发明名称：一种基于多知识库的表格实体链接方法

申请人：　东南大学

发明人：　吴天星　漆桂林　刘太云　严晟嘉　朴智新　许亮　王瑞明

（东南大学计算机学院、软件学院）

第一发明人：吴天星，身份证号：**321102199006181514**

联系人：吴天星

联系电话：**15077889931**，办公室电话：**025-52090910**

E-mail：wutianxing618@163.com

漆桂林：**13376069256　025-52090910　　gqi@seu.edu.cn**

代理人：　　刘琦

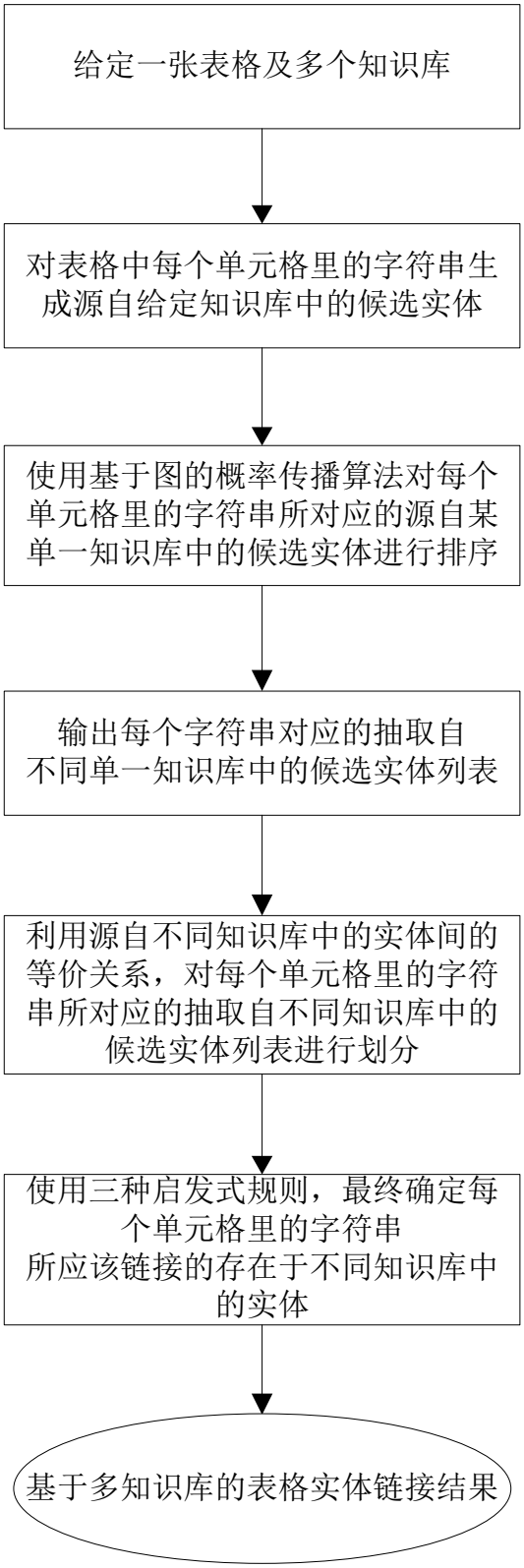
电话：　　**13951853948**

E-mail:　　**liuqisylawyer@163.com**

说 明 书 摘 要

本发明公开了一种基于多知识库的表格实体链接方法，主要用于解决表格中的实体链接问题。本发明首先对于给定表格中每个单元格里的字符串生成候选实体，这些候选实体抽取自给定的知识库。然后，提出一种通用的基于图的概率传播算法对每个单元格里的字符串对应的候选实体进行排序，该方法可作用于任意单一的知识库。再根据基于不同单一知识库的候选实体排序的结果，利用源自不同知识库中的实体间的等价关系，对每个单元格里的字符串所对应的抽取自不同知识库中的已排序候选实体进行划分。最后使用三种启发式规则以最终确定每个单元格里的字符串所应该链接的存在于不同知识库中的实体，从而完成基于多知识库的表格实体链接的任务。

摘 要 附 图


给定一张表格及多个知识库
对表格中每个单元格里的字符串生成源自给定知识库中的候选实体
使用基于图的概率传播算法对每个单元格里的字符串所对应的源自某单一知识库中的候选实体进行排序
输出每个字符串对应的抽取自不同单一知识库中的候选实体列表
利用源自不同知识库中的实体间的等价关系，对每个单元格里的字符串所对应的抽取自不同知识库中的候选实体列表进行划分
使用三种启发式规则，最终确定每个单元格里的字符串所应该链接的存在于不同知识库中的实体
基于多知识库的表格实体链接结果

权 利 要 求 书

1．一种基于多知识库的表格实体链接方法，其特征在于，该方法包括如下步骤：

1）每次从知识库集合 **K**={KB_1,KB_2,…,KB_z…,KB_n}中选定一个单一知识库 KB_z，按照如下方法从所述单一知识库 KB_z 中抽取候选实体，构建候选实体列表，最终得到每个单一知识库构建的候选实体列表：

利用同义词典 BabelNet 与字符串匹配规则，将表格 T 中所有单元格里的字符串 s 生成源自单一知识库 KB_z 的候选实体，每个字符串 s 对应多个候选实体；

利用基于图的概率传播算法对表格 T 中每个字符串 s 所对应的候选实体进行排序，得到候选实体列表；

2）将每个字符串 s 所对应的 n 个候选实体列表中的所有实体划分成多个实体集合，这些实体集合可分为两类：第一类中的每个集合里的实体数量 num∈{2,3,…,n}，每个集合中的实体分别源自不同的候选实体列表，且这些实体两两之间均存在等价关系；第二类中的每个集合中的实体数量均为 1，每个集合中的实体仅源自一个候选实体列表且与源自其他候选实体列表中的每个实体之间均不存在等价关系；

3）针对每个字符串所对应的多个不同的实体集合，使用三种启发式规则为每个字符串 s 选择一个实体集合中的所有实体作为该字符串 s 所应该链接的存在于不同知识库中的实体，从而完成表格实体链接。

2．根据权利要求 1 所述的基于多知识库的表格实体链接方法，其特征在于，所述步骤 1）中，按照如下方式生成源自单一知识库 KB_z 的候选实体：

1-a）为单一知识库 KB_z 中的每个实体查找其在同义词典 BabelNet 中的所有同义词，并构建每个实体对应的同义词集合；

1-b）对每个字符串 s 进行分词，得到零散片段{$w_1(s)$,$w_2(s)$,…,$w_v(s)$,…,$w_k(s)$}，其中 $w_v(s)$表示对字符串 s 分词后的第 v 个片段，v∈{1,2,…,k}，k 为对字符串 s 分词后得到零散片段的总数量；

