(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105512245 B
(45)授权公告日 2018.08.21

(21)申请号 201510862236.0

(22)申请日 2015.11.30

(65)同一申请的已公布的文献号
申请公布号 CN 105512245 A

(43)申请公布日 2016.04.20

(73)专利权人 青岛智能产业技术研究院
地址 266109 山东省青岛市高新区创业大厦B座26楼

(72)发明人 倪晓春 曾帅 张杰 袁勇 王飞跃

(74)专利代理机构 青岛联信知识产权代理事务所(普通合伙) 37227
代理人 徐艳艳

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

(56)对比文件
CN 104657425 A,2015.05.27,
CN 103309990 A,2013.09.18,
CN 104217296 A,2014.12.17,
CN 104657425 A,2015.05.27,
CN 104268197 A,2015.01.07,
CN 102110140 A,2011.06.29,
US 9141916 B1,2015.09.22,

审查员 李小敏

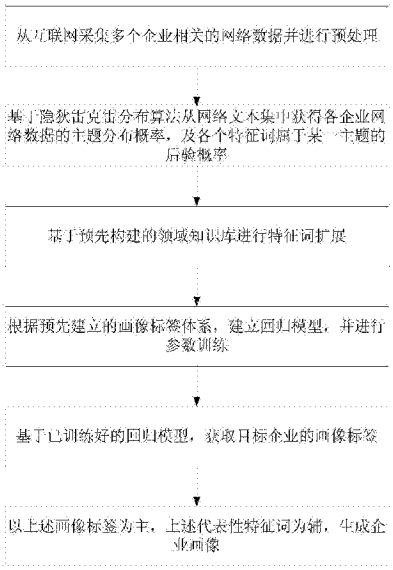
权利要求书2页 说明书4页 附图3页

(54)发明名称

一种基于回归模型建立企业画像的方法

(57)摘要

本发明提供一种基于回归模型建立企业画像的方法，包括从互联网采集企业相关的网络数据并进行预处理，获得各企业的网络文本集及由文本特征词构成的特征词库；基于隐狄雷克雷分布算法从网络文本集中获得各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率；基于预先构建的领域知识库进行特征词扩展；根据预先建立的画像标签体系，建立回归模型，并进行参数训练；基于已训练好的回归模型，获取目标企业的画像标签；以上述画像标签为主，上述代表性特征词为辅，生成企业画像。该方法充分利用社会情报文本的潜在语义信息，弥补传统企业画像方法的不足，丰富企业网络形象层次，从多粒度勾勒网民对企业的认知感。该方法准确度高、易于实现、易于扩展，具有良好的运用前景和可观的市场价值。



CN 105512245 B

1.一种基于回归模型建立企业画像的方法，其特征在于，该方法包括以下步骤：

步骤一、获取企业的代表性特征词，具体步骤如下：

S1：从互联网采集企业的企业舆情相关数据并进行预处理，获得企业的网络文本集及由文本特征词构成的特征词库；

S2：基于隐狄雷克雷分布算法从网络文本集中获得企业舆情相关数据的主题分布概率，及各个特征词属于某一主题的后验概率；

S2：基于预先构建的领域知识库进行特征词扩展；

步骤二、获取企业的画像标签，具体包括，建立画像标签体系及回归模型，选取企业样本，根据画像标签体系对回归模型进行参数训练；并基于已训练好的回归模型，获取目标企业的画像标签；

步骤三、以步骤二获取的画像标签为主，步骤一获取的代表性特征词为辅，生成企业画像。

2.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤S1中，按如下步骤进行：

1)企业舆情相关数据采集，包括新闻、微博、微信、论坛多通道社会情报；

2)基于XPATH或者正则表达式的方法将文本中包含的非文本数据和冗余信息删除，获得各企业的结构化的网络文本集；

3)对特定企业相关的网络文本进行分词处理，利用统计算法从分词结果中获取各企业相关网络数据的文本特征词。

3.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤S2中，按如下步骤进行：

1)基于预先构建的与企业领域相关的自定义词典，对企业相关的网络文本进行分词处理，利用统计算法从分词结果中获取各企业相关网络数据的文本特征词；

2)将分词后的文本转化为词的向量空间表示，生成稀疏矩阵，同时创建：一个字典(wordIndex,word)，一个词频数统计表(wordIndex,count)，以及一个文档频率表(wordIndex,DFCount)；

