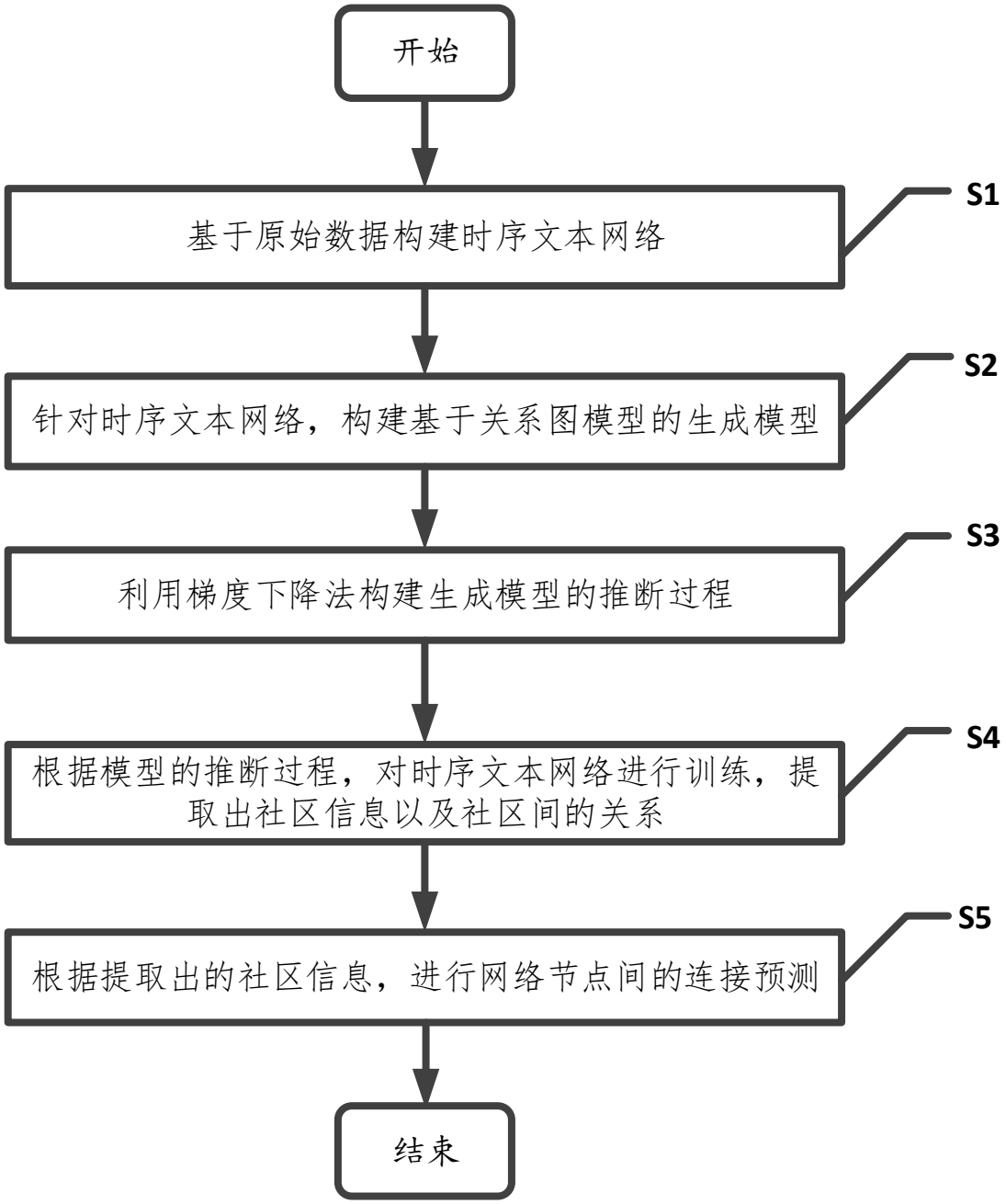
说　明　书　摘　要

本发明公开了一种基于时序文本网络的社区检测与用户关系预测方法，包括：基于原始数据构建时序文本网络；针对时序文本网络，构建基于关系图模型的生成模型；利用梯度下降法构建生成模型的推断过程；根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系；根据提取出的社区信息，进行网络节点间的连接预测。本发明构建了全新的社区检测方法，并提出了社区相关度的概念，大幅提升了社区检测的准确性和解释性。

摘　要　附　图


开始
基于原始数据构建时序文本网络
S1
针对时序文本网络，构建基于关系图模型的生成模型
S2
利用梯度下降法构建生成模型的推断过程
S3
根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系
S4
根据提取出的社区信息，进行网络节点间的连接预测
S5
结束

5

权 利 要 求 书

1.一种基于时序文本网络的社区检测与用户关系预测方法，其特征在于，包括如下步骤：

步骤S1：基于原始数据构建时序文本网络；

步骤S2：针对时序文本网络，构建基于关系图模型的生成模型；

步骤S3：利用梯度下降法构建生成模型的推断过程；

步骤 S4：根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系，其中社区指表现出较高相关性的点的集合，社区间的关系指的是社区之间的相似度；

步骤S5：根据提取出的社区信息，进行网络节点间的连接预测。

2.根据权利要求 1 所述的基于时序文本网络的社区检测与用户关系预测方法，其特征在于，所述步骤S1包括：

步骤S101：将顶点集V设为空集，将边集E设为空集；

步骤S102：将原始数据集中的每一篇文章加到顶点集V中；

步骤S103：顶点集V中的每一篇文章对应一个标签T，该标签是指每一篇文章的发表时间；

步骤S104：将原始数据集中文章间的链接关系加到边集E中；

步骤S105：(V，E；T)的集合构成图G，图G为时序文本网络。

3.根据权利要求 2 所述的基于时序文本网络的社区检测与用户关系预测方法，其特征在于，所述步骤S2包括：

步骤S201：定义节点u与节点v之间通过社区i、j产生连接的概率：

$$p(u,v,i,j)=\left(1-\exp(-F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度；**错误!未找到引用源。**表示节点**错误!未找到**

错误!未找到引用源。与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点 u 的时间戳;**错误!未找到引用源。**表示节点 v 的时间戳;

步骤 S202：定义节点 u 与节点 v 之间通过任意两个社区产生连接的概率为：

$$p(u,v)=\left(1-\exp(-\sum_{i,j}F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v)$$
$$=\left(1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度；$\mathbf{F_u^T}$表示节点**错误!未找到引用源。**与所有社区的连接强度的向量的转置；$\mathbf{F_v}$表示节点 v 与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;

步骤 S203：针对时序文本网络，根据步骤 S202 定义的公式，生成时序文本网络**错误!未找到引用源。**：

$$G^P=(V\bigcup V_\omega, E\bigcup E_{\omega d}; T\bigcup T_\omega)$$

其中，V、E、T 分别是时序文本网络中的节点集合、边集合以及时间戳集合；**错误!未找到引用源。**代表一个单词；存在于**错误!未找到引用源。**的边(**错误!未找到引用源。**,**错误!未找到引用源。**)代表单词 i 存在于文章 j 中；**错误!未找到引用源。**代表单词的时间戳，被设置成 0；对