1-c）使用字符串匹配规则为表格 T 中所有单元格里的字符串生成候选实体，该规则为：如果知识库 KB_z 中的某个实体 e 及 e 的某个同义词包含经过分词后的字符串 s 的某个零散片段 $w_v(s)$，则将该实体 e 作为字符串 s 的一个候选实体。

3．根据权利要求 1 所述的基于多知识库的表格实体链接方法，其特征在于，所述步骤 1）中对表格 T 中每个字符串 s 所对应的候选实体进行排序的具体流程为：

1-1）按照如下方式为表格 T 构建实体消岐图 G：将表格 T 中每个单元格里的字符串作为一个字符串节点，将每个候选实体作为一个实体节点，将字符串——实体边

作为一条存在于每个字符串与其对应的一个候选实体之间的无向边，将一条实体——实体边作为一条存在于 G 中任意两个实体节点之间的无向边；

1-2）计算所述实体消岐图 G 中每个字符串与其对应的每个候选实体之间的字符串——实体语言学相似度、字符串——实体上下文相似度，并根据这两种相似度计算每条字符串——实体边的权重；

1-3）计算实体消岐图 G 中任意实体之间的实体——实体三元组相似度与实体——实体上下文相似度，并根据这两种相似度计算每条实体——实体边的权重；

1-4）利用如下公式进行迭代概率传播，直至向量 **R** 收敛：

$$\mathbf{R} = ((1-\mathrm{b}) \times \frac{\mathbf{E}}{\mathrm{m}} + \mathrm{b} \times \mathbf{A}) \times \mathbf{R}$$

其中 m 为所构建的实体消岐图 G 中节点的总量，**E** 是一个 m×m 的全 1 矩阵，b 是一个接近 1 的常数，b∈[0.8,1)，**R** 是一个 m×1 的向量<r_1,r_2,…,r_m>，r_j 为 G 中第 j 个节点所关联到的概率值，j∈{1,2,…,m}；R 的初始值计算方式如下：若第 j 个节点为字符串节点，则 r_j=1/m，它表示该字符串节点的重要度；若第 j 个节点为实体节点，则 r_j=0，它表示该一字符串链接到该实体的概率值；A 是一个 m×m 邻接矩阵，表示方式如下：

$$\mathbf{A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} & \cdots & \mathrm{A}_{1\mathrm{m}} \\ \mathrm{A}_{21} & \mathrm{A}_{22} & \cdots & \mathrm{A}_{2\mathrm{m}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{A}_{\mathrm{m}1} & \mathrm{A}_{\mathrm{m}2} & \cdots & \mathrm{A}_{\mathrm{mm}} \end{bmatrix}$$

其中 A_{xy} 表示从实体消岐图 G 中的第 x 个节点到第 y 个节点的转移概率，x∈{1,2,…,m}，y∈{1,2,…,m}，A_{xy} 的定义如下：

$$\mathrm{A}_{\mathrm{xy}} = \begin{cases} \dfrac{\mathrm{W}_{\mathrm{se}}(\mathrm{x},\mathrm{y})}{\mathrm{W}_{\mathrm{se}}(\mathrm{x},*)}，若\ \mathrm{x} \neq \mathrm{y}\ 且\ \mathrm{x}\ 是字符串节点 \\ \mathrm{a} \times \dfrac{\mathrm{W}_{\mathrm{ee}}(\mathrm{x},\mathrm{y})}{\mathrm{W}_{\mathrm{ee}}(\mathrm{x},*)}，若\ \mathrm{x} \neq \mathrm{y}\ 且\ \mathrm{x}\ 与\ \mathrm{y}\ 均是实体节点 \\ (1-\mathrm{a}) \times \dfrac{\mathrm{W}_{\mathrm{se}}(\mathrm{y},\mathrm{x})}{\mathrm{W}_{\mathrm{se}}(*,\mathrm{x})}，若\ \mathrm{x} \neq \mathrm{y}\ 且\ \mathrm{x}\ 是实体节点以及\ \mathrm{y}\ 是字符串节点 \\ 0，若\ \mathrm{x} = \mathrm{y} \end{cases}$$

其中 W_{se}(x,y)表示字符串节点 x 与实体节点 y 之间的字符串——实体边权重，W_{se}(y,x)表示字符串节点 y 与实体节点 x 之间的字符串——实体边权重，W_{se}(x,*)表示字符串节点 x 与其相邻的每个实体节点之间的字符串——实体边权重的总和，W_{se}(*,x)表示实体节点 x 与其相邻的每个字符串节点之间的字符串——实体边权重的总和，

$W_{ee}(x,y)$表示实体节点 x、y 之间的实体——实体边权重，$W_{ee}(x,*)$表示实体节点 x 与其相邻的每个实体节点之间的实体——实体边权重的总和，a 是一个常数，$a\in(0,1)$；

1-5）所述向量 **R** 收敛后，根据候选实体所在的实体节点所关联的概率值，对字符串 s 对应的候选实体进行降序排列，从而得到候选实体列表。

4．根据权利要求 1、2 或 3 所述的基于多知识库的表格实体链接方法，其特征在于，所述步骤 3）中的三种启发式规则分别为：

规则一：如果在字符串 s 对应的多个实体集合中，存在一个集合 Set，与其他实体集合相比，Set 中所有实体在各自对应的候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，且集合 Set 中实体的数量 num 不小于所有给定知识库的数量的一半，则选择集合 Set 中的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则二：如果在字符串 s 对应的多个实体集合中，存在 g 个集合，g>1，这 g 个集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 相等，最高值 hr 也相等，且与其他实体集合相比，这 g 个集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，此外这 g 个集合中每个集合里实体的数量均不小于所有给定知识库的数量的一半，则随机选择这 g 个集合中的一个集合里的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则三：如果在字符串 s 对应的每个实体集合中实体的数量均小于所有给定知识库的数量的一半，则取出在字符串 s 所对应的 n 个候选实体列表，将每个列表中排名第一的实体作为 s 所应该链接的存在于不同知识库中的实体。

说 明 书

一种基于多知识库的表格实体链接方法

技术领域

本发明属于实体链接领域，涉及一种基于多知识库的表格实体链接方法。

背景技术

当前的万维网中存在大量的拥有高质量关系型数据的 HTML 表格，这些表格被视为从万维网中进行知识抽取的重要来源。为了实现语义万维网的愿景，许多工作尝试挖掘表格中潜在的语义信息，将给定表格中的内容表示成 RDF 三元组。对表格内容进行语义信息挖掘的首要步骤即为实体链接，实体链接是识别表格中每个单元格里的字符串的真正含义，并将这些字符串分别链接向给定知识库中的实体。如果不能正确识别表格中潜在的实体，那么将很难从给定表格的内容中挖掘出正确的 RDF 三元组，所以对表格进行实体链接是具有非常大的研究意义与实用价值的工作。