3)索引创建，将字符串转换为数字编号，生成docIndex，即文档索引<文档id,文档名>，以及数字化的矩阵，即<文档id,文档名:{词汇id:tfidf列表}>；

4)设定隐含主题数、迭代次数运行算法，生成隐狄雷克雷分布算法模型；

5)利用生成的隐狄雷克雷分布算法模型推导出各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率。

4.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤S3中，领域知识库的构建过程如下：

1)数据采集：从特定领域的百科网站和专业字典网站获取所有领域词条页面信息与领域词条数据；

2)信息抽取：从领域百科获取的领域词条页面信息中抽取领域词条的相关信息；

3)百科词条关系构建：构建百科词条的正向及反向关系，即依据百科网站的词条的相关词条获得与相关词条相关的词条集合；

4)领域知识集成：依据词条名称将源于不同领域百科网站的词条进行合并和去重，将

百科网站的数据集成结果与领域专业字典获得的词汇进行合并和去重。

5.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤S3中，特征词的扩展过程如下：

1)基于步骤S2输出结果中选取高概率主题的代表性特征词；

2)再基于预先构建的领域知识库进行同义词、近义词、关联词扩展，从语义上对主题包含的特征词进行扩展，从而丰富企业特征词。

6.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤二中，所述画像标签体系的建立过程为：

1)数据采集：从特定领域的百科网站获取领域词条页面信息；

2)信息抽取：从领域词条页面信息中抽取领域词条的标签数据；

3)标签融合：将源于不同领域百科网站的词条标签进行合并和去重；

4)人工校准：由领域专家对融合后的标签进行过滤和梳理，构建领域画像标签体系。

7.根据权利要求1所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤二中，所述回归模型为：

$$Y=\frac{1}{1+\exp(-\omega_i^T\cdot X+\omega_0^T)} \quad (1)$$

其中，因变量$Y=(y_0,\cdots,y_m)^T$为画像标签概率，$y_i\in[0,1]$，m为画像标签的个数，自变量$X=(x_0,\cdots,x_n)^T$为目标企业相关网络文本的主题分布概率，$x_i\in[0,1]$，n为主题的个数，ω_i^T为回归系数矩阵，ω_0^T为残差矩阵。

8.根据权利要求6所述的一种基于回归模型建立企业画像的方法，其特征在于，步骤二中，所述回归模型的训练方法为：

根据预先建立画像标签体系，选取部分企业的网络数据进行人工标注，获取这些企业的画像标签$Y=(y_0,\cdots,y_m)^T$，$y_i\in\{0,1\}$，其中$y_i=1$表示企业具有该标签，$y_i=0$表示企业不具有该标签，对应于公式(1)，以这些企业的主题概率分布作为自变量，以标注的画像标签为因变量，拟合求解回归系数ω_i^T和残差ω_0^T。

9.根据权利要求7所述的一种基于回归模型建立企业画像的方法，其特征在于，基于回归模型的画像标签获取方法为：以目标企业的主题概率分布作为自变量，输入到训练好的回归模型，得到其画像标签概率$Y=(y_0,\cdots,y_m)^T$，若$y_i\geq0.5$，则判定企业具有该标签，否则判定企业不具有该标签。

一种基于回归模型建立企业画像的方法

技术领域

[0001] 本发明涉及模式识别领域技术领域，具体地说，涉及一种基于回归模型建立企业画像的方法。

背景技术

[0002] 随着移动互联网、物联网等新技术的迅速发展，人类进入数据时代。过去由媒体垄断的传播资源，今天被数以亿计的个体所分享；传播速度以秒传计；组织传播与个体传播、媒体传播与自媒体传播在不断融合与分化的状态中推动企业舆情的发展。

[0003] 企业形象和声誉在互联网上以碎片化方式呈现。如何从全媒体海量数据中获得企业在网民中的认知度，绘制出全面的企业网络形象，建立企业的画像，成为企业迫切需求。

发明内容

[0004] 为了解决上述问题，本发明提供一种基于回归模型建立企业画像的方法，其具体的技术方案如下：

[0005] 一种基于回归模型建立企业画像的方法，该方法包括以下步骤：

[0006] 步骤一、获取企业的代表性特征词，具体步骤如下：

[0007] S1:从互联网采集企业的企业舆情相关数据并进行预处理，获得企业的网络文本集及由文本特征词构成的特征词库；

[0008] S2:基于隐狄雷克雷分布算法从网络文本集中获得企业舆情相关数据的主题分布概率，及各个特征词属于某一主题的后验概率；

[0009] S3:基于预先构建的领域知识库进行特征词扩展；

[0010] 步骤二、获取企业的画像标签，具体包括，建立画像标签体系及回归模型，选取企业样本，根据画像标签体系对回归模型进行参数训练；并基于已训练好的回归模型，获取目标企业的画像标签；