于该网络中任意两点，根据 S202 所定义的概率，预测两点间是否有边存在。

4. 根据权利要求 3 所述的基于时序文本网络的社区检测与用户关系预测方法，其特征在于，所述步骤 S3 包括：

步骤 S301：利用块坐标梯度下降法，对于对每个节点 u，假设对 $\forall v \neq u, \mathbf{F}_v$ 不变且 $\boldsymbol{\eta}$ 不变，首先更新 $\mathbf{F}_u$，即 $\hat{\mathbf{F}}_\mathbf{u} = \arg\max_{\mathbf{F}_\mathbf{u} \geq 0} l(\mathbf{F}_\mathbf{u})$，$l(\mathbf{F}_\mathbf{u})$ 为针对 $\mathbf{F}_\mathbf{u}$ 的对数似然函数，具体地：

$$l(\mathbf{F}_\mathbf{u}) = \sum_{v \in inN(u)} log(1-\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})) - \sum_{\substack{v \notin N(u) \\ t(v)<t(u)}} \mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u} + \sum_{v' \in outN(u)} log(1-\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{v'})) - \sum_{\substack{v' \notin N(u) \\ t(v')>t(u)}} \mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{v'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathrm{v}$ 和 $\mathbf{F}_{v'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵，$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S302：利用梯度下降法，根据如下公式可以进行对 F 的更新：

$$F_{uk}^{new} \leftarrow \max\{0, F_{uk}^{old} + \alpha_{F_u}(\nabla l(\mathbf{F}_\mathbf{u}))_k\}$$

其中 α_{F_u} 为利用回溯搜索算法计算所得步长；F_{uk}^{new} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新后的值；F_{uk}^{old} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新前的值；$\nabla l(\mathbf{F}_\mathbf{u}))$ 为更新 $\mathbf{F}_\mathbf{u}$ 时所用的梯度，具体的：

$$\nabla l(\mathbf{F}_\mathbf{u})) = \sum_{v \in inN(u)} \frac{\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})}{1-\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})}\boldsymbol{\eta}^\mathbf{T}\mathbf{F}_\mathbf{v} - \sum_{\substack{v \notin N(u) \\ t(v)<t(u)}} \boldsymbol{\eta}^\mathbf{T}\mathbf{F}_\mathbf{v} + \sum_{v' \in outN(u)} \frac{\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{v'})}{1-\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{v'})}\boldsymbol{\eta}\mathbf{F}_{v'} - \sum_{\substack{v' \notin N(u) \\ t(v')>t(u)}} \boldsymbol{\eta}\mathbf{F}_{v'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathrm{v}$ 和 $\mathbf{F}_{v'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵；$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$、$\boldsymbol{\eta}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S303：$\mathbf{F}$ 更新完成后，假设 $\mathbf{F}$ 不变，根据如下公式可以进行对 $\boldsymbol{\eta}$ 的更新：

$$\eta_{ij}^{new} \leftarrow \max\{0, \eta_{ij}^{old} + \alpha_{\eta}\left(\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta})\right)_{ij}\}$$

其中α_{η}为利用回溯搜索算法计算所得步长；η_{ij}^{new}为$\boldsymbol{\eta}$矩阵中第 i 行第 j 列更新后的值；η_{ij}^{old}为$\boldsymbol{\eta}$矩阵中第 i 行第 j 列更新前的值；$\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta})$为更新$\boldsymbol{\eta}$时所用的梯度，具体的：

$$\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta}) = \sum_{(u \to v) \in E} \frac{exp(-\mathbf{F}_u^{\mathrm{T}} \boldsymbol{\eta} \mathbf{F}_v)}{1 - exp(-\mathbf{F}_u^{\mathrm{T}} \boldsymbol{\eta} \mathbf{F}_v)} \mathbf{F}_u \mathbf{F}_v^{\mathrm{T}} - \sum_{\substack{(u \to v) \notin E \\ t(u) < t(v)}} \mathbf{F}_u \mathbf{F}_v^{\mathrm{T}}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$为对应的转置矩阵；$(u \to v)$表示从点 u 指向点 v 的边；t(u)与 t(v)分别表示点 u 与点 v 的时间戳；

步骤 S304：计算$\nabla l(\mathbf{F}_u))$和$\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta})$的时间复杂度分别为$O(N)$和$O(N^2)$，为降低时间复杂度、提高可计算性，采取如下近似：

$$\sum_{\substack{v \in N(u) \\ t(v) < t(u)}} h^{\mathrm{T}} \mathbf{F}_v = \sum_{t(v) < t(u)} h^{\mathrm{T}} \mathbf{F}_v - \sum_{v \in inN(u)} h^{\mathrm{T}} \mathbf{F}_v,$$

$$\sum_{\substack{v \in N(u) \\ t(v) > t(u)}} h \mathbf{F}_{v} = \sum_{t(v) > t(u)} h \mathbf{F}_{v} - \sum_{v \in outN(u)} h \mathbf{F}_{v},$$

$$\sum_{\substack{(u \to v) \notin E \\ t(u) < t(v)}} \mathbf{F}_u \mathbf{F}_v^{\mathrm{T}} = \sum_{\substack{(u \to v) \\ t(u) < t(v)}} \mathbf{F}_u \mathbf{F}_v^{\mathrm{T}} - \sum_{\substack{(u \to v) \in E \\ t(u) < t(v)}} \mathbf{F}_u \mathbf{F}_v^{\mathrm{T}}.$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合；N(u)表示 inN(u)和 outN(u)的并集，$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$、$\boldsymbol{\eta}^{\mathrm{T}}$为对应的转置矩阵；$(u \to v)$表示从点 u 指向点 v 的边；t(u)、t(v)、t(v’)分别表示点 u、点 v 与点 v’的时间戳；

计算$\nabla l(\mathbf{F}_u))$和$\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta})$的时间复杂度分别为$O(|N(u)|)$、$O(|E|)$，总时间复杂度为$O(|E|)$，中$|N(u)|$表示集合$N(u)$包含的节点的个数；$|E|$表示网络

中边的条数。

5.根据权利要求 4 所述的基于时序文本网络的社区检测与用户关系预测方法，其特征在于，所述步骤 S4 包括：

步骤 S401：从数据文件中读取数据，并根据步骤 S1 构建时序文本网络；

步骤 S402：初始化用户与社区间的联系强度矩阵 F；基于向网络中的导率模型，如果节点 u 的入邻居 inN(u)有比所有点 v∈outN(u)的入邻居 inN(v)有更小的导率，则该入邻居 inN(u)在邻近是最小的；对于属于一个在邻近最小的邻域 k 内的节点 u'，初始化节点 u'与一个社区 k 之间的联系强度 Fu'k = 1，否则令 Fu'k = 0；为了初始化 η，设置主对角线上的项为 0.9，其他项为 0.1；