近年来，国内外研究人员为了解决表格实体链接的问题，提出了许多相关系不同的方法，并研制出若干实用系统，包括 Mulwad 等人提出的基于语义信息传递的方法，英国谢菲尔德大学研制的 TableMiner 系统，清华大学研制的 LIEGE 系统以及美国西北大学研制的 TabEL 系统等。但是目前现有的表格实体链接的方法与系统存在两个主要的问题：1）许多方法或系统依赖于基于特定信息的特征，比如列标题与知识库中的实体类型，但是大多数抽取自万维网中的表格均没有列标题，同时许多知识库也没有实体类型这样的语义信息，这导致这些方法与系统并不通用，实用性较差；2）所有目前方法与系统均是针对单一知识库进行表格实体链接，但是这并不能保证表格实体链接的质量，很多表格中的实体并不存在于某一单一知识库中，那么仅针对单一知识库进行实体链接是不合理的。

LIEGE 系统首先对维基百科站点的实体页面，重定向页面，去歧义页面以及超链接信息进行了统计，得到一个关于表格中单元格里的字符串和知识库实体的词典。然后从词典中为字符串生成候选实体集合，最后利用一种迭代联合消歧算法完成实体链接。但是 LIEGE 系统仅能对列表型表格（一列多行）进行基于任意单一知识库的实体链接，大大减弱了该系统的实用性。

TabEL 系统首先利用统计万维网与维基百科中所有实体的相关信息，然后得到相

应的先验概率，并依照此概率为给定表格中每个单元格里的字符串生成候选实体，之后定义了多种不同的特征，最后综合这些特征值，使用一种基于最大似然概率的联合实体消岐方法，进行表格实体链接。TabEL 比 LIEGE 更加先进，原因是 TabEL 能够对多行多列的表格进行基于任意单一知识库的实体链接，但是该系统还是不能完成基于多知识库的表格实体链接的任务，由于许多字符串所应该链接的实体不存在于给定的单一知识库中，导致使用 TabEL 系统进行表格实体链接的质量依旧不能令人满意。此外，该系统依赖于不同来源计算得到的先验概率，而每个来源本身就是有所侧重，导致获取的先验概率并不客观，容易影响表格实体链接的质量。

发明内容

技术问题：本发明提供一种对于给定的一张表格以及任意多个知识库，能够自动化地确定表格中每个单元格里的字符串所应该链接的存在于多个不同知识库中的实体的基于多知识库的表格实体链接方法。

技术方案：本发明的基于多知识库的表格实体链接方法，首先通过一种利用同义词典 BabelNet 与字符串匹配规则的方法，为表格中每个单元格里的字符串生成抽取自给定知识库中的候选实体，然后设计一种通用的且不依赖于任何特定信息的基于图的概率传播算法，对每个单元格里的字符串对应的抽取自不同知识库的候选实体进行排序，之后利用源自不同知识库中的实体间的等价关系对每个字符串所对应的抽取自不同知识库中的已排序候选实体进行划分，最后使用三种启发式规则确定每个字符串所应该链接的存在于不同知识库中的实体。

本发明的基于多知识库的表格实体链接方法，包括如下步骤：

1）每次从知识库集合 $\mathbf{K}=\{KB_1,KB_2,\ldots,KB_z\ldots,KB_n\}$中选定一个单一知识库 KB_z，按照如下方法从所述单一知识库 KB_z 中抽取候选实体，构建候选实体列表，最终得到每个单一知识库构建的候选实体列表：

利用同义词典 BabelNet 与字符串匹配规则，将表格 T 中所有单元格里的字符串 s 生成源自单一知识库 KB_z 的候选实体，每个字符串 s 对应多个候选实体；

利用基于图的概率传播算法对表格 T 中每个字符串 s 所对应的候选实体进行排序，得到候选实体列表；

2）将每个字符串 s 所对应的 n 个候选实体列表中的所有实体划分成多个实体集合，这些实体集合可分为两类：第一类中的每个集合里的实体数量 $num\in\{2,3,\ldots,n\}$，

每个集合中的实体分别源自不同的候选实体列表，且这些实体两两之间均存在等价关系；第二类中的每个集合中的实体数量均为 1，每个集合中的实体仅源自一个候选实体列表且与源自其他候选实体列表中的每个实体之间均不存在等价关系；

3）针对每个字符串所对应的多个不同的实体集合，使用三种启发式规则为每个字符串 s 选择一个实体集合中的所有实体作为该字符串 s 所应该链接的存在于不同知识库中的实体，从而完成表格实体链接。

本发明方法的优选方案中，所述步骤 1）中，按照如下方式生成源自单一知识库 KB_z 的候选实体：

1-a）为单一知识库 KB_z 中的每个实体查找其在同义词典 BabelNet 中的所有同义词，并构建每个实体对应的同义词集合；

1-b）对每个字符串 s 进行分词，得到零散片段$\{w_1(s),w_2(s),\ldots,w_v(s),\ldots,w_k(s)\}$，其中 $w_v(s)$表示对字符串 s 分词后的第 v 个片段，$v\in\{1,2,\ldots,k\}$，k 为对字符串 s 分词后得到零散片段的总数量；

1-c）使用字符串匹配规则为表格 T 中所有单元格里的字符串生成候选实体，该规则为：如果知识库 KB_z 中的某个实体 e 及 e 的某个同义词包含经过分词后的字符串 s 的某个零散片段 $w_v(s)$，则将该实体 e 作为字符串 s 的一个候选实体。

本发明方法的优选方案中，所述步骤 1）中对表格 T 中每个字符串 s 所对应的候选实体进行排序的具体流程为：

1-1）按照如下方式为表格 T 构建实体消岐图 G：将表格 T 中每个单元格里的字符串作为一个字符串节点，将每个候选实体作为一个实体节点，将字符串——实体边作为一条存在于每个字符串与其对应的一个候选实体之间的无向边，将一条实体——实体边作为一条存在于 G 中任意两个实体节点之间的无向边；

1-2）计算所述实体消岐图 G 中每个字符串与其对应的每个候选实体之间的字符串——实体语言学相似度、字符串——实体上下文相似度，并根据这两种相似度计算每条字符串——实体边的权重；

1-3）计算实体消岐图 G 中任意实体之间的实体——实体三元组相似度与实体——实体上下文相似度，并根据这两种相似度计算每条实体——实体边的权重；

1-4）利用如下公式进行迭代概率传播，直至向量 **R** 收敛：

$$\mathbf{R}=((1-\mathrm{b})\times\frac{\mathbf{E}}{\mathrm{m}}+\mathrm{b}\times\mathbf{A})\times\mathbf{R}$$

