



(12) 发明专利申请

(10) 申请公布号 CN 103745014 A

(43) 申请公布日 2014. 04. 23

(21) 申请号 201410042782. 5

(22) 申请日 2014. 01. 29

(71) 申请人 中国科学院计算技术研究所

地址 100190 北京市海淀区中关村科学院南路 6 号

(72) 发明人 梁英 胡开先 许洪波 程学旗  
张国清

(74) 专利代理机构 北京泛华伟业知识产权代理有限公司 11280

代理人 王勇

(51) Int. Cl.

G06F 17/30 (2006. 01)

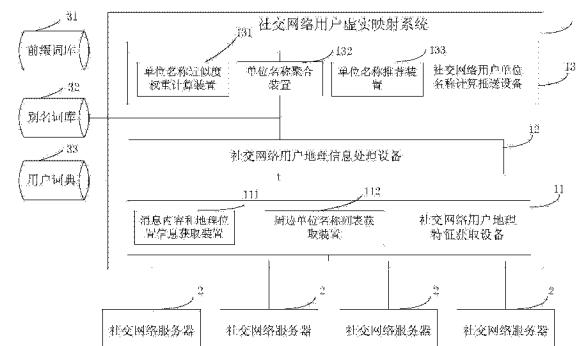
权利要求书3页 说明书11页 附图3页

(54) 发明名称

一种社交网络用户虚实映射方法和系统

(57) 摘要

本发明提供一种社交网络用户虚实映射方法和系统，所述方法包括：根据社交网络用户的唯一标识获取该社交网络用户的地理位置信息，并且获取该地理位置信息对应的地理位置的周边单位名称列表；以及将所述周边单位名称列表中的每个单位名称与所述社交网络用户发布的消息内容进行匹配，根据匹配程度选择一个或多个单位名称。本发明可根据社交网络用户的唯一标识将该用户映射到一个或多个工作单位，提高了社交网络用户虚实映射的精确度。



1. 一种社交网络用户虚实映射方法,包括:

步骤 1)、根据社交网络用户的唯一标识获取该社交网络用户的地理位置信息,并且获取该地理位置信息对应的地理位置的周边单位名称列表;

步骤 2)、将所述周边单位名称列表中的每个单位名称与所述社交网络用户发布的消息内容进行匹配,根据匹配程度选择一个或多个单位名称。

2. 根据权利要求 1 所述的方法,其中,步骤 1) 还包括:

对所述周边单位名称列表中的每个单位名称进行分词,得到该单位名称的分词结果。

3. 根据权利要求 2 所述的方法,在步骤 2) 中,将周边单位名称列表中的每个单位名称与社交网络用户发布的消息内容进行匹配包括:

步骤 21)、将每个单位名称的全称与所述社交网络用户发布的消息内容进行匹配,如果匹配成功则使用下式计算该单位名称的近似度权重:

$$\text{weight}(\text{str}) = \text{word.size}(\text{str}) * \text{factor}^{\text{matchtime}(\text{sstr})}$$

其中, str 表示单位名称, weight(str) 表示单位名称的近似度权重, word.size(str) 表示单位名称的长度, factor 表示乘数因子, matchtimes(str) 表示单位名称与消息内容的匹配成功次数;

步骤 22)、如果匹配不成功,则将该单位名称的分词结果中除该单位名称的全称外的每个分词与所述社交网络用户发布的消息内容进行匹配,将每个分词的匹配成功次数之和作为该单位名称的近似度权重。

4. 根据权利要求 3 所述的方法,其中,步骤 2) 还包括:

步骤 23)、合并近似度权重相同且具有共同的最大前缀的单位名称,使得所述共同的最大前缀包含在前缀词库中或者其长度达到预定长度;其中,合并后的单位名称为所述共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和,所述前缀词库用于存放指示地理位置的前缀词。

5. 根据权利要求 4 所述的方法,其中,步骤 23) 包括:

步骤 231)、对于一种近似度权重,新建一棵 Trie 树;

步骤 232)、将具有该近似度权重的单位名称插入所述 Trie 树,得到具有共同的最大前缀的单位名称;

步骤 233)、如果该共同的最大前缀包含在前缀词库中或者其长度达到预定长度,则合并具有该共同的最大前缀且具有该近似度权重的单位名称;其中,合并后的单位名称为该共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和,所述前缀词库用于存放指示地理位置的前缀词;