步骤 S403：每轮次根据公式更新 F 与 η，首先针对每个节点 u，根据梯度公式更新节点 u 与所有社区之间的联系强度向量 $\mathbf{F_u}$，梯度公式如下：

$$\nabla l(\mathbf{F_u}) = \sum_{v\in inN(u)} \frac{\exp(-\mathbf{F_v^T}\boldsymbol{\eta}\mathbf{F_u})}{1-\exp(-\mathbf{F_v^T}\boldsymbol{\eta}\mathbf{F_u})}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{F_v} - \sum_{\substack{v\notin N(u)\\ t(v)<t(u)}} \boldsymbol{\eta}^{\mathrm{T}}\mathbf{F_v} + \sum_{v'\in outN(u)} \frac{\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_{v'}})}{1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_{v'}})}\boldsymbol{\eta}\mathbf{F_{v'}} - \sum_{\substack{v'\notin N(u)\\ t(v')>t(u)}} \boldsymbol{\eta}\mathbf{F_{v'}}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F_u^T}$、$\mathbf{F_v^T}$、$\boldsymbol{\eta}^{\mathbf{T}}$为对应的转置矩阵；

F 更新完成后，根据梯度公式更新社区间的联系矩阵 η，梯度公式如下：

$$\nabla_{\eta} l(\mathbf{F},\boldsymbol{\eta}) = \sum_{(u\to v)\in E} \frac{exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})}{1-exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})}\mathbf{F_u}\mathbf{F_v^T} - \sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}} \mathbf{F_u}\mathbf{F_v^T}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节

点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F_u^T}$、$\mathbf{F_v^T}$为对应的转置矩阵；$(u \to v)$表示从点 u 指向点 v 的边；t(u)与 t(v)分别表示点 u 与点 v 的时间戳；

步骤 S404：经过一定轮次后，判定每个节点与社区间的隶属关系，针对每个社区 k，设定一个阈值δ_k，具体设定方法如下：

$$\delta_k = \sqrt{-\frac{log(1-1/N)}{\eta_{kk}}}$$

其中 N 为节点总数；η_{kk}为社区间联系矩阵$\boldsymbol{\eta}$第 k 行第 k 列的分量，对于节点 u 与社区 k，若联系强度F_{uk}大于社区 k 的阈值δ_k，则认为节点 u 隶属于社区 k。

6. 根据权利要求 5 所述的基于时序文本网络的社区检测与用户关系预测方法，其特征在于，所述步骤 S5 包括：

步骤 S501：对选定的文本数据集进行训练，提取出节点与社区间联系强度矩阵 F，以及社区间的联系关系矩阵$\boldsymbol{\eta}$；

步骤 S502：读取矩阵 F 与矩阵$\boldsymbol{\eta}$；

步骤 S503：根据步骤 S2 所定义的公式计算节点 u 与节点 v 之间边的存在概率。

说　　明　　书

基于时序文本网络的社区检测与用户关系预测方法

技术领域

本发明涉及到时序文本网络探社区检测领域，具体地，涉及一种基于时序文本网络的社区检测与用户关系预测方法。

背景技术

网络是一个强大的语言，它能够阐释社会、自然以及学术领域中的数据关系。一个理解网络的方法是定义和分析一组有着相同属性的节点。这样的一组节点可以被解释为社交网络中的组织单位，或者引用网络中的相同领域。探测社区问题就是在网络中寻找这样的一组节点的研究任务。传统的方法大都基于一个节点只属于一个社区这个假设，集中寻找离散社区。那么在除去这个假设的情况下，交叉社区检测问题变得越来越普遍并在最近引起了越来越多的关注。

尽管在过去网络中的交叉多等级社区问题已经被讨论过，但在一个大的网络中定义一个有意义的社区网络依旧是个艰难的任务。大多数方法很难应用于大型网络，并且在缺少有信服力的标准情况下，对检测出的社区进行评估极其困难。因此，尽管网络问题已经被广泛的研究，小型网络中的社区的存在和特性已经被熟知，在特大型网络中定义交叉社区的方法依旧不甚清晰。

探测重叠社区一般有两种形式的信息可以利用。第一种是链型结构，例如边的有无。经典方法大都集中于这种形式的信息，并致力于获取一组节点，这些节点之间的连接相比于外部网络而言更为紧密。第二种是节点属性，包括在线的用户档案，预先存在的蛋白质功能和论文的文本内容。由于链接结构中普遍存在的噪音，同时基于这两种方法检测社区信息的方法已经越来越受欢迎。

发明内容

针对现有技术中的缺陷，本发明的目的是提供一种基于时序文本网

络的社区检测与用户关系预测方法，研究在时序文本网络中探测交叉社区的问题，在时序文本网络识别有意义的社区为后续应用开发提供了有用的知识。

为实现上述目的，本发明是根据以下技术方案实现的：

一种基于时序文本网络的社区检测与用户关系预测方法，包括如下步骤：

步骤 S1：基于原始数据构建时序文本网络；

步骤 S2：针对时序文本网络，构建基于关系图模型的生成模型；

步骤 S3：利用梯度下降法构建生成模型的推断过程；

步骤 S4：根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系，其中社区指表现出较高相关性的点的集合，社区间的关系指的是社区之间的相似度；

步骤 S5：根据提取出的社区信息，进行网络节点间的连接预测。

上述技术方案中，所述步骤 S1 包括：

步骤 S101：将顶点集 V 设为空集，将边集 E 设为空集；

步骤 S102：将原始数据集中的每一篇文章加到顶点集 V 中；

步骤 S103：顶点集 V 中的每一篇文章对应一个标签 T，该标签是指每一篇文章的发表时间；

步骤 S104：将原始数据集中文章间的链接关系加到边集 E 中；

步骤 S105：(V，E；T)的集合构成图 G，图 G 为时序文本网络。

上述技术方案中，所述步骤 S2 包括：

步骤 S201：定义节点 u 与节点 v 之间通过社区 i、j 产生连接的概率：

$$p(u,v,i,j)=\left(1-\exp(-F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度；**错误!未找到引用源。**表示节点**错误!未找到**

引用源。与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点 u 的时间戳;**错误!未找到引用源。**表示节点 v 的时间戳;

步骤 S202：定义节点 u 与节点 v 之间通过任意两个社区产生连接的概率为：

$$p(u,v)=\left(1-\exp(-\sum_{i,j}F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v)$$
$$=\left(1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度；$\mathbf{F_u^T}$表示节点**错误!未找到引用源。**与所有社区的连接强度的向量的转置；$\mathbf{F_v}$表示节点 v 与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;