其中 m 为所构建的实体消岐图 G 中节点的总量，**E** 是一个 m×m 的全 1 矩阵，b 是一个接近 1 的常数，b∈[0.8,1)，**R** 是一个 m×1 的向量<r_1,r_2,…,r_m>，r_j 为 G 中第 j 个节点所关联到的概率值，j∈{1,2,…,m}；R 的初始值计算方式如下：若第 j 个节点为字符串节点，则 r_j=1/m，它表示该字符串节点的重要度；若第 j 个节点为实体节点，则 r_j=0，它表示该一字符串链接到该实体的概率值；A 是一个 m×m 邻接矩阵，表示方式如下：

$$\mathbf{A}=\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix}$$

其中 A_{xy} 表示从实体消岐图 G 中的第 x 个节点到第 y 个节点的转移概率，x∈{1,2,…,m}，y∈{1,2,…,m}，A_{xy} 的定义如下：

$$A_{xy}=\begin{cases} \dfrac{W_{se}(x,y)}{W_{se}(x,*)}，若\ x\neq y\ 且\ x\ 是字符串节点 \\ a\times\dfrac{W_{ee}(x,y)}{W_{ee}(x,*)}，若\ x\neq y\ 且\ x\ 与\ y\ 均是实体节点 \\ (1-a)\times\dfrac{W_{se}(y,x)}{W_{se}(*,x)}，若\ x\neq y\ 且\ x\ 是实体节点以及\ y\ 是字符串节点 \\ 0，若\ x=y \end{cases}$$

其中 W_{se}(x,y)表示字符串节点 x 与实体节点 y 之间的字符串——实体边权重，W_{se}(y,x)表示字符串节点 y 与实体节点 x 之间的字符串——实体边权重，W_{se}(x,*)表示字符串节点 x 与其相邻的每个实体节点之间的字符串——实体边权重的总和，W_{se}(*,x)表示实体节点 x 与其相邻的每个字符串节点之间的字符串——实体边权重的总和，W_{ee}(x,y)表示实体节点 x、y 之间的实体——实体边权重，W_{ee}(x,*)表示实体节点 x 与其相邻的每个实体节点之间的实体——实体边权重的总和，a 是一个常数，a∈(0,1)；

1-5）所述向量 **R** 收敛后，根据候选实体所在的实体节点所关联的概率值，对字符串 s 对应的候选实体进行降序排列，从而得到候选实体列表。

本发明方法的优选方案中，所述步骤 3）中的三种启发式规则分别为：

规则一：如果在字符串 s 对应的多个实体集合中，存在一个集合 Set，与其他实体集合相比，Set 中所有实体在各自对应的候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，且集合 Set 中实体的数量 num 不小于所有给定知识库的数量的一半，则选择集合 Set 中的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则二：如果在字符串 s 对应的多个实体集合中，存在 g 个集合，g>1，这 g 个

集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 相等，最高值 hr 也相等，且与其他实体集合相比，这 g 个集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，此外这 g 个集合中每个集合里实体的数量均不小于所有给定知识库的数量的一半，则随机选择这 g 个集合中的一个集合里的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则三：如果在字符串 s 对应的每个实体集合中实体的数量均小于所有给定知识库的数量的一半，则取出在字符串 s 所对应的 n 个候选实体列表，将每个列表中排名第一的实体作为 s 所应该链接的存在于不同知识库中的实体。

本发明提出的基于多知识库的表格实体链接方法，不依赖于任何特定信息且可以利用任意多种不同的知识库进行表格实体链接，很好地克服了现有方法或系统的弱点，在实体链接的质量上也有了较大的提升。

有益效果：本发明与现有技术相比，具有以下优点：

相比 Mulwad 等人提出的基于语义信息传递的方法，本发明不依赖于列标题等特定信息对给定表格进行整体建模，从而完成基于多知识库的表格实体链接任务，实用性更强，针对不管是否存在列标题的表格均能进行实体链接。此外，本发明对于外部信息的依赖更小，仅需使用任意给定知识库中的 RDF 三元组计算字符串与实体之间的相似程度，这个需求是极易满足的，而 Mulwad 等人提出的方法严重依赖于 Wikitology 的查询功能，一旦 Wikitology 的查询功能失效或者 Wikitology 不再开放，则他们的方法也就无法完成实体链接的任务

相比于英国谢菲尔德大学研制的 TableMiner 系统，本发明使用一种基于图的概率传播算法为给定表格中每个字符串的候选实体进行排序，该算法强调表格中任意单元格中的字符串之间均存在潜在的关系，从而选择这种联合消歧的方式捕捉字符串之间的关联，从而一次性完成表格整体的实体链接。而 TableMiner 不考虑同一表格中字符串之间的潜在关联，仅以给定字符串为中心，单独为每个字符串进行实体链接，不仅效率上不如本发明，而且在割裂了表格中字符串之间的关联后，实体链接的质量也并不高。

相比于清华大学研制的 LIEGE 系统，发明不仅能够针对多行多列的表格进行基于任意单一知识库的实体链接，而且还利用不同知识库中实体间的等价关系提出三种启发式规则，从而完成基于多知识库的实体链接任务。LIEGE 系统的设计仅针对列表型表格，即一列多行表格，提出一系列基于维基百科的特征进行单一知识库的实体链

接，局限性太强，而本发明从基于同义词典与字符串匹配规则的候选实体生成方法，到基于图的概率传播算法的候选实体排序方法，再到三种进行多知识库实体链接的启发式规则，整体对外部信息的依赖较小，所设计的特征都是通用易得，方法局限性小，适合各种场景下的表格实体链接任务。

相比于美国西北大学研制的 TabEL 系统，本发明可以不依赖于任何先验概率对多行多列的表格进行基于多知识库的实体链接。TabEL 系统利用统计万维网与维基百科中所有实体的相关信息，然后得到相应的先验概率，并依照此概率为给定表格中每个单元格里的字符串生成候选实体，这种方式得到的先验概率其实是不准确的，因为万维网与维基百科并不能涵盖这个世界的方方面面，他们更多地还是体现当前世界流行的内容。而本发明使用的候选实体生成与排序方法的并不考虑这样的先验概率，这样可以大大减少基于流行内容的先验概率对表格实体链接带来的谬误。此外，TabEL 系统还是不能完成基于多知识库的表格实体链接的任务，由于许多字符串所应该链接的实体不存在于给定的单一知识库中，导致使用 TabEL 系统进行表格实体链接的质量依旧不能令人满意。而本发明提出的基于不同知识库中实体间的等价关系的启发式规则，可以完成基于多知识库的表格实体链接任务，大大提高链接的准确性与覆盖率。