[0011] 步骤三、以步骤二获取的画像标签为主，步骤一获取的代表性特征词为辅，生成企业画 像。

[0012] 进一步，步骤S1中，按如下步骤进行：

[0013] 1)企业舆情相关数据采集，包括新闻、微博、微信、论坛多通道社会情报；

[0014] 2)基于XPATH或者正则表达式的方法将文本中包含的非文本数据和冗余信息删除，获得各企业的结构化的网络文本集；

[0015] 3)对特定企业相关的网络文本进行分词处理，利用统计算法从分词结果中获取各企业相关网络数据的文本特征词。

[0016] 进一步，步骤S2中，按如下步骤进行：

[0017] 1)基于预先构建的与企业领域相关的自定义词典，对企业相关的网络文本进行分词处理，利用统计算法从分词结果中获取各企业相关网络数据的文本特征词；

[0018] 2)将分词后的文本转化为词的向量空间表示，生成稀疏矩阵，同时创建：一个字典

(wordIndex,word)，一个词频数统计表(wordIndex,count)，以及一个文档频率表(wordIndex,DFCount)；

[0019] 3)索引创建，将字符串转换为数字编号，生成docIndex，即文档索引<文档id，文档名>，以及数字化的矩阵，即<文档id,文档名:{词汇id:tfidf列表}>；

[0020] 4)设定隐含主题数、迭代次数运行算法，生成隐狄雷克雷分布算法模型；

[0021] 5)利用生成的隐狄雷克雷分布算法模型推导出各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率。

[0022] 进一步，步骤S3中，领域知识库的构建过程如下：

[0023] 1)数据采集：从特定领域的百科网站和专业字典网站获取所有领域词条页面信息与领域词条数据；

[0024] 2)信息抽取：从领域百科获取的领域词条页面信息中抽取领域词条的相关信息(包括：词条标题、词条分类、词条推荐、同义词、近义词等)；

[0025] 3)百科词条关系构建：构建百科词条的正向及反向关系，即依据百科网站的词条的相关词条获得与相关词条相关的词条集合；

[0026] 4)领域知识集成：依据词条名称将源于不同领域百科网站的词条进行合并和去重，将百科网站的数据集成结果与领域专业字典获得的词汇进行合并和去重。

[0027] 进一步，步骤S3中，特征词的扩展过程如下：

[0028] 1)基于步骤S2输出结果中选取高概率主题的代表性特征词；

[0029] 2)再基于预先构建的领域知识库进行同义词、近义词、关联词扩展，从语义上对主题包含的特征词进行扩展，从而丰富企业特征词。

[0030] 进一步，步骤二中，所述回归画像标签体系的建立过程为：

[0031] 1)数据采集：从特定领域的百科网站获取领域词条页面信息；

[0032] 2)信息抽取：从领域词条页面信息中抽取领域词条的标签数据(包括：词条分类、段落标题、词条属性等)；

[0033] 3)标签融合：将源于不同领域百科网站的词条标签进行合并和去重；

[0034] 4)人工校准：由领域专家对融合后的标签进行过滤和梳理，构建领域画像标签体系。

[0035] 进一步，步骤二中，所述回归模型为：

[0036] $$Y=\frac{1}{1+\exp(-\omega_1^T X+\omega_0^T)} \quad (1)$$

[0037] 其中，因变量$Y=(y_0,\ldots,y_m)^T$为画像标签概率，$y_i\in[0,1]$，m为画像标签的个数，自变量$X=(x_c,\ldots,x_n)^T$为目标企业相关网络文本的主题分布概率，$x_i\in[0,1]$，n为主题的个数，ω_1^T为回归系数矩阵，ω_0^T为残差矩阵。

[0038] 进一步，步骤二中，所述回归模型的训练方法为：

[0039] 根据预先建立画像标签体系，选取部分企业的网络数据进行人工标注，获取这些企业的画像标签$Y=(y_0,\ldots,y_m)^T$，$y_i\in\{0,1\}$，其中$y_i=1$表示企业具有该标签，$y_i=0$表示企业不具有该标签。对应于公式(1)，以这些企业的主题概率分布作为自变量，以标注的画像标签为因变量，拟合求解回归系数ω_1^T和残差ω_0^T。