步骤 S203：针对时序文本网络，根据步骤 S202 定义的公式，生成时序文本网络**错误!未找到引用源。**：

$$G^P=(V\bigcup V_\omega, E\bigcup E_{\omega d}; T\bigcup T_\omega)$$

其中，V、E、T 分别是时序文本网络中的节点集合、边集合以及时间戳集合；**错误!未找到引用源。**代表一个单词；存在于**错误!未找到引用源。**的边(**错误!未找到引用源。**,**错误!未找到引用源。**)代表单词 i 存在于文章 j 中；**错误!未找到引用源。**代表单词的时间戳，被设置成 0；对

于该网络中任意两点，根据 S202 所定义的概率，预测两点间是否有边存在。

上述技术方案中，所述步骤 S3 包括：

步骤 S301：利用块坐标梯度下降法，对于对每个节点 u，假设对 $\forall v \neq u, \mathbf{F}_v$ 不变且 $\boldsymbol{\eta}$ 不变，首先更新 $\mathbf{F}_u$，即 $\hat{\mathbf{F}}_\mathbf{u} = \arg\max_{\mathbf{F}_\mathbf{u} \geq 0} l(\mathbf{F}_\mathbf{u})$，$l(\mathbf{F}_\mathbf{u})$ 为针对 $\mathbf{F}_\mathbf{u}$ 的对数似然函数，具体地：

$$l(\mathbf{F}_\mathbf{u}) = \sum_{v \in inN(u)} log(1-\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})) - \sum_{\substack{v \notin N(u) \\ t(v)<t(u)}} \mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u} + \sum_{v' \in outN(u)} log(1-\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'})) - \sum_{\substack{v' \notin N(u) \\ t(v')>t(u)}} \mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，如图 3 所示，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathbf{v}$ 和 $\mathbf{F}_{\mathbf{v}'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵，$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S302：利用梯度下降法，根据如下公式可以进行对 F 的更新：

$$F_{uk}^{new} \leftarrow \max\{0, F_{uk}^{old} + \alpha_{F_u}(\nabla l(\mathbf{F}_\mathbf{u}))_k\}$$

其中 α_{F_u} 为利用回溯搜索算法计算所得步长；F_{uk}^{new} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新后的值；F_{uk}^{old} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新前的值；$\nabla l(\mathbf{F}_\mathbf{u}))$ 为更新 $\mathbf{F}_\mathbf{u}$ 时所用的梯度，具体的：

$$\nabla l(\mathbf{F}_\mathbf{u})) = \sum_{v \in inN(u)} \frac{\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})}{1-\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_\mathbf{u})}\boldsymbol{\eta}^\mathbf{T}\mathbf{F}_\mathbf{v} - \sum_{\substack{v \notin N(u) \\ t(v)<t(u)}} \boldsymbol{\eta}^\mathbf{T}\mathbf{F}_\mathbf{v} + \sum_{v' \in outN(u)} \frac{\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'})}{1-\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T}\boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'})}\boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'} - \sum_{\substack{v' \notin N(u) \\ t(v')>t(u)}} \boldsymbol{\eta}\mathbf{F}_{\mathbf{v}'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathbf{v}$ 和 $\mathbf{F}_{\mathbf{v}'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵；$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$、$\boldsymbol{\eta}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S303：$\mathbf{F}$ 更新完成后，假设 $\mathbf{F}$ 不变，根据如下公式可以进行对 $\boldsymbol{\eta}$ 的更新：

$$\eta_{ij}^{new} \leftarrow \max\{0, \eta_{ij}^{old} + \alpha_\eta \left(\nabla_\eta l(\mathbf{F}, \boldsymbol{\eta})\right)_{ij}\}$$

其中α_η为利用回溯搜索算法计算所得步长；η_{ij}^{new}为$\boldsymbol{\eta}$矩阵中第 i 行第 j 列更新后的值；η_{ij}^{old}为$\boldsymbol{\eta}$矩阵中第 i 行第 j 列更新前的值；$\nabla_\eta l(\mathbf{F},\boldsymbol{\eta})$为更新$\boldsymbol{\eta}$时所用的梯度，具体的：

$$\nabla_\eta l(\mathbf{F},\boldsymbol{\eta})=\sum_{(u\to v)\in E}\frac{exp(-\mathbf{F}_u^{\mathrm{T}}\boldsymbol{\eta}\mathbf{F}_v)}{1-exp(-\mathbf{F}_u^{\mathrm{T}}\boldsymbol{\eta}\mathbf{F}_v)}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}-\sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$为对应的转置矩阵；$(u\to v)$表示从点 u 指向点 v 的边；t(u)与 t(v)分别表示点 u 与点 v 的时间戳；

步骤 S304：计算$\nabla l(\mathbf{F}_u))$和$\nabla_\eta l(\mathbf{F},\boldsymbol{\eta})$的时间复杂度分别为$O(N)$和$O(N^2)$，为降低时间复杂度、提高可计算性，采取如下近似：

$$\sum_{\substack{v\in N(u)\\ t(v)<t(u)}}h^{\mathrm{T}}\mathbf{F}_v=\sum_{t(v)<t(u)}h^{\mathrm{T}}\mathbf{F}_v-\sum_{v\in inN(u)}h^{\mathrm{T}}\mathbf{F}_v,$$

$$\sum_{\substack{v\in N(u)\\ t(v)>t(u)}}h\mathbf{F}_v=\sum_{t(v)>t(u)}h\mathbf{F}_v-\sum_{v\in outN(u)}h\mathbf{F}_v$$

$$\sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}=\sum_{\substack{(u\to v)\\ t(u)<t(v)}}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}-\sum_{\substack{(u\to v)\in E\\ t(u)<t(v)}}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}.$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合；N(u)表示 inN(u)和 outN(u)的并集，$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$、$\boldsymbol{\eta}^{\mathrm{T}}$为对应的转置矩阵；$(u\to v)$表示从点 u 指向点 v 的边；t(u)、t(v)、t(v’)分别表示点 u、点 v 与点 v’的时间戳；

计算$\nabla l(\mathbf{F}_u))$和$\nabla_\eta l(\mathbf{F},\boldsymbol{\eta})$的时间复杂度分别为$O(|N(u)|)$、$O(|E|)$，总时间复杂度为$O(|E|)$，中$|N(u)|$表示集合$N(u)$包含的节点的个数；$|E|$表示网络中边的条数。

上述技术方案中，所述步骤 S4 包括：

步骤 S401：从数据文件中读取数据，并根据步骤 S1 构建时序文本网络；

步骤 S402：初始化用户与社区间的联系强度矩阵 F；基于向网络中的导率模型，如果节点 u 的入邻居 inN(u)有比所有点 v∈outN(u)的入邻居 inN(v)有更小的导率，则该入邻居 inN(u)在邻近是最小的；对于属于一个在邻近最小的邻域 k 内的节点 u'，初始化节点 u'与一个社区 k 之间的联系强度 Fu'k = 1，否则令 Fu'k = 0；为了初始化 η，设置主对角线上的项为 0.9，其他项为 0.1；