步骤 234)、销毁所述 Trie 树;

步骤 235)、对于未处理的近似度权重,返回步骤 231) 进行处理。

6. 根据权利要求 4 所述的方法,其中,步骤 2) 还包括:

步骤 24)、合并近似度权重不同且单位名称相同或者互为别名的单位名称;其中,合并后的单位名称为所合并的单位名称中的任何一个,其近似度权重为所合并的单位名称的近似度权重之和。

7. 根据权利要求 3-6 中任何一个所述的方法,在步骤 2) 中,根据匹配程度选择一个或多个单位名称包括:

将单位名称按照近似度权重进行降序排列,选择前N个单位名称并推送;其中N为正整数。

8. 根据权利要求2-6中任何一个所述的方法,在步骤1)中,对周边单位名称列表中的每个单位名称进行分词包括:

对所述周边单位名称列表中的每个单位名称进行中文分词,删除长度为一个字的分词;以及

对所述周边单位名称列表中的每个单位名称进行二元组分词。

9. 根据权利要求2-6中任何一个所述的方法,其中,步骤1)还包括:

如果对单位名称进行分词后得到的分词具有别名,则将该别名加入该单位名称的分词结果。

10. 根据权利要求1所述的方法,其中,步骤1)包括:

步骤11)、根据社交网络用户的唯一标识从社交网络服务器获取关于该社交网络用户的返回信息,从中得到该社交网络用户发布的消息内容和地理位置信息;

步骤12)、根据所述社交网络用户的地理位置信息从社交网络服务器获取该地理位置信息对应的地理位置的周边单位名称列表。

11. 根据权利要求10所述的方法,其中社交网络用户的地理位置信息包括该社交网络用户发布消息的地理位置信息和该社交网络用户签到的地理位置信息。

12. 根据权利要求10或11所述的方法,其中,步骤11)还包括:

统一所述社交网络用户的地理位置信息的精确度;以及

按照出现次数降序排序所述社交网络用户的地理位置信息,选择前M个地理位置信息;其中M为正整数。

13. 根据权利要求12所述的方法,其中,步骤12)包括:

根据所选择的M个地理位置信息,从社交网络服务器获取对应的地理位置的周边单位名称列表。

14. 一种社交网络用户虚实映射系统(1),包括:

社交网络用户地理特征获取设备(11),用于根据社交网络用户的唯一标识获取该社交网络用户的地理位置信息,并且获取该地理位置信息对应的地理位置的周边单位名称列表;以及

社交网络用户单位名称计算推送设备(13),用于将所述周边单位名称列表中的每个单位名称与所述社交网络用户发布的消息内容进行匹配,根据匹配程度选择一个或多个单位名称。

15. 根据权利要求14所述的系统(1),其中,所述系统还包括:

社交网络用户地理信息处理设备(12),用于对所述周边单位名称列表中的每个单位名称进行分词,得到该单位名称的分词结果。

16. 根据权利要求15所述的系统(1),其中,所述社交网络用户单位名称计算推送设备(13)用于将每个单位名称的全称与所述社交网络用户发布的消息内容进行匹配,如果匹配成功则使用下式计算该单位名称的近似度权重:

$$\text{weight}(\text{str}) = \text{word.size}(\text{str}) * \text{factor}^{\text{matchtime}(\text{sstr})}$$

其中, str 表示单位名称, weight(str) 表示单位名称的近似度权重, word.size(str)

表示单位名称的长度, factor 表示乘数因子, matchtimes(str) 表示单位名称与消息内容的匹配成功次数;如果匹配不成功,则将该单位名称的分词结果中除该单位名称的全称外的每个分词与所述社交网络用户发布的消息内容进行匹配,将每个分词的匹配成功次数之和作为该单位名称的近似度权重。

17. 根据权利要求 16 所述的系统(1),其中,所述社交网络用户单位名称计算推送设备(13)还用于合并近似度权重相同且具有共同的最大前缀的单位名称,使得所述共同的最大前缀包含在前缀词库中或者其长度达到预定长度;其中,合并后的单位名称为所述共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和,所述前缀词库用于存放指示地理位置的前缀词。