经过实验分析证明，利用本发明提出的基于多知识库的表格实体链接方法，可以完成基于任意的多个知识库的表格实体链接任务。无论是准确率、召回率还是 F 值，本发明在这些评价指标上都优于目前最先进的表格实体链接方法及系统。

附图说明

图 1 是本发明的基本过程的示意图；

图 2 是本发明中从单一知识库中抽取候选实体的流程图；

图 3 是本发明中基于图的概率传播算法的流程图。

具体实施方式

以下结合实施例和说明书附图，详细说明本发明的实施过程。

本发明是基于多知识库的表格实体链接方法，包括以下 3 个步骤：

1）每次从知识库集合 $\mathbf{K}=\{KB_1,KB_2,\ldots,KB_z\ldots,KB_n\}$中选定一个单一知识库 KB_z，按照如下方法从所述单一知识库 KB_z 中抽取候选实体，构建候选实体列表，最终得到每个单一知识库构建的候选实体列表，详细步骤如下：

由于将知识库中数百万的实体均作为每个字符串的候选实体是不切实际的，所以需要使用一种高效且低成本的方法为每个字符串快速选定若干个可能的候选实体，以便进一步使用更加复杂的方法对得到的候选实体进行进一步的判定。本发明为了尽可能在筛选候选实体时保证覆盖率，首先使用同义词典 BabelNet 与字符串匹配规则，将表格 T 中所有单元格里的字符串 s 生成源自单一知识库 KB_z 的候选实体，每个字符串 s 对应多个候选实体，这里结合图 2 说明候选实体的生成过程：

（1）为单一知识库 KB_z 中的每个实体查找其在同义词典 BabelNet 中的所有同义词，并构建每个实体对应的同义词集合；

（2）对每个字符串 s 进行分词，得到零散片段 $\{w_1(s),w_2(s),\ldots,w_v(s),\ldots,w_k(s)\}$，其中 $w_v(s)$表示对字符串 s 分词后的第 v 个片段，$v\in\{1,2,\ldots,k\}$，k 为对字符串 s 分词后得到零散片段的总数量，比如字符串“Michael Jordan”对应两个片段“Michael”与“Jordan”；

（3）使用字符串匹配规则为表格 T 中所有单元格里的字符串生成候选实体，该规则为：如果知识库 KB_z 中的某个实体 e 及 e 的某个同义词包含经过分词后的字符串 s 的某个零散片段 $w_v(s)$，则将该实体 e 作为字符串 s 的一个候选实体，比如给定抽取自 KB_1 中的实体“Michael Jeffrey Jordan”与“Michael Irwin Jordan”均可判定为字符串“Michael Jordan”的候选实体。

在对给定表格 T 中每个单元格里的字符串生成抽取自知识库 KB_z 中的候选实体后，为了最终确定每个字符串所应该链接的实体，需要对每个字符串的候选实体进行排序，即通常所认为的实体消岐工作。一般而言，不难发现表格中同一行或者同一列单元格里的字符串之间存在一定的关系，换句话说，即可认为一个表格中任意两个单元格中的字符串之间存在潜在的关联，所以本发明选择使用一种通用的基于图的概率传播算法为给定表格 T 中所有单元格里的字符串进行联合消岐，即同时为所有字符串各自的候选实体进行排序，该方法可以作用于任何单一的知识库，不依赖于任何特定的表格信息或特定知识库中的特定信息。

这里结合图 3 说明本发明提出的为表格 T 中每个字符串 s 所对应的候选实体进行排序的基于图的概率传播算法：

（1）按照如下方式为表格 T 构建实体消岐图 G：将表格 T 中每个单元格里的字符串作为一个字符串节点，将每个候选实体作为一个实体节点，将字符串——实体边作为一条存在于每个字符串与其对应的一个候选实体之间的无向边，将一条实体——

实体边作为一条存在于 G 中任意两个实体节点之间的无向边；

（2）计算实体消岐图 G 中每个字符串与其对应的每个候选实体之间的字符串——实体语言学相似度、字符串——实体上下文相似度，这两种相似度的计算同样不依赖于任何特定信息，是通用的从不同角度衡量字符串与实体之间的相似程度，并根据这两种相似度计算每条字符串——实体边的权重，计算方式如下：

（2a）字符串——实体语言学相似度：给定字符串 s 与实体 e，它们之间的字符串——实体语言学相似度 linSim(s,e)的定义如下所示：

$$\mathrm{linSim}(\mathrm{s},\mathrm{e}) = 1 - \frac{\mathrm{EditDistance}(\mathrm{s},\mathrm{l}(\mathrm{e}))}{\max\{|\mathrm{s}|,|\mathrm{l}(\mathrm{e})|\}}$$

其中 l(e)是实体 e 的标签字符串，|s|和|l(e)|分别表示字符串 s 的长度与实体 e 的标签字符串长度，$\mathrm{EditDistance}(\mathrm{s},\mathrm{l}(\mathrm{e}))$表示字符串 s 与实体 e 的标签字符串之间的编辑距离；

（2b）字符串——实体上下文相似度：给定字符串 s，取出与 s 所在单元格处于同一行及同一列的单元格中的所有字符串，再对这些字符串进行分词，收集这些字符串各自对应的若干零散片段，所有收集到的零散片段构成了字符串 s 的上下文集合 scSet(s)；给定实体 e，查询 e 所在的知识库 KB_z，取出所有 e 作为主语或宾语的三元组，并收集这些三元组中的所有除 e 以外的作为主语或宾语的实体，之后对这些实体的字符串标签进行分词，将这些字符串标签各自对应的零散片段均放置于集合 ecSet(e)中，ecSet(e)表示实体 e 的上下文集合；对于给定的字符串 s 与实体 e，它们之间的字符串——实体上下文相似度 secSim(s,e)如下所示：

$$\mathrm{secSim}(\mathrm{s},\mathrm{e}) = \frac{|\mathrm{scSet}(\mathrm{s}) \cap \mathrm{ecSet}(\mathrm{e})|}{|\mathrm{scSet}(\mathrm{s}) \cup \mathrm{ecSet}(\mathrm{e})|}$$

其中$|\mathrm{scSet}(\mathrm{s}) \cap \mathrm{ecSet}(\mathrm{e})|$表示字符串 s 与实体 e 各自的上下文集合的交集大小，$|\mathrm{scSet}(\mathrm{s}) \cup \mathrm{ecSet}(\mathrm{e})|$表示字符串 s 与实体 e 各自的上下文集合的并集大小。