[0040] 进一步，基于回归模型的画像标签获取方法为：以目标企业的主题概率分布作为

自变量，输入到训练好的回归模型，得到其画像标签概率$Y=(y_0,...,y_m)^T$，若$y_i \geq 0.5$，则判定企业具有该标签，否则判定企业不具有该标签。

[0041] 本发明所提供的一种基于回归模型建立企业画像的方法，具有以下优点：

[0042] 本发明提出了基于回归模型对企业舆情进行建模实施企业画像的方法，是一种基于主题概率分布实施企业画像方法。传统的企业画像方法采用统计的方法提取高频词汇作为画像标签，忽略文本的潜在语义信息。而基于回归模型的企业画像方法，是利用文本自身潜在语义作为特征，不依赖于孤立词语相似度对比，具有更好的通用性与易用性，能够更好的表达出文本潜在语义结构，从而达到更好的企业画像效果。

[0043] 本发明依据新闻、微博、微信、论坛多通道数据，挖掘网络文本潜在语义信息，分层次多粒度勾勒企业网络形象特征，为企业观察、理解和应对复杂的舆论生态环境提供了工具和条件，具有良好的运用前景和可观的市场价值。

附图说明

[0044] 图1是本发明方法的流程图。

[0045] 图2是本发明方法的实现流程图。

[0046] 图3是根据本发明实施企业画像效果示意图。

具体实施方式

[0047] 下面结合附图及本发明的实施例对本发明的一种基于回归模型建立企业画像的方法作进一步详细的说明。

[0048] 本发明所提出的一种基于回归模型建立企业画像的方法包括以下步骤：

[0049] 步骤1，从互联网采集企业相关的网络数据并进行预处理，获得各企业的网络文本集及由文本特征词构成的特征词库。

[0050] 互联网产生了海量的企业相关的异构文本数据(新闻、博客、论坛、微薄、微信等全媒体数据)，这些文本基本都是半结构HTML格式，且包含大量的非文本数据，需要将这些无用信息过滤掉。采用基于XPATH或者正则表达式的方法将这些信息从每个文本中删除，统一处理为结构化信息，且每家企业信息融合在一起。以青岛知名企业为例，总计采集1000家青岛企业相关的网络数据，将这些半结构HTML数据结构化信息清洗与整理后，获得各企业的网络文本集；

[0051] 基于预先构建的与企业领域相关的自定义词典，对特定企业相关的网络文本进行分词处理，利用统计算法从分词结果中获取各企业相关网络数据的文本特征词。

[0052] 步骤2，基于隐狄雷克雷分布算法从网络文本集中获得各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率。

[0053] 将分词后的文本转化为词的向量空间表示，生成稀疏矩阵，同时创建：一个字典(wordIndex,word)，一个词频数统计表(wordIndex,count)，以及一个文档频率表(wordIndex,DFCount)，并基于最大的文档频率DF百分比移除高频语汇；

[0054] 为了方便计算，进行索引创建，将字符串转换为数字编号，生成docIndex，即文档索引<文档id,文档名>，以及数字化的矩阵，即<文档id,文档名:{词汇id:tfidf列表}>；

[0055] 设定隐含主题数、迭代次数运行算法，生成隐狄雷克雷分布算法模型；

[0056] 利用生成的隐狄雷克雷分布算法模型推导出各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率，例如：topic_0

[0057] 家电[p(家电|topic_0)]=0.155923

[0058] 智能[p(智能|topic_0)]=0.078596

[0059] 物流[p(物流|topic_0)]=0.006325

[0060] 步骤3，基于预先构建的领域知识库进行特征词扩展。

[0061] 领域知识库是基于领域百科以及领域专业字典构建的，构建过程如下所述：

[0062] 数据采集：从特定领域的百科网站和专业字典网站获取所有领域词条页面信息与领域词条数据；

[0063] 信息抽取：从领域百科获取的领域词条页面信息中抽取领域词条的相关信息(包括：词条标题、词条分类、词条推荐、同义词、近义词等)；

[0064] 百科词条关系构建：构建百科词条的正向及反向关系，即依据百科网站的词条的相关词条获得与相关词条相关的词条集合；

[0065] 领域知识集成：依据词条名称将源于不同领域百科网站的词条进行合并和去重，将百科网站的数据集成结果与领域专业字典获得的词汇进行合并和去重。

[0066] 构建完成领域知识库后，可以进行特征词扩展，具体步骤如下：

[0067] 首先，基于步骤2输出结果中选取高概率主题的代表性特征词，再基于预先构建的领域知识库进行同义词、近义词、关联词扩展，从语义上对主题包含的特征词进行扩展，从而丰富企业特征词。