18. 根据权利要求 16 所述的系统(1),其中,所述社交网络用户单位名称计算推送设备(13)还用于合并近似度权重不同且单位名称相同或者互为别名的单位名称;其中,合并后的单位名称为所合并的单位名称中的任何一个,其近似度权重为所合并的单位名称的近似度权重之和。

## 一种社交网络用户虚实映射方法和系统

### 技术领域

[0001] 本发明涉及计算机数据挖掘分析领域,尤其涉及一种社交网络用户虚实映射方法和系统。

### 背景技术

[0002] 随着互联网的不断发展,社交网络在人们生活中扮演着重要的角色,微博、微信、人人网等已经成为人们获取信息、展示自我和营销推广的重要手段。通过社交网络,人们可以方便地以虚拟身份自由发表观点和意见,每个人都是信息的生产者和消费者,形成“自媒体”。匿名的社交网络在一定程度上保护了用户的隐私,但同样也带来了很多问题。比如,不易追踪网络虚假消息的发布者、不易定位危害国家治安言论的发布者、不易在网络中追查违法犯罪行为等。尽管我国推出了网络实名制注册政策,但面对无边界的网络社会,实名制需要全国统一,甚至需要与世界接轨,因此在实施过程中困难重重。此外,即使是采用了实名制注册也是后台实名,而前台显示仍使用昵称的匿名形式。因此,在网络的虚拟性、匿名性等特征下,根据用户在社交网络中的虚拟身份来识别用户的真实身份,即实现用户的虚实映射,具有积极的社会意义。

[0003] 目前,针对社交网络中用户虚实映射的研究主要包括以下两个方面:一类是基于网络 IP 地址定位网络设备,如通过分析移动设备 IP 地址,网络设备 IP 地址及台式电脑 IP 地址等来获取设备所在的省市信息;另一类是通过人物特征属性对某个用户群体进行识别,用户的特征属性可以包括性别、年龄段、居住地、毕业院校、性格、星座、爱好、职业等,通过挖掘网络数据来识别用户特征属性,可以将拥有相同或相似特征属性的用户群挖掘出来,以便为网络营销、电商广告提供服务。

[0004] 然而,由于实践中难以获得社交网络用户的 IP 信息,因此前一类方法的适用范围受到一定限制,不能满足 IP 缺失的社交网络用户的虚实映射需求;而第二类方法面向拥有相似特征的用户群,并不是面向个人用户,其偏向于挖掘用户的特征属性分类,并不能识别用户的真实身份。

### 发明内容

[0005] 为解决现有技术中存在的问题,本发明提供一种社交网络用户虚实映射方法,所述方法包括:

[0006] 步骤 1)、根据社交网络用户的唯一标识获取该社交网络用户的地理位置信息,并且获取该地理位置信息对应的地理位置的周边单位名称列表;

[0007] 步骤 2)、将所述周边单位名称列表中的每个单位名称与所述社交网络用户发布的消息内容进行匹配,根据匹配程度选择一个或多个单位名称。

[0008] 在一个实施例中,步骤 1) 还包括:对所述周边单位名称列表中的每个单位名称进行分词,得到该单位名称的分词结果。

[0009] 在一个实施例中,在步骤 2) 中,将周边单位名称列表中的每个单位名称与社交网

络用户发布的消息内容进行匹配包括：

[0010] 步骤 21)、将每个单位名称的全称与所述社交网络用户发布的消息内容进行匹配，如果匹配成功则使用下式计算该单位名称的近似度权重：

[0011]  $\text{weight}(\text{str}) = \text{word.size}(\text{str}) * \text{factor}^{\text{matchtime}(\text{sstr})}$

[0012] 其中， $\text{str}$  表示单位名称， $\text{weight}(\text{str})$  表示单位名称的近似度权重， $\text{word.size}(\text{str})$  表示单位名称的长度， $\text{factor}$  表示乘数因子， $\text{matchtimes}(\text{str})$  表示单位名称与消息内容的匹配成功次数；

[0013] 步骤 22)、如果匹配不成功，则将该单位名称的分词结果中除该单位名称的全称外的每个分词与所述社交网络用户发布的消息内容进行匹配，将每个分词的匹配成功次数之和作为该单位名称的近似度权重。