步骤 S403：每轮次根据公式更新 F 与 η，首先针对每个节点 u，根据梯度公式更新节点 u 与所有社区之间的联系强度向量 $\mathbf{F_u}$，梯度公式如下：

$$\nabla l(\mathbf{F_u})) = \sum_{v \in inN(u)} \frac{\exp(-\mathbf{F_v^T \eta F_u})}{1-\exp(-\mathbf{F_v^T \eta F_u})} \boldsymbol{\eta}^{\mathrm{T}} \mathbf{F_v} - \sum_{\substack{v \notin N(u) \\ t(v)<t(u)}} \boldsymbol{\eta}^{\mathrm{T}} \mathbf{F_v} + \sum_{v' \in outN(u)} \frac{\exp(-\mathbf{F_u^T \eta F_{v'}})}{1-\exp(-\mathbf{F_u^T \eta F_{v'}})} \boldsymbol{\eta} \mathbf{F_{v'}} - \sum_{\substack{v' \notin N(u) \\ t(v')>t(u)}} \boldsymbol{\eta} \mathbf{F_{v'}}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F_u^T}$、$\mathbf{F_v^T}$、$\boldsymbol{\eta}^{\mathrm{T}}$为对应的转置矩阵；

F 更新完成后，根据梯度公式更新社区间的联系矩阵 η，梯度公式如下：

$$\nabla_{\eta} l(\mathbf{F}, \boldsymbol{\eta}) = \sum_{(u \to v) \in E} \frac{exp(-\mathbf{F_u^T \eta F_v})}{1-exp(-\mathbf{F_u^T \eta F_v})} \mathbf{F_u F_v^T} - \sum_{\substack{(u \to v) \notin E \\ t(u)<t(v)}} \mathbf{F_u F_v^T}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F_u^T}$、$\mathbf{F_v^T}$为对应的转置矩阵；$(u \to v)$表示从点 u 指向点 v 的边；

t(u)与t(v)分别表示点u与点v的时间戳；

步骤S404：经过一定轮次后，判定每个节点与社区间的隶属关系，针对每个社区k，设定一个阈值δ_k，具体设定方法如下：

$$\delta_k = \sqrt{-\frac{log(1-1/N)}{\eta_{kk}}}$$

其中N为节点总数；η_{kk}为社区间联系矩阵$\boldsymbol{\eta}$第k行第k列的分量，对于节点u与社区k，若联系强度F_{uk}大于社区k的阈值δ_k，则认为节点u隶属于社区k。

上述技术方案中，所述步骤S5包括：

步骤S501：对选定的文本数据集进行训练，提取出节点与社区间联系强度矩阵F，以及社区间的联系关系矩阵$\boldsymbol{\eta}$；

步骤S502：读取矩阵F与矩阵$\boldsymbol{\eta}$；

步骤S503：根据步骤S2所定义的公式计算节点u与节点v之间边的存在概率。

本发明与现有技术相比，具有如下有益效果：

本发明基于时序文本网络中的网络结构信息和文本信息，同时提取出了节点与社区间的隶属关系和社区间的联系关系，弥补了现有技术在分析节点连接原因上的不足；本发明构建了全新的社区检测模型，考虑了社区间的联系关系，同时提供了一种新的文本信息在社区检测中的应用方法，提高了社区检测的效率和准确性。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它附图。

图 1 为本发明的预测方法流程图；

图 2 为本发明构造的适用的时序文本网络的示意图；

图 3 为本发明的出、入邻居示意图；

图 4 为本发明构造的词聚类的词云示意图。

具体实施方式

为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。

根据本发明提供的基于时序文本网络的社区检测与用户关系预测方法，涉及整理含时序文本网络的自动化程序、基于生成模型的新型社区检测方法、新型方法的推断过程和参数估计、社区成员及社区间关系提取、节点间连接预测；具体地，如图 1 所示，包括如下步骤：

步骤 S1：基于原始数据构建时序文本网络；

步骤 S2：针对时序文本网络，构建基于关系图模型的生成模型；

步骤 S3：利用梯度下降法构建生成模型的推断过程；

步骤 S4：根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系，社区指的是表现出较高相关性的点的集合，社区间的关系指的是社区之间的相似度；

步骤 S5：根据提取出的社区信息，进行网络节点间的连接预测并绘制词云。

如图 2 所示，所述步骤 S1 包括：从互联网上获得公开的时序文本数据集，从数据集中抽取出时序文本网络，例如在论文网络中以论文的发表时间作为时序信息、以论文的标题和摘要作为文本信息，在社交网络中以用户推送的短文内容作为文本信息、推送时间座位时序信息，在超链接的网页网络中以网页标题和主要文字作为文本信息、网页更新时间作为时序信息；从数据集中抽取出链接信息，例如在论文网络中以论文

的参考文献作为链接信息，在社交网络中以转发行为作为链接信息，在超链接的网页中以网页的链接作为链接信息；将提取出的信息生成 csv 格式的文件，具体地：

步骤 S101：将顶点集 V 设为空集，将边集 E 设为空集，将图 G 设为 V,E 的集合；

步骤 S102：将原始数据集中的每一篇文章加到顶点集 V 中；

步骤 S103：顶点集 V 中的每一篇文章对应一个标签 T，该标签是指每一篇文章的发表时间；

步骤 S104：将原始数据集中文章间的链接关系加到边集 E 中。

步骤 S105：(V, E; T)的集合构成图 G，即为时序文本网络

所述步骤 S2 包括：对时序文本网络结构中的文本和链接的生成过程进行建模，生成模型是指在已知参数的条件下，假设文章生成过程服从的模型；具体地：

步骤 S201：定义节点 u 与节点 v 之间通过社区 i、j 产生连接的概率：

$$p(u,v,i,j)=\left(1-\exp(-F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点 u 的时间戳;**错误!未找到引用源。**表示节点 v 的时间戳；

步骤 S202：定义节点 u 与节点 v 之间通过任意两个社区产生连接的概率：

$$p(u,v)=\left(1-\exp(-\sum_{i,j}F_{ui}\eta_{ij}F_{vj})\right)\delta(u\to v)$$
$$=\left(1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})\right)\delta(u\to v),$$

$$\delta(u\to v)=\begin{cases}1 & \text{if } t(u)<t(v)\\ 0 & \text{otherwise.}\end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;$\mathbf{F_u^T}$表示节点**错误!未找到引用源。**与所有社区的连接强度的向量的转置;$\mathbf{F_v}$表示节点 v 与所有社区的连接强度的向量;$\boldsymbol{\eta}$表示社区间的相似度的矩阵;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;

步骤 S203：针对时序文本网络 G，根据步骤 S202 定义的公式，生成时序文本网络**错误!未找到引用源。**:

$$G^P=(V\bigcup V_\omega, E\bigcup E_{\omega d}; T\bigcup T_\omega)$$

其中，V，E，T 分别是时序文本网络中的节点集合，边集合以及时间戳集合。**错误!未找到引用源。**代表一个单词；存在于**错误!未找到引用源。**的边(**错误!未找到引用源。**,**错误!未找到引用源。**)代表单词 i 存在于文章 j 中；**错误!未找到引用源。**代表单词的时间戳，被设置成 0；对于该网络中任意两点，根据 S202 所定义的概率，预测两点间是否有边存在。