[0068] 步骤4，本发明利用Logistic回归模型建立企业网络文本的主题概率分布X和企业画像标签Y的回归关系，从而估计企业的画像标签。所述Logistic回归模型用公式可表示为：

[0069] $$Y=\frac{1}{1+\exp(-\omega_1^T X+\omega_0^T)} \quad (2)$$

[0070] 其中，ω_1^T为回归系数矩阵，ω_0^T为残差矩阵。

[0071] 随机从1000家企业中选取100家，根据预先建立画像标签体系$(d_0,\ldots,d_m)^T$，

[0072] 对这些企业的网络数据进行人工标注，获取企业画像标签$Y=(y_0,\ldots,y_m)^T$。其中$y_i\in\{0,1\}$与d_i一一对应，$y_i=1$表示企业具有标签d_i，$y_i=0$表示企业不具有该标签d_i。以这些企业的主题概率分布作为自变量，以标注的画像标签为因变量，通过最大似然估计法求解回归系数ω_1^T。

[0073] 步骤5，对未被人工标注的其余900家企业，依次选择其一作为目标企业，以目标企业的主题概率分布作为自变量，输入到训练好的回归模型，得到其画像标签概率 $Y=(y_0,\ldots,y_m)^T$，若$y_i\geq 0.5$，则判定该企业具有标签d_i，否则判定该企业不具有标签d_i。