[0014] 在一个实施例中，步骤 2) 还包括：

[0015] 步骤 23)、合并近似度权重相同且具有共同的最大前缀的单位名称，使得所述共同的最大前缀包含在前缀词库中或者其长度达到预定长度；其中，合并后的单位名称为所述共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和，所述前缀词库用于存放指示地理位置的前缀词。

[0016] 在进一步的实施例中，步骤 23) 包括：

[0017] 步骤 231)、对于一种近似度权重，新建一棵 Trie 树；

[0018] 步骤 232)、将具有该近似度权重的单位名称插入所述 Trie 树，得到具有共同的最大前缀的单位名称；

[0019] 步骤 233)、如果该共同的最大前缀包含在前缀词库中或者其长度达到预定长度，则合并具有该共同的最大前缀且具有该近似度权重的单位名称；其中，合并后的单位名称为该共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和，所述前缀词库用于存放指示地理位置的前缀词；

[0020] 步骤 234)、销毁所述 Trie 树；

[0021] 步骤 235)、对于未处理的近似度权重，返回步骤 231) 进行处理。

[0022] 在一个实施例中，步骤 2) 还包括：

[0023] 步骤 24)、合并近似度权重不同且单位名称相同或者互为别名的单位名称；其中，合并后的单位名称为所合并的单位名称中的任何一个，其近似度权重为所合并的单位名称的近似度权重之和。

[0024] 在一个实施例中，在步骤 2) 中，根据匹配程度选择一个或多个单位名称包括：将单位名称按照近似度权重进行降序排列，选择前 N 个单位名称并推送；其中 N 为正整数。

[0025] 在一个实施例中，对周边单位名称列表中的每个单位名称进行分词包括：对所述周边单位名称列表中的每个单位名称进行中文分词，删除长度为一个字的分词；以及对所述周边单位名称列表中的每个单位名称进行二元组分词。

[0026] 在一个实施例中，对周边单位名称列表中的每个单位名称进行分词还包括：如果对单位名称进行分词后得到的分词具有别名，则将该别名加入该单位名称的分词结果。

[0027] 在一个实施例中，步骤 1) 包括：

[0028] 步骤 11)、根据社交网络用户的唯一标识从社交网络服务器获取关于该社交网络用户的返回信息，从中得到该社交网络用户发布的消息内容和地理位置信息；

[0029] 步骤 12)、根据所述社交网络用户的地理位置信息从社交网络服务器获取该地理位置信息对应的地理位置的周边单位名称列表。其中社交网络用户的地理位置信息包括该社交网络用户发布消息的地理位置信息和该社交网络用户签到的地理位置信息。

[0030] 在进一步的实施例中,步骤 11) 还包括 :统一所述社交网络用户的地理位置信息的精确度 ;以及按照出现次数降序排序所述社交网络用户的地理位置信息,选择前 M 个地理位置信息 ;其中 M 为正整数。

[0031] 在进一步的实施例中,步骤 12) 包括 :根据所选择的 M 个地理位置信息,从社交网络服务器获取对应的地理位置的周边单位名称列表。

[0032] 根据本发明的一个实施例,还提供一种社交网络用户虚实映射系统,包括 :

[0033] 社交网络用户地理特征获取设备,用于根据社交网络用户的唯一标识获取该社交网络用户的地理位置信息,并且获取该地理位置信息对应的地理位置的周边单位名称列表 ;以及

[0034] 社交网络用户单位名称计算推送设备,用于将所述周边单位名称列表中的每个单位名称与所述社交网络用户发布的消息内容进行匹配,根据匹配程度选择一个或多个单位名称。

[0035] 在一个实施例中,所述系统还包括 :

[0036] 社交网络用户地理信息处理设备,用于对所述周边单位名称列表中的每个单位名称进行分词,得到该单位名称的分词结果。

[0037] 在一个实施例中,所述社交网络用户单位名称计算推送设备用于将每个单位名称的全称与所述社交网络用户发布的消息内容进行匹配,如果匹配成功则使用下式计算该单位名称的近似度权重 :

[0038]  $\text{weight}(\text{str}) = \text{word.size}(\text{str}) * \text{factor}^{\text{matchtime}(\text{str})}$