所述步骤 S3 包括：构建生成模型的推断过程，估计生成模型中的参数，通过已知的文本信息、链接信息和时序信息去推断隐含的参数；本发明采用梯度下降法进行推断，具体地：

步骤 S301：利用块坐标梯度下降法，对于对每个节点 u，假设对

$\forall v \neq u, \mathbf{F}_v$ 不变且 $\boldsymbol{\eta}$ 不变，首先更新 $\mathbf{F}_u$，即 $\hat{\mathbf{F}}_\mathbf{u} = \arg\max_{\mathbf{F}_\mathbf{u} \geq 0} l(\mathbf{F}_\mathbf{u})$。其中 $l(\mathbf{F}_\mathbf{u})$ 为针对 $\mathbf{F}_\mathbf{u}$ 的对数似然函数，具体地：

$$l(\mathbf{F}_\mathbf{u}) = \sum_{v \in inN(u)} log(1 - \exp(-\mathbf{F}_\mathbf{v}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_\mathbf{u})) - \sum_{\substack{v \notin N(u) \\ t(v) < t(u)}} \mathbf{F}_\mathbf{v}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_\mathbf{u} + \sum_{v' \in outN(u)} log(1 - \exp(-\mathbf{F}_\mathbf{u}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_{v'})) - \sum_{\substack{v' \notin N(u) \\ t(v') > t(u)}} \mathbf{F}_\mathbf{u}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_{v'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathrm{v}$ 和 $\mathbf{F}_{v'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵；$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S302：利用梯度下降法，根据如下公式可以进行对 F 的更新：

$$F_{uk}^{new} \leftarrow \max\{0, F_{uk}^{old} + \alpha_{F_u} (\nabla l(\mathbf{F}_\mathbf{u}))_k\}$$

其中 α_{F_u} 为利用回溯搜索算法计算所得步长；F_{uk}^{new} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新后的值；F_{uk}^{old} 为 $\mathbf{F}_\mathbf{u}$ 向量第 k 个分量更新前的值；$\nabla l(\mathbf{F}_\mathbf{u}))$ 为更新 $\mathbf{F}_\mathbf{u}$ 时所用的梯度，具体的：

$$\nabla l(\mathbf{F}_\mathbf{u})) = \sum_{v \in inN(u)} \frac{\exp(-\mathbf{F}_\mathbf{v}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_\mathbf{u})}{1 - \exp(-\mathbf{F}_\mathbf{v}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_\mathbf{u})} \boldsymbol{\eta}^\mathbf{T} \mathbf{F}_\mathbf{v} - \sum_{\substack{v \notin N(u) \\ t(v) < t(u)}} \boldsymbol{\eta}^\mathbf{T} \mathbf{F}_\mathbf{v} + \sum_{v' \in outN(u)} \frac{\exp(-\mathbf{F}_\mathbf{u}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_{v'})}{1 - \exp(-\mathbf{F}_\mathbf{u}^\mathbf{T} \boldsymbol{\eta} \mathbf{F}_{v'})} \boldsymbol{\eta} \mathbf{F}_{v'} - \sum_{\substack{v' \notin N(u) \\ t(v') > t(u)}} \boldsymbol{\eta} \mathbf{F}_{v'}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F}_\mathbf{u}$、$\mathbf{F}_\mathrm{v}$ 和 $\mathbf{F}_{v'}$ 分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵；$\mathbf{F}_\mathbf{u}^\mathbf{T}$、$\mathbf{F}_\mathbf{v}^\mathbf{T}$、$\boldsymbol{\eta}^\mathbf{T}$ 为对应的转置矩阵；

步骤 S303：$\mathbf{F}$ 更新完成后，假设 $\mathbf{F}$ 不变，根据如下公式可以进行对 $\boldsymbol{\eta}$ 的更新：

$$\eta_{ij}^{new} \leftarrow \max\{0, \eta_{ij}^{old} + \alpha_\eta \left(\nabla_\eta l(\mathbf{F}, \boldsymbol{\eta})\right)_{ij}\}$$

其中 α_η 为利用回溯搜索算法计算所得步长；η_{ij}^{new} 为 $\boldsymbol{\eta}$ 矩阵中第 i 行第 j 列更新后的值；η_{ij}^{old} 为 $\boldsymbol{\eta}$ 矩阵中第 i 行第 j 列更新前的值；$\nabla_\eta l(\mathbf{F}, \boldsymbol{\eta})$ 为更新 $\boldsymbol{\eta}$ 时所用的梯度，具体的：

$$\nabla_{\eta} l(\mathbf{F},\boldsymbol{\eta}) = \sum_{(u\to v)\in E} \frac{exp(-\mathbf{F}_u^{\mathrm{T}}\boldsymbol{\eta}\mathbf{F}_v)}{1-exp(-\mathbf{F}_u^{\mathrm{T}}\boldsymbol{\eta}\mathbf{F}_v)}\mathbf{F}_u\mathbf{F}_v^{\mathrm{T}} - \sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}} \mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$为对应的转置矩阵；$(u\to v)$表示从点 u 指向点 v 的边；t(u)与 t(v)分别表示点 u 与点 v 的时间戳；

步骤 S304：计算$\nabla l(\mathbf{F}_u))$和$\nabla_{\eta} l(\mathbf{F},\boldsymbol{\eta})$的时间复杂度分别为 $O(N)$和$O(N^2)$，为降低时间复杂度、提高可计算性，采取如下近似：

$$\sum_{\substack{v\in N(u)\\ t(v)<t(u)}} h^{\mathrm{T}}\mathbf{F}_v = \sum_{t(v)<t(u)} h^{\mathrm{T}}\mathbf{F}_v - \sum_{v\in inN(u)} h^{\mathrm{T}}\mathbf{F}_v,$$

$$\sum_{\substack{v\in N(u)\\ t(v)>t(u)}} h\mathbf{F}_v = \sum_{t(v)>t(u)} h\mathbf{F}_v - \sum_{v\in outN(u)} h\mathbf{F}_v,$$

$$\sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}} \mathbf{F}_u\mathbf{F}_v^{\mathrm{T}} = \sum_{\substack{(u\to v)\\ t(u)<t(v)}} \mathbf{F}_u\mathbf{F}_v^{\mathrm{T}} - \sum_{\substack{(u\to v)\in E\\ t(u)<t(v)}} \mathbf{F}_u\mathbf{F}_v^{\mathrm{T}}.$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合；N(u)表示 inN(u)和 outN(u)的并集，$\mathbf{F}_u$、$\mathbf{F}_v$和$\mathbf{F}_{v'}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F}_u^{\mathrm{T}}$、$\mathbf{F}_v^{\mathrm{T}}$、$\boldsymbol{\eta}^{\mathrm{T}}$为对应的转置矩阵；$(u\to v)$表示从点 u 指向点 v 的边；t(u)、t(v)、t(v’)分别表示点 u、点 v 与点 v’的时间戳；