（2c）字符串——实体边的权重：给定实体消岐图 G 中一个字符串节点 node(s)，该节点表示字符串 s，同时给定一个与该字符串节点相邻的实体节点 node(e)，该实体节点表示实体 e，node(s)与 node(e)之间的字符串——实体边的权重 $\mathrm{W_{se}}$(node(s),node(e))的定义如下所示：

$$\mathrm{W_{se}}(\mathrm{node}(\mathrm{s}),\mathrm{node}(\mathrm{e})) = \alpha_1 \times \mathrm{linSim}(\mathrm{s},\mathrm{e}) + \beta_1 \times \mathrm{secSim}(\mathrm{s},\mathrm{e}) + \gamma_1$$

其中，$\alpha_1 + \beta_1 + \gamma_1 = 1$，$\alpha_1 \in (0,1)$，$\beta_1 \in (0,1)$，$\gamma_1 \in (0,1)$且$\alpha_1 \gg \gamma_1$，$\beta_1 \gg \gamma_1$；这里经过多次实验，决定令 $\gamma_1 = 0.01$，$\alpha_1 = \beta_1 = 0.445$，不难发现

$W_{se}(node(s), node(e))$的最小值为 0.01，这是为了在后续的概率传播的过程中保证实体消岐图 G 的连通性。

（3）计算实体消岐图 G 中任意实体之间的实体——实体三元组相似度与实体——实体上下文相似度，这两种相似度的计算同样不依赖于任何特定信息，是通用的从不同角度衡量字符串与实体之间的相似程度，并根据这两种相似度计算每条实体——实体边的权重，计算方式如下：

（3a）实体——实体三元组相似度：给定两个源自同一知识库 KB_z 的实体 e_1 与 e_2，它们之间的三元组相似度 $triSim(e_1,e_2)$的定义如下所示：

$$triSim(e_1, e_2) = \begin{cases} 1，若e_1与e_2同时出现在KB_z中的某个三元组中 \\ 0，其他情况 \end{cases}$$

（3b）实体——实体上下文相似度：给定两个源自同一知识库 KB_z 的实体 e_1 与 e_2，查询 KB_z，取出所有 e_1 作为主语或宾语的三元组，并收集这些三元组中的所有除 e_1 以外的作为主语或宾语的实体，之后对这些实体的字符串标签进行分词，将这些字符串标签各自对应的零散片段均放置于集合 $ecSet(e_1)$中，$ecSet(e_1)$表示实体 e_1 的上下文集合，以同样的方式构建实体 e_2 的上下文集合 $ecSet(e_2)$，实体 e_1 与 e_2 间的实体——实体上下文相似度 $eecSim(e_1,e_2)$的定义如下所示：

$$eecSim(e_1, e_2) = \frac{|ecSet(e_1) \cap ecSet(e_2)|}{|ecSet(e_1) \cup ecSet(e_2)|}$$

其中$|ecSet(e_1) \cap ecSet(e_2)|$表示实体 e_1 与 e_2 各自的上下文集合的交集大小，$|ecSet(e_1) \cup ecSet(e_2)|$表示字符串 s 与实体 e 各自的上下文集合的并集大小。

（3c）实体——实体边的权重：给定实体消岐图 G 中任意两个实体节点 $node(e_1)$与 $node(e_2)$，这两个节点分别表示实体 e_1 与 e_2，$node(e_1)$ 与 $node(e_2)$之间的实体——实体边的权重 $W_{ee}(node(e_1),node(e_2))$的定义如下所示：

$$W_{ee}(node(e_1), node(e_2)) = \alpha_2 \times triSim(e_1, e_2) + \beta_2 \times eecSim(e_1, e_2) + \gamma_2$$

其中，$\alpha_2 + \beta_2 + \gamma_2 = 1$，$\alpha_2 \in (0,1)$，$\beta_2 \in (0,1)$，$\gamma_2 \in (0,1)$且$\alpha_2 \gg \gamma_2$，$\beta_2 \gg \gamma_2$；这里经过多次实验，决定令 $\gamma_2 = 0.01$，$\alpha_2 = \beta_2 = 0.445$，不难发现 $W_{ee}(node(e_1), node(e_2))$的最小值为 0.01，这同样是为了在后续的概率传播的过程中保证实体消岐图 G 的连通性。

（4）利用如下公式进行迭代概率传播，直至向量 **R** 收敛：

$$\mathbf{R} = ((1 - b) \times \frac{\mathbf{E}}{m} + b \times \mathbf{A}) \times \mathbf{R}$$

其中 m 为所构建的实体消岐图 G 中节点的总量，**E** 是一个 m×m 的全 1 矩阵，b 是一个接近 1 的常数，b∈[0.8,1)，经过多次实验，本发明最终令 b=0.85；**R** 是一个 m×1 的向量<r_1,r_2,…,r_m>，r_j 为 G 中第 j 个节点所关联到的概率值，j∈{1,2,…,m}，R 的初始值计算方式如下：若第 j 个节点为字符串节点，则 r_j=1/m，它表示该字符串节点的重要度；若第 j 个节点为实体节点，则 r_j=0，它表示该一字符串链接到该实体的概率值；A 是一个 m×m 邻接矩阵，表示方式如下：

$$\mathbf{A}=\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix}$$

其中 A_{xy} 表示从实体消岐图 G 中的第 x 个节点到第 y 个节点的转移概率，x∈{1,2,…,m}，y∈{1,2,…,m}，A_{xy} 的定义如下：

$$A_{xy}=\begin{cases} \dfrac{W_{se}(x,y)}{W_{se}(x,*)}，若\ x\neq y\ 且\ x\ 是字符串节点 \\ a\times\dfrac{W_{ee}(x,y)}{W_{ee}(x,*)}，若\ x\neq y\ 且\ x\ 与\ y\ 均是实体节点 \\ (1-a)\times\dfrac{W_{se}(y,x)}{W_{se}(*,x)}，若\ x\neq y\ 且\ x\ 是实体节点以及\ y\ 是字符串节点 \\ 0，若\ x=y \end{cases}$$

其中 W_{se}(x,y)表示字符串节点 x 与实体节点 y 之间的字符串——实体边权重，W_{se}(y,x)表示字符串节点 y 与实体节点 x 之间的字符串——实体边权重，W_{se}(x,*)表示字符串节点 x 与其相邻的每个实体节点之间的字符串——实体边权重的总和，W_{se}(*,x)表示实体节点 x 与其相邻的每个字符串节点之间的字符串——实体边权重的总和，W_{ee}(x,y)表示实体节点 x、y 之间的实体——实体边权重，W_{ee}(x,*)表示实体节点 x 与其相邻的每个实体节点之间的实体——实体边权重的总和，a 是一个常数，a∈(0,1)，经过多次实验，本发明最终令 a=0.5。