[0039] 其中, str 表示单位名称, weight(str) 表示单位名称的近似度权重, word.size(str) 表示单位名称的长度, factor 表示乘数因子, matchtimes(str) 表示单位名称与消息内容的匹配成功次数 ;如果匹配不成功,则将该单位名称的分词结果中除该单位名称的全称外的每个分词与所述社交网络用户发布的消息内容进行匹配,将每个分词的匹配成功次数之和作为该单位名称的近似度权重。

[0040] 在一个实施例中,所述社交网络用户单位名称计算推送设备还用于合并近似度权重相同且具有共同的最大前缀的单位名称,使得所述共同的最大前缀包含在前缀词库中或者其长度达到预定长度 ;其中,合并后的单位名称为所述共同的最大前缀且其近似度权重为所合并的单位名称的近似度权重之和,所述前缀词库用于存放指示地理位置的前缀词。

[0041] 在一个实施例中,所述社交网络用户单位名称计算推送设备还用于合并近似度权重不同且单位名称相同或者互为别名的单位名称 ;其中,合并后的单位名称为所合并的单位名称中的任何一个,其近似度权重为所合并的单位名称的近似度权重之和。

[0042] 采用本发明可以达到如下的有益效果 :

[0043] 本发明可根据社交网络用户的唯一标识将该社交网络用户映射到一个或多个工作单位,实现了社交网络用户到其工作单位的虚实映射。根据从社交网络用户地理位置信息得到的单位名称与该用户发布的消息内容的匹配程度来计算每个单位名称的可能性,提高了识别社交网络用户工作单位的准确性。对具有共同的最大前缀以及相同或互为别名

的工作单位名称进行聚合，极大程度地减少了冗余重复，进一步提高了社交网络用户虚实映射的准确性。此外，地理位置信息除了考虑用户发布消息的位置，还考虑了用户的签到位置，同样提高了用户虚实映射的准确性。

### 附图说明

- [0044] 图 1 是根据本发明一个实施例的社交网络用户虚实映射方法的流程图；
- [0045] 图 2 是根据本发明一个实施例的单位名称聚合方法的流程图；
- [0046] 图 3 是根据本发明一个实施例的构建前缀树的方法示意图；以及
- [0047] 图 4 是根据本发明一个实施例的社交网络用户虚实映射系统的框图。

### 具体实施方式

[0048] 下面结合附图和具体实施方式对本发明进行说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0049] 根据本发明的一个实施例，提供一种社交网络用户虚实映射方法。概括而言，该方法首先根据社交网络用户的唯一标识在社交网络服务器上获取该用户的地理位置信息，并且利用获取的地理位置信息在社交网络服务器上获取周边的单位名称列表；其次，根据所获取的每个单位名称与该用户发布的消息内容(例如，用户的发言内容、发表的博文内容等)的匹配程度得到每个单位名称的近似度权重；接着，对周边单位名称列表中的单位名称进行聚合，以基于聚合后的近似度权重来推送一个或多个工作单位名称。参考图 1 并以微博用户为例，该方法可使用以下步骤进行描述：

- [0050] 步骤 S101：输入微博用户的唯一标识
- [0051] 步骤 S102：获取该微博用户发布的消息内容(即发表的博文内容)和地理位置信息
- [0052] 在一个实施例中，可根据输入的微博用户唯一标识向社交网络服务器(在本实施例中为微博服务器)发出请求，获得批准后，在该微博服务器上抓取微博用户发表的博文内容和地理位置信息。如果没有抓取到关于该微博用户的博文内容或地理位置的信息，则虚实映射过程结束；如果抓取成功，则由微博服务器返回包括用户的博文内容、发表博文的地理位置以及用户签到的地理位置等返回信息，其中后面两种信息构成微博用户的地理位置信息。在进一步的实施例中，可以仅抓取特定时间区间内的信息，例如抓取工作时间段内的信息。
- [0053] 在一个实施例中，在抓取成功并获得返回信息后，可首先统一地理位置信息的精确度。例如，由于地理位置信息通常表示为经纬度，则可以统一经纬度数据的小数点后的位数。继而统计微博用户的每个地理位置信息出现的次数，将该出现次数作为指标对地理位置信息进行降序排序，选取前 N 个作为频繁地理位置信息。
- [0054] 步骤 S103：获取周边单位名称列表
- [0055] 将上一步中获取的地理位置信息，逐条发送给微博服务器，定位该地理位置信息对应的地理位置并且获取该地理位置的周边单位名称列表。在另一个实施例中，可以将频繁地理位置信息逐条发送给微博服务器，并且获取对应的周边单位名称列表。
- [0056] 步骤 S104：对获取到的周边单位名称列表进行分词
- [0057] 在一个实施例中，可采用本领域技术人员公知的分词方法对周边单位名称列表中