[0074] 步骤6，以步骤5获取的画像标签为主，步骤3获取的代表性特征词为辅，生成企业画像，如图3可以看到青岛海尔的企业画像。

[0075] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

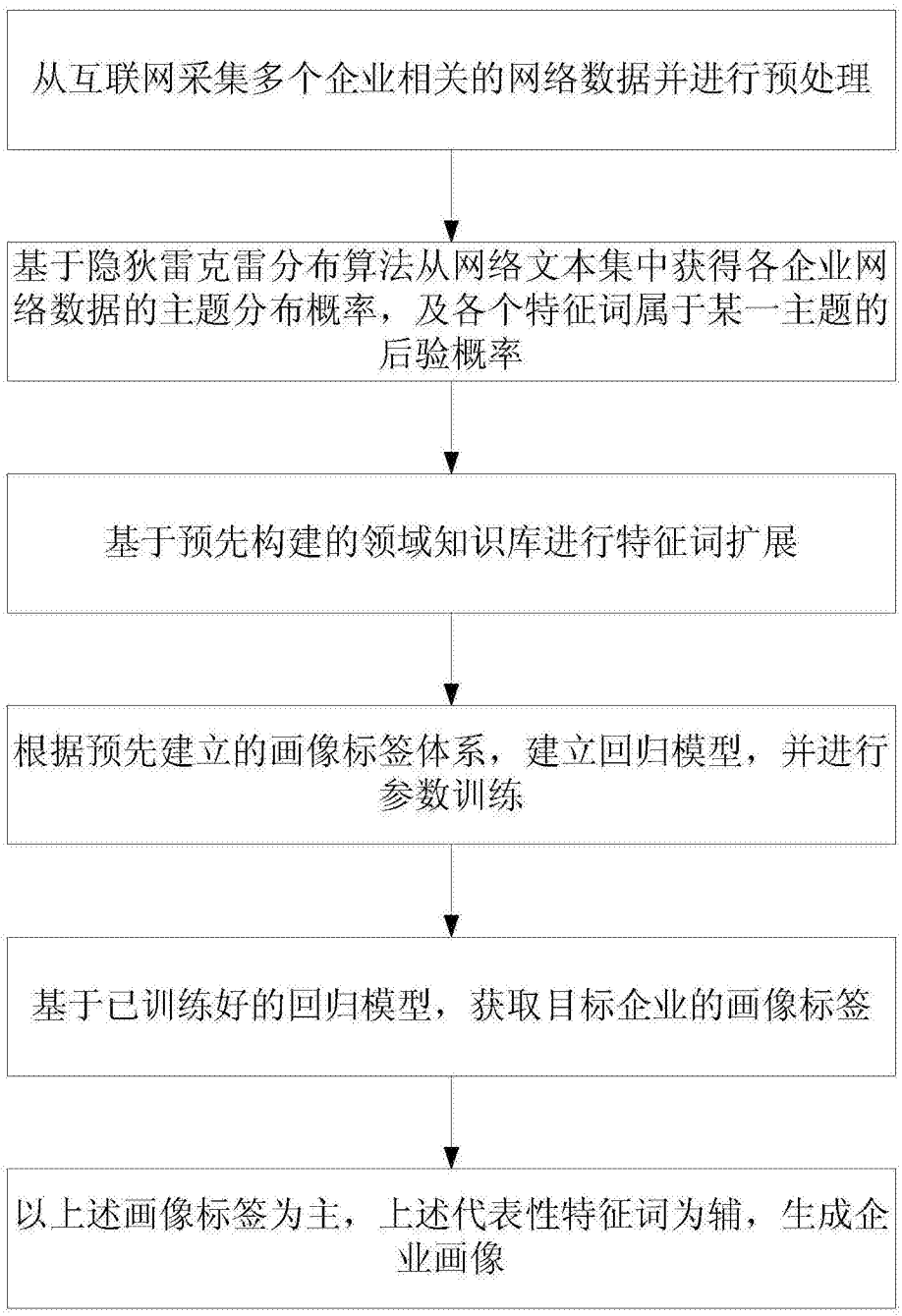

从互联网采集多个企业相关的网络数据并进行预处理
基于隐狄雷克雷分布算法从网络文本集中获得各企业网络数据的主题分布概率，及各个特征词属于某一主题的后验概率
基于预先构建的领域知识库进行特征词扩展
根据预先建立的画像标签体系，建立回归模型，并进行参数训练
基于已训练好的回归模型，获取目标企业的画像标签
以上述画像标签为主，上述代表性特征词为辅，生成企业画像