另外，根据马尔可夫链的收敛性定义，需要保证矩阵 A 非周期，所以本发明在任意两个节点之间增加一条特殊的无向边，这些特殊的边上所关联的转移概率为一极小值，这个值由常数 b 控制；**R** 收敛后，给定表格 T 中任意一个单元格里的字符串 s 及其对应的候选实体，根据这些候选实体所在的实体节点所关联的概率值，对字符串 s 对应的候选实体进行降序排列，从而得到表格 T 中每个单元格里的字符串所对应的已排序候选实体列表。

（5）所述向量 **R** 收敛后，根据候选实体所在的实体节点所关联的概率值，对字

符串 s 对应的候选实体进行降序排列，从而得到候选实体列表。

基于单一知识库的表格实体链接并不总能确保一个良好覆盖率，一种直观的解决方案是分别进行基于不同单一知识库的表格实体链接以提高覆盖率，但是这种方法带来的问题是同一字符串所链接到的不同知识库中的实体间并不具备等价关系，即面临着许多冲突，因此本发明使用如下方法以提高表格实体链接的覆盖率并且能够解决基于不同单一知识库的表格实体链接的结果间的冲突问题。

2）将每个字符串 s 所对应的 n 个候选实体列表中的所有实体划分成多个实体集合，这些实体集合可分为两类：第一类中的每个集合里的实体数量 num∈{2,3,…,n}，每个集合中的实体分别源自不同的候选实体列表，且这些实体两两之间均存在等价关系；第二类中的每个集合中的实体数量均为 1，每个集合中的实体仅源自一个候选实体列表且与源自其他候选实体列表中的每个实体之间均不存在等价关系；

3）针对每个字符串所对应的多个不同的实体集合，使用三种启发式规则为每个字符串 s 选择一个实体集合中的所有实体作为该字符串 s 所应该链接的存在于不同知识库中的实体，从而完成表格实体链接。：

下面介绍本发明提出的三种启发式规则如下：

规则一：如果在字符串 s 对应的多个实体集合中，存在一个集合 Set，与其他实体集合相比，Set 中所有实体在各自对应的候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，且集合 Set 中实体的数量 num 不小于所有给定知识库的数量的一半，则选择集合 Set 中的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则二：如果在字符串 s 对应的多个实体集合中，存在 g 个集合，g>1，这 g 个集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 相等，最高值 hr 也相等，且与其他实体集合相比，这 g 个集合中每个集合里的所有实体在各自候选实体列表中的排名的平均值 ar 与最高值 hr 均最高，此外这 g 个集合中每个集合里实体的数量均不小于所有给定知识库的数量的一半，则随机选择这 g 个集合中的一个集合里的所有实体为 s 所应该链接的存在于不同知识库中的实体；

规则三：如果在字符串 s 对应的每个实体集合中实体的数量均小于所有给定知识库的数量的一半，则取出在字符串 s 所对应的 n 个候选实体列表，将每个列表中排名第一的实体作为 s 所应该链接的存在于不同知识库中的实体。

为了争取同时获得全局与局部最优的实体链接结果，本发明提出的三种不同的启发式规则不仅考虑了每个字符串对应的每个实体集合中所有实体的平均排名与最高

排名，还有每个集合中实体的数量，即覆盖这些相同含义的实体的知识库的数量。如果给定集合中实体的数量低于所有给定知识库数量的一半，那么意味着该集合中的拥有相同含义的实体仅被很少的知识库所覆盖，所以若最终选择这个集合中的实体以解决基于不同单一知识库的实体链接结果间的冲突是不符合全局最优的设想的。

上述实施例仅是本发明的优选实施方式，应当指出：对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和等同替换，这些对本发明权利要求进行改进和等同替换后的技术方案，均落入本发明的保护范围。

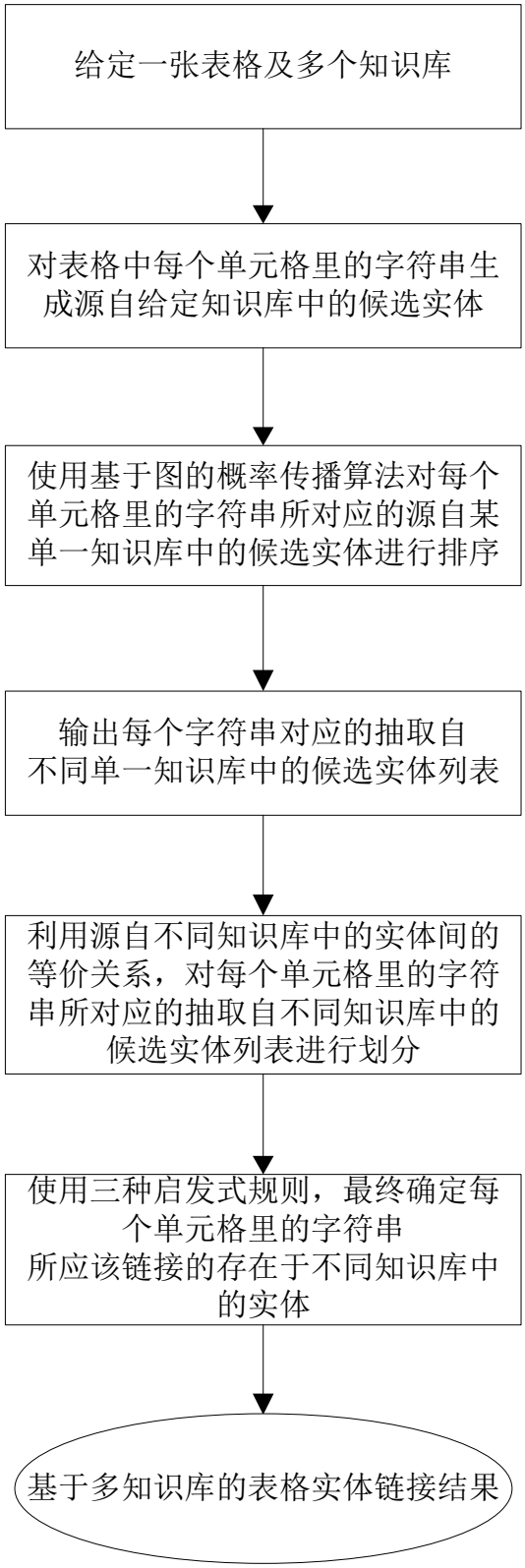

给定一张表格及多个知识库
对表格中每个单元格里的字符串生成源自给定知识库中的候选实体
使用基于图的概率传播算法对每个单元格里的字符串所对应的源自某单一知识库中的候选实体进行排序
输出每个字符串对应的抽取自不同单一知识库中的候选实体列表
利用源自不同知识库中的实体间的等价关系，对每个单元格里的字符串所对应的抽取自不同知识库中的候选实体列表进行划分
使用三种启发式规则，最终确定每个单元格里的字符串所应该链接的存在于不同知识库中的实体
基于多知识库的表格实体链接结果