的单位名称进行分词,得到每个单位名称对应的分词结果(包括一个或多个分词,其中分词可包括单位名称的全称)。举例来说,如果单位名称为 Entity B,则其分词结果可包括:Entity B、Entity 以及 B。

[0058] 在一个优选的实施例中,可首先对周边单位名称列表中的单位名称进行中文分词,在中文分词结束后再继续对这些单位名称进行二元组分词。其中,中文分词过程结束后,可移除长度为单个字的那些分词。

[0059] 在一个实施例中,在中文分词过程中还可以参考用户词典,参考用户词典可使分词结果更符合客户需要。其中,用户词典是指用户维护的一个词典,参考该词典是指用户在分词过程中将特定的词分成一个分词,而不是按照默认的方法进行分词。

[0060] 在另一个实施例中,如果在分词过程中发现得到的某个分词在别名词库中拥有别名(例如单位名称缩略词),则将该分词的所有别名也加入该单位名称对应的分词结果中。其中,别名词库是由用户维护的一个词库,用于保存多个单位名称的别名,其帮助识别分词和信息处理过程中拥有别名的实体和互为别名的实体,以达到提高匹配效率的目的。

[0061] 步骤 S105 :计算每个单位名称的近似度权重

[0062] 在周边单位名称列表的分词过程结束后,可计算该周边单位名称列表中的每个单位名称的近似度权重,即计算单位名称与微博用户发表的博文内容的匹配程度。

[0063] 在一个实施例中,可首先将单位名称对应的分词结果与博文内容进行匹配,能够完全匹配的单位名称的近似度权重高,而基本匹配的单位名称的近似度权重低。其中,完全匹配表示单位名称的全称在博文内容中得到匹配(如 Entity B 成功匹配);而基本匹配表示单位名称的分词结果中的分词(不包括单位名称的全称)在博文内容中得到匹配(如 Entity 或 B 成功匹配)。

[0064] 在一个实施例中,可根据下式来计算完全匹配的单位名称的近似度权重:

[0065]  $\text{weight}(\text{str}) = \text{word.size}(\text{str}) * \text{factor}^{\text{matchtime}(\text{sstr})}$  (1)

[0066] 其中, str 代表输入字符串,例如单位名称;weight(str) 代表输入字符串的近似度权重,word.size(str) 代表输入字符串的长度,factor 代表乘数因子,matchtimes(str) 代表输入字符串与博文内容的匹配次数。假定输入字符串为 Entity B,则 word.size 等于 8,设置乘数因子为 1.1,如果完全匹配次数为 10 次,那么计算得到的近似度权重约为 20。

[0067] 在另一个实施例中,计算基本匹配的单位名称的近似度权重包括:计算该单位名称对应的分词结果中每个分词(不包括整个单位名称全称)与博文内容的匹配次数之和。

[0068] 举例来说,假定输入的某微博用户的唯一标识为 A,由 A 获取的周边单位名称列表中包括 Entity B、Entity C、Entity D 等,并且其中,

[0069] Entity B 的分词结果是 Entity B,Entity, B ;

[0070] Entity C 的分词结果是 Entity C,Entity, C ;

[0071] Entity D 的分词结果是 Entity D,Entity, D.....

[0072] 则可采用如下算法来计算单位名称的近似度权重:

[0073]

---

### **算法: entityWeight**

---

[0074]

---

**输入:** entityBuf

**输出:** weight

---

//读取分词后的周边单位名称列表并保存

1: readSaveClassifiedEntities( entityBuf, entitylist); /\* entityBuf 保存分词后的周边单位名称列表, entitylist 是周边单位名称列表的文件 \*/

//对列表中的每个单位名称根据分词计算近似度权重 weight

2: **foreach** entity in entityBuf as Entity B, Entity C .....

3: **foreach** classified word of entity as Entity B, Entity, B .....