这样，计算$\nabla l(\mathbf{F}_u))$和$\nabla_{\eta} l(\mathbf{F},\boldsymbol{\eta})$的时间复杂度分别为$O(|N(u)|)$、$O(|E|)$，总时间复杂度为$O(|E|)$。其中$|N(u)|$表示集合$N(u)$包含的节点的个数；$|E|$表示网络中边的条数；

所述步骤 S4 包括：训练时序文本网络，并根据得到的参数来计算节点与社区间的隶属关系以及社区间的联系关系，计算得到的关系强度用于步骤 S5 中的节点连接预测，具体地：

步骤 S401：从数据文件中读取数据，并根据步骤 1 构建时序文本网

络；

步骤 S402：初始化用户与社区间的联系强度矩阵 F。基于有向网络中的导率模型，如果节点 u 的入邻居 inN(u)有比所有点 v ∈ outN(u)的入邻居 inN(v)有更小的导率，则该入邻居 inN(u)在邻近是最小的。对于属于这样一个在邻近最小的邻域 k 内的节点 u'，初始化节点 u'与一个社区 k 之间的联系强度 Fu'k = 1，否则令 Fu'k = 0。为了初始化 η，设置主对角线上的项为 0.9，其他项为 0.1；

步骤 S403：每轮次根据公式更新 F 与 η，首先针对每个节点 u，根据梯度公式更新节点 u 与所有社区之间的联系强度向量$\mathbf{F_u}$，梯度公式如下：

$$\nabla l(\mathbf{F_u})) = \sum_{v\in inN(u)} \frac{\exp(-\mathbf{F_v^T}\boldsymbol{\eta}\mathbf{F_u})}{1-\exp(-\mathbf{F_v^T}\boldsymbol{\eta}\mathbf{F_u})}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{F_v} - \sum_{\substack{v\notin N(u)\\ t(v)<t(u)}} \boldsymbol{\eta}^{\mathrm{T}}\mathbf{F_v} + \sum_{v'\in outN(u)} \frac{\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_{v'}})}{1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_{v'}})}\boldsymbol{\eta}\mathbf{F_{v'}} - \sum_{\substack{v'\notin N(u)\\ t(v')>t(u)}} \boldsymbol{\eta}\mathbf{F_{v'}}$$

其中 inN(u)和 outN(u)表示进入 u 节点和从 u 节点发出的节点的集合，$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量，$\boldsymbol{\eta}$表示社区间的相似度的矩阵。$\mathbf{F_u^T}$、$\mathbf{F_v^T}$、$\boldsymbol{\eta}^{\mathbf{T}}$为对应的转置矩阵。

F 更新完成后，根据梯度公式更新社区间的联系矩阵$\boldsymbol{\eta}$，梯度公式如下：

$$\nabla_{\eta} l(\mathbf{F},\boldsymbol{\eta}) = \sum_{(u\to v)\in E} \frac{exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})}{1-exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})}\mathbf{F_u}\mathbf{F_v^T} - \sum_{\substack{(u\to v)\notin E\\ t(u)<t(v)}} \mathbf{F_u}\mathbf{F_v^T}.$$

其中 E 表示时序文本网络中所有边的集合；$\mathbf{F_u}$、$\mathbf{F_v}$和$\mathbf{F_{v'}}$分别表示 u 节点、v 节点和 v’节点与所有社区的连接强度的向量；$\boldsymbol{\eta}$表示社区间的相似度的矩阵；$\mathbf{F_u^T}$、$\mathbf{F_v^T}$为对应的转置矩阵；$(u \to v)$表示从点 u 指向点 v 的边；t(u)与 t(v)分别表示点 u 与点 v 的时间戳；

步骤 S4.4：经过一定轮次后，判定每个节点与社区间的隶属关系。针对每个社区 k，设定一个阈值δ_k，具体设定方法如下：

$$\delta_k = \sqrt{-\frac{log(1-1/N)}{\eta_{kk}}}$$

其中 N 为节点总数；η_{kk} 为社区间联系矩阵 $\boldsymbol{\eta}$ 第 k 行第 k 列的分量。对于节点 u 与社区 k，若联系强度 F_{uk} 大于社区 k 的阈值 δ_k，则认为节点 u 隶属于社区 k。

所述步骤 S5 包括：根据前述步骤提取出的节点与社区间的隶属关系以及社区间的联系关系进行节点连接预测并绘制词云，如图 4 所示，具体地：

步骤 S501：对选定的文本数据集进行训练，提取出节点与社区间联系强度矩阵 F，以及社区间的联系关系矩阵 $\boldsymbol{\eta}$；

步骤 S502：读取矩阵 F 与矩阵 $\boldsymbol{\eta}$；

步骤 S503：根据步骤 S2 所定义的公式计算节点 u 与节点 v 之间边的存在概率。具体公式如下：

$$\begin{aligned} p(u,v) &= \left(1-\exp(-\sum_{i,j} F_{ui}\eta_{ij}F_{vj})\right)\delta(u \to v) \\ &= \left(1-\exp(-\mathbf{F_u^T}\boldsymbol{\eta}\mathbf{F_v})\right)\delta(u \to v), \end{aligned}$$

$$\delta(u \to v) = \begin{cases} 1 & \text{if } t(u) < t(v) \\ 0 & \text{otherwise.} \end{cases}$$

其中**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示节点**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度;**错误!未找到引用源。**表示社区**错误!未找到引用源。**与社区**错误!未找到引用源。**的连接强度；$\mathbf{F_u^T}$ 表示节点**错误!未找到引用源。**与所有社区的连接强度的向量的转置；$\mathbf{F_v}$ 表示节点 v 与所有社区的连接强度的向量；$\boldsymbol{\eta}$ 表示社区间的相似度的矩阵;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳;**错误!未找到引用源。**表示节点**错误!未找到引用源。**的时间戳；

步骤 S504：对任意一个社区 k,找出隶属于这个社区的所有词以及每

说　明　书　摘　要

本发明公开了一种基于时序文本网络的社区检测与用户关系预测方法，包括：基于原始数据构建时序文本网络；针对时序文本网络，构建基于关系图模型的生成模型；利用梯度下降法构建生成模型的推断过程；根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系；根据提取出的社区信息，进行网络节点间的连接预测。本发明构建了全新的社区检测方法，并提出了社区相关度的概念，大幅提升了社区检测的准确性和解释性。