图1

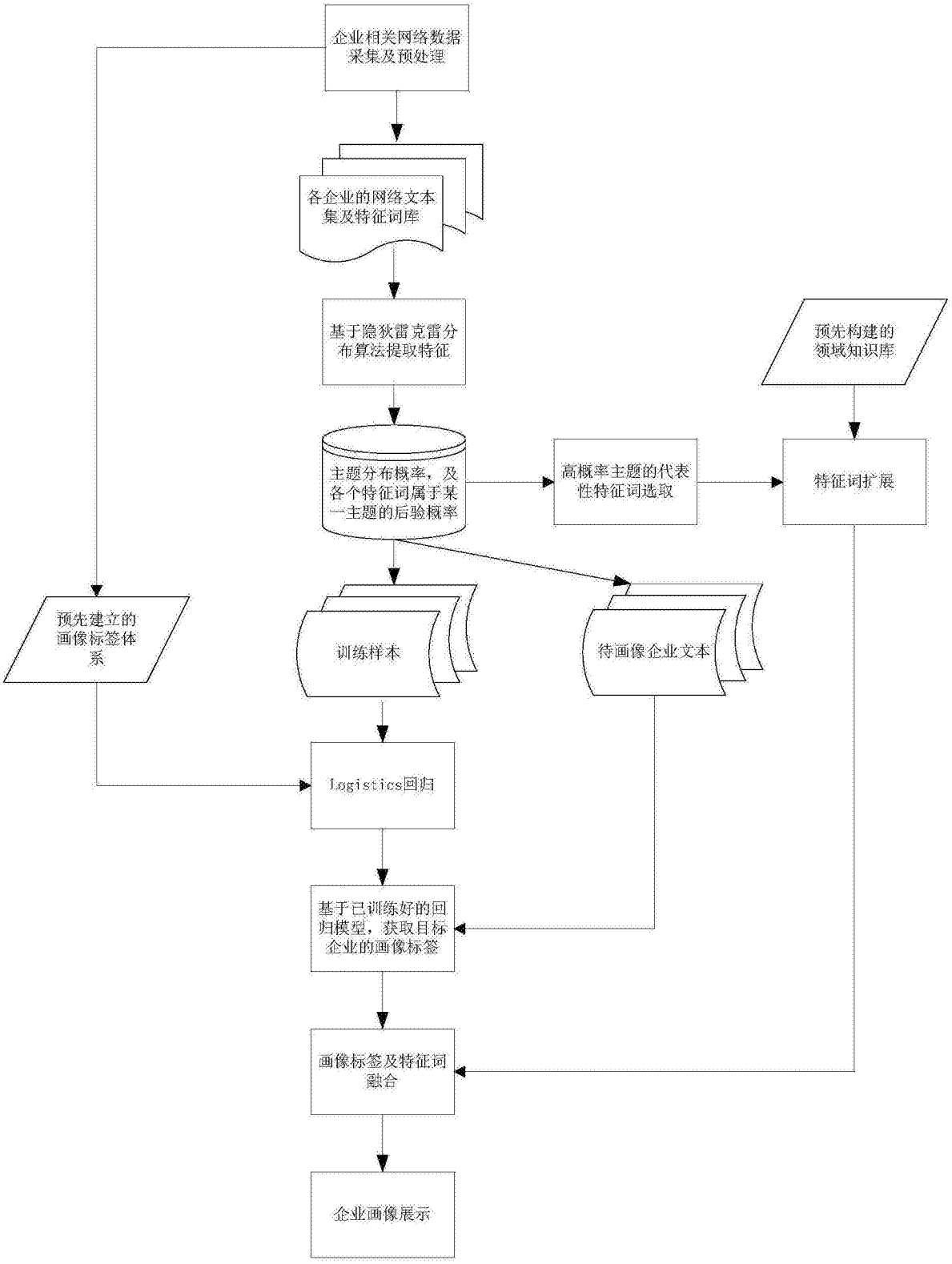

企业相关网络数据采集及预处理
各企业的网络文本集及特征词库
基于隐狄雷克雷分布算法提取特征
预先构建的领域知识库
主题分布概率，及各个特征词属于某一主题的后验概率
高概率主题的代表性特征词选取
特征词扩展
预先建立的画像标签体系
训练样本
待画像企业文本
Logistics回归
基于已训练好的回归模型，获取目标企业的画像标签
画像标签及特征词融合
企业画像展示

图2

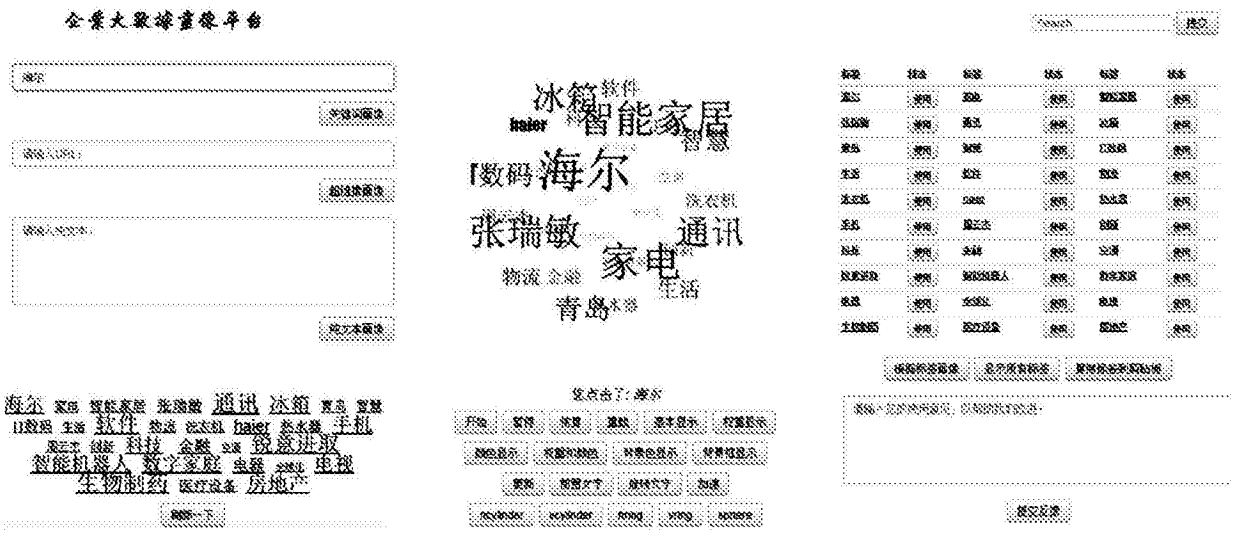


图3