4:     weight = weight + matchAndCount(word, matchtimes);

5: **end for**

6: **end for**/\* matchAndCount 函数用于计算近似度权重 weight, 其针对完全匹配或基本匹配, 分别采用相应的方法进行计算, matchtimes 表示匹配次数 \*/

---

[0075] 步骤 S106 :对单位名称进行聚合

[0076] 在一个实施例中, 聚合单位名称可包括聚合周边单位名称列表中拥有共同最大前缀(即最长公共前缀)的单位名称以及使用别名词库聚合互为别名或者相同的单位名称, 并且统计聚合后的单位名称列表中每个单位名称的近似度权重。其中, 共同的最大前缀需满足以下条件才有效: 即其是前缀词库中包括的前缀词或者满足用户预先设定的长度。

[0077] 前缀词库用于存放用户指定的关于地理位置的前缀词, 包括单位名称、地理位置以及地址等。其作用是判定所得单位名称的共同最大前缀是否有效, 使得该共同最大前缀有效时才执行聚合。

[0078] 步骤 S107 :推送该微博用户最有可能的工作单位名称

[0079] 将聚合后的单位名称按照近似度权重进行降序排列, 输出前 N 项。其中, N 可以是用户预先指定的一个值。

[0080] 现参考图 2 进一步描述步骤 S106, 根据本发明的一个实施例, 聚合单位名称首先对具有相同近似度权重的单位名称进行聚合, 再对具有不同相似度权重的单位名称进行聚合, 包括以下子步骤:

[0081] 步骤 S201 :逐个输入具有相同近似度权重的单位名称。

[0082] 步骤 S202 :将具有相同近似度权重的每个单位名称插入一棵前缀树。

[0083] 对于当前处理的近似度权重, 新建一个前缀树(具有相同近似度权重的单位名称插入同一个前缀树), 该前缀树可以采用 Trie 树结构。如图 3 所示, 在每个前缀树中, 有一个根节点、多个中间节点和叶子节点。其中, 拥有儿子或者兄弟的非根节点被称为中间节点, 没有儿子和兄弟的节点被称作叶子节点。

[0084] 节点定义:

[0085]

```
struct node{
    string key;
    int matchtimes;
    int depth;
    node * parent;
    node * leftchild;
    node * nextsiblin;
}
```

[0086] 其中 key 代表节点的值, matchtimes 代表匹配次数, depth 代表节点的深度, parent 代表指向父节点的指针, leftchild 代表指向左儿子的指针, nextsiblin 代表指向右兄弟的指针。

[0087] 接着,将步骤 S201 中输入的各个单位名称逐条插入到新建的前缀树中。以近似度权重相同的单位名称 Entity B、Entity C、Firm E 和 Firm FG 为例,图 3 示出了该插入过程:

[0088] 首先插入单位名称“Entity B”,插入前缀树的过程包括将“Entity B”中的每个字分别插入到前缀树中,直到处理结束。

[0089] 接着处理下个单位名称,下个单位名称是“Entity C”,将其插入到前缀树,注意到“Entity”已经存在于前缀树中,因此不再插入新的节点,只是增加组成“Entity”的六个字母和一个空格的 7 个节点的匹配次数。当处理到“C”这个字时发现与当前节点“B”不相同,则新建“B”的右兄弟节点“C”。

[0090] 对于单位名称“Firm E”和“Firm FG”,也采用同样的处理方式。首先检查当前节点的值是否匹配当前处理的字,如果不匹配则新建兄弟节点,并将指针指向兄弟节点的子节点,如果匹配则将当前节点的匹配次数加 1,将指针指向该节点的子节点,如果当前值为空,则直接将当前处理的字符值赋给当前节点。

[0091] 步骤 S203 :在前缀树中找共同的最大前缀。

[0092] 共同的最大前缀是指一个或多个(近似度权重相同的)单位名称具有相同的前缀,该相同的前缀可包括一个或多个字,取最大的相同前缀即得到共同最大前缀。在图 3 的示例中,共同最大前缀包括“Entity”和“Firm”。在进一步的实施例中,可在本步骤去除最后一个空格,得到共同最大前缀“Entity”和“Firm”。