的文本内容。 第三，基于实证观察，本发明提出了一个可以利用时序文本网络中所有的信息来源并可以囊括数以百万计的网络节点的概率生成模型。

以上对本发明的具体实施例进行了描述。需要理解的是，本发明并不局限于上述特定实施方式，本领域技术人员可以在权利要求的范围内做出各种变化或修改，这并不影响本发明的实质内容。在不冲突的情况下，本申请的实施例和实施例中的特征可以任意相互组合。

说　明　书　附　图

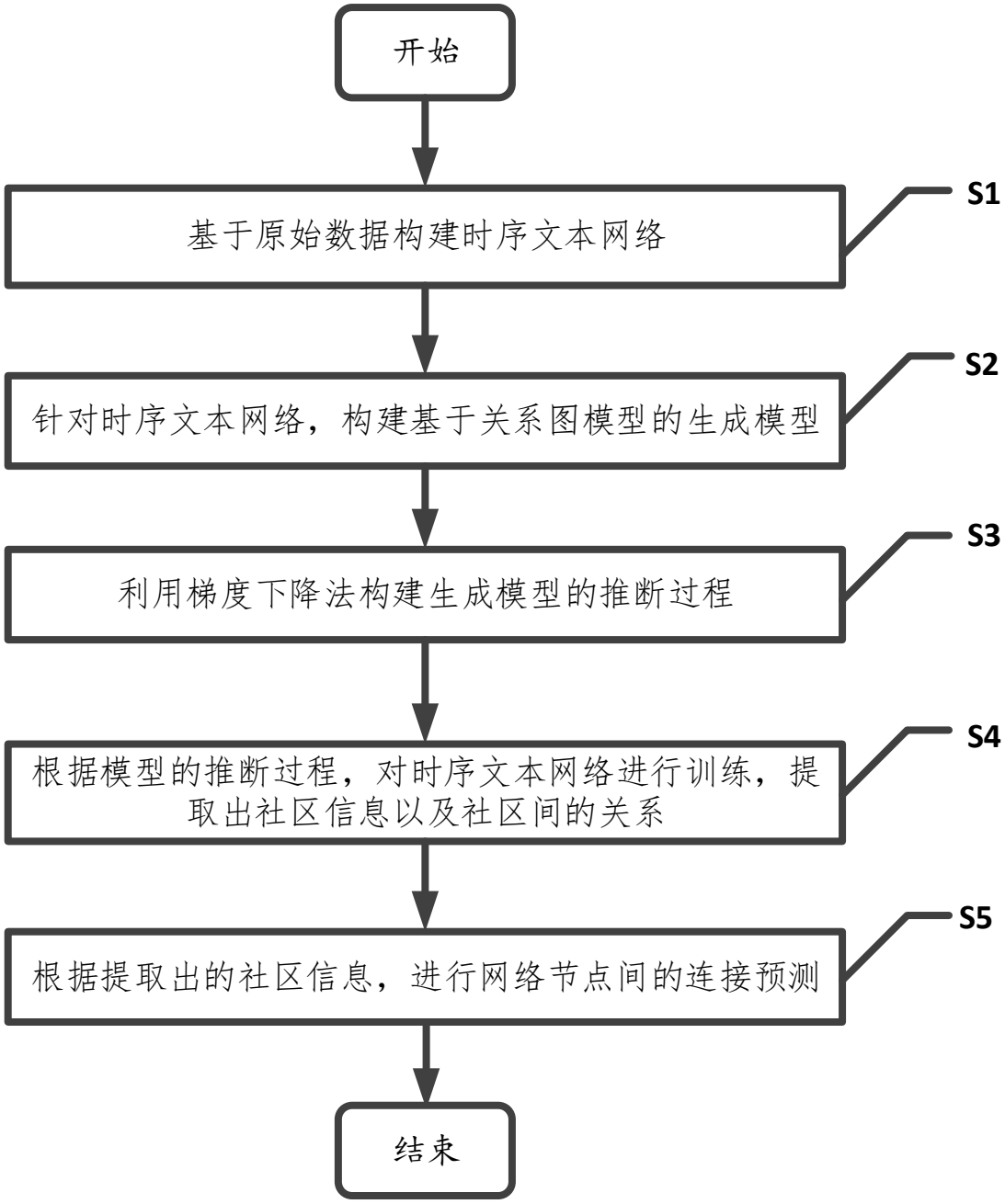

开始
基于原始数据构建时序文本网络
S1
针对时序文本网络，构建基于关系图模型的生成模型
S2
利用梯度下降法构建生成模型的推断过程
S3
根据模型的推断过程，对时序文本网络进行训练，提取出社区信息以及社区间的关系
S4
根据提取出的社区信息，进行网络节点间的连接预测
S5
结束

图 1

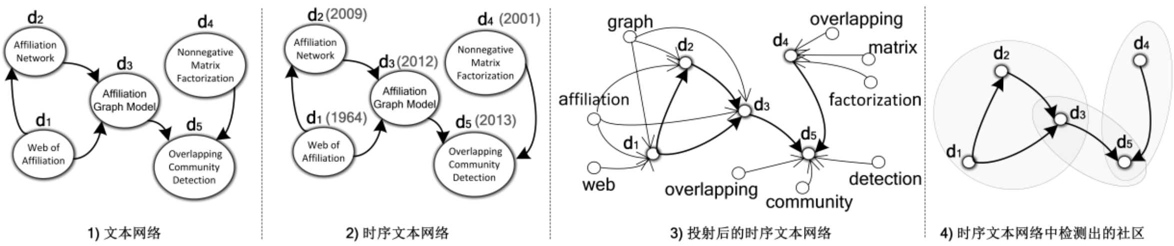

d2
Affiliation Network
d3
Affiliation Graph Model
d4
Nonnegative Matrix Factorization
d1
Web of Affiliation
d5
Overlapping Community Detection
1) 文本网络
d2 (2009)
d3 (2012)
d4 (2001)
d1 (1964)
d5 (2013)
2) 时序文本网络
graph
overlapping
matrix
affiliation
factorization
web
overlapping
community
detection
3) 投射后的时序文本网络
4) 时序文本网络中检测出的社区

图 2

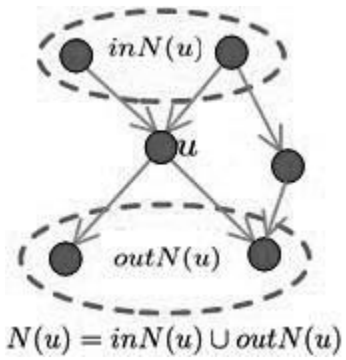

inN(u)
u
outN(u)
N(u) = inN(u) ∪ outN(u)

图 3


risk
patient
associate
disease
control
studies
coronaritis
arterial
mortal
bypass
graft
diabet
liver
fatti
oral
cardiovascular
outcome
population
serum
stent
increass
pump
cardiac
women
chronic
earlier
elev
surgeries
obesity
adult
marker
sever
left
adverse
prevail
independ
care
predictor
morbid
venous
cirrhosis
thrombosis
nation
vein
year
late
cohort
alcohol

图 4