图 1

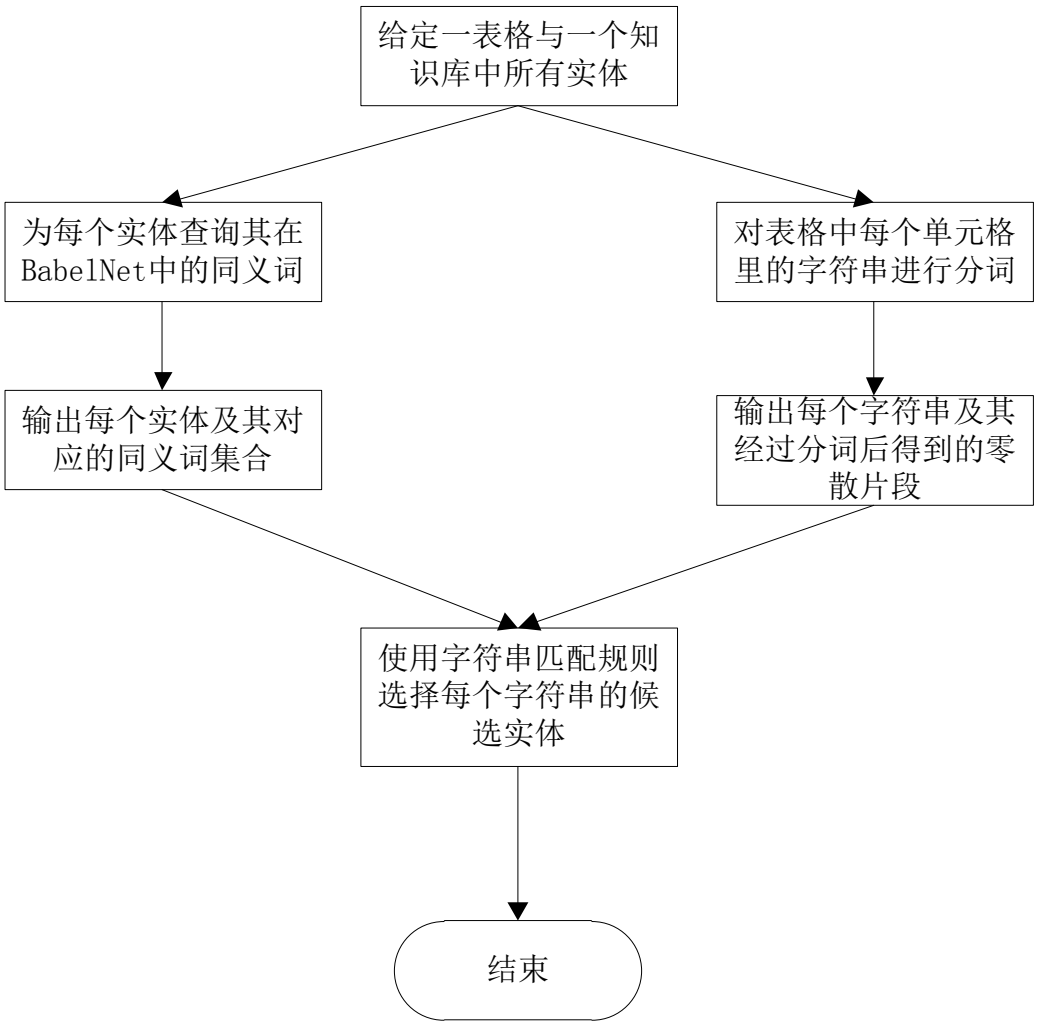

给定一表格与一个知识库中所有实体
为每个实体查询其在BabelNet中的同义词
对表格中每个单元格里的字符串进行分词
输出每个实体及其对应的同义词集合
输出每个字符串及其经过分词后得到的零散片段
使用字符串匹配规则选择每个字符串的候选实体
结束

图 2

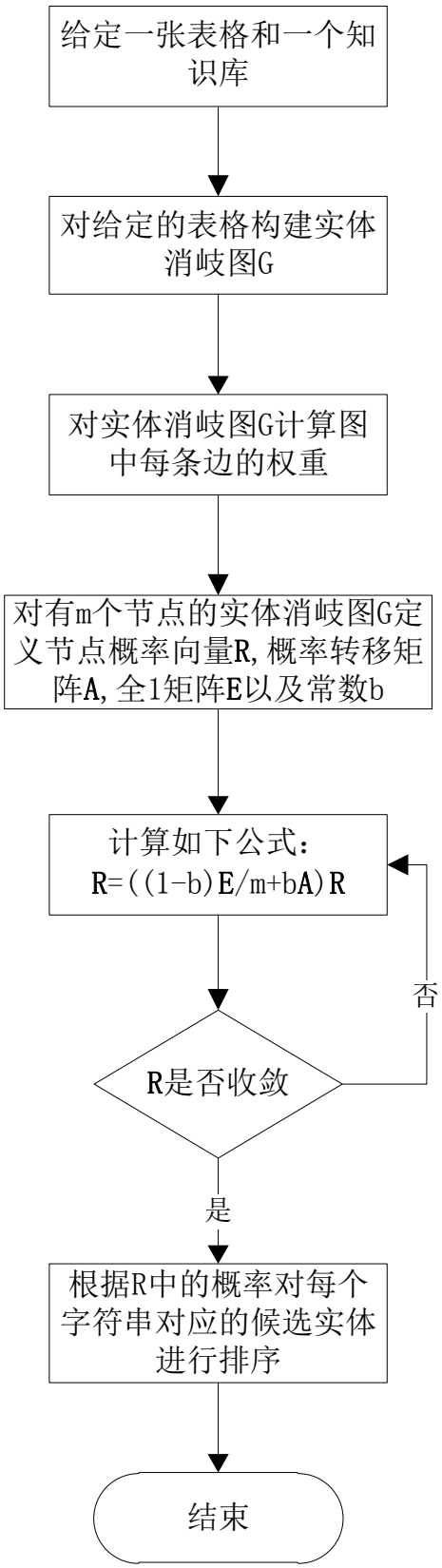

给定一张表格和一个知识库
对给定的表格构建实体消岐图G
对实体消岐图G计算图中每条边的权重
对有m个节点的实体消岐图G定义节点概率向量R,概率转移矩阵A,全1矩阵E以及常数b
计算如下公式:
R=((1-b)E/m+bA)R
R是否收敛
否
是
根据R中的概率对每个字符串对应的候选实体进行排序
结束

图 3