[0093] 步骤 S204 :聚合拥有共同最大前缀的单位名称,并且合并这些单位名称的近似度权重,得到聚合后的单位名称以及对应的近似度权重。

[0094] 在本步骤中,只聚合共同最大前缀属于前缀词库或者满足长度要求的那些单位名称。例如,对于共同最大前缀“Entity”和“Firm”,如果前缀词库中包括这两个词,则将单位名称“Entity B”和“Entity C”合并为单位名称“Entity”,计算“Entity”的近似度权重为“Entity B”(或“Entity C”)近似度权重与匹配次数(即 2)的乘积(或者看成“Entity B”的近似度权重与“Entity C”的近似度权重之和),同理可计算聚合后的单位名称“Firm”的

权重。又例如,对于共同最大前缀“Entity”和“Firm”,如果前缀词库中不包括这两个词,设置共同最大前缀需要满足5个字母长度,则可以聚合前缀为“Entity”的单位名称,而不聚合前缀为“Firm”的单位名称。

[0095] 如果没有共同最大前缀,或者具有共同最大前缀但该共同最大前缀不属于前缀词库且不满足共同最大前缀的长度要求,则保持原本的单位名称及其近似度权重不变。

[0096] 在一个实施例中,在完成聚合后,还要删除所构建的前缀树。

[0097] 在步骤S105计算每个单位名称的近似度权重后,会得到多种近似度权重。步骤S201-S204仅聚合了一种近似度权重的单位名称,对于未被处理的近似度权重,重复步骤S201-S204,直到所有的近似度权重均已处理。

[0098] 步骤S205:聚合不同近似度权重间单位名称相同或互为别名的单位名称,并且合并其近似度权重。

[0099] 可以参考上述别名词库来逐个比对不同近似度权重对应的单位名称,如果发现它们相同或者互为别名就将其合并,并将权重相加作为合并后的单位名称的权重。这样做可以最大程度地去除最终结果中的重复,以便提高映射的准确率。

[0100] 在一个实施例中,聚合单位名称的算法描述如下:

---

**算法:** entityAggregate

---

**输入:** entityKeyValue

---

**输出:** entityKeyValue

---

//读取前缀词库并保存

1: readSavePrefixDict(prefixBuf, prefixDict); /\* prefixBuf 保存前缀词库, prefixDict 是前缀词库的文件 \*/

//逐个输入拥有相同权重的单位名称

2: **foreach** 权重 in entityKeyValue /\* entityKeyValue 是保存 entityBuf 中每一条单位名称对应的近似度权重的变量 \*/

3: newTree(); //新建树

//把具有相同权重的每个单位名称插入前缀树

4: **foreach** 相同权重的单位名称 as Entity B, Entity C, Firm E .....

5: insertTree( Entity B); /\* 分别把 Entity B、Entity C...插入前缀树 \*/

6: **end for**

7: getMaxPrefix(root); //获得最大前缀

8: getMaxMatchtime(root); //获得前缀匹配次数

//聚合拥有共同的最大前缀的单位名称并且合并其近似度权重

9: **if**( maxPrefix > lengthLimit || maxPrefix in prefixDic)

10: aggregateEntityandWeight();

11: **end if**/\*如果共同最大前缀满足长度要求或者属于前缀词库, 则合并单位名称并且合并对应的近似度权重; 如果不满足该条件, 则保持单位名称和近似度权重不变; 没有共同最大前缀的单位名称及其近似度权重保持不变\*/

12: deleteTree(); //删除前缀树

/\*聚合不同权重间相同或互为别名的单位名称并合并近似度权重, 如果不同权重之间的单位名称相同或者互为别名, 则聚合单位名称和近似度权重\*/

13: **end for**

14: removeRedundance( entityKeyValue);

15: readSaveSimilarWords( simiBuf, similarwords); /\* simiBuf 保存别名词库的变量, similarwords 是别名词库文件 \*/

---

[0101]

[0102] 根据本发明的一个实施例, 还提供一种社交网络用户虚实映射系统, 如图 4 所示, 社交网络用户虚实映射系统 1 包括: 社交网络用户地理特征获取设备 11, 社交网络用户地

理信息处理设备 12 和社交网络用户单位名称计算推送设备 13。以下分别对系统 1 中的各个设备进行详细描述。

[0103] 一、社交网络用户地理特征获取设备 11

[0104] 社交网络用户地理特征获取设备 11 包括两个部分,分别是消息内容和地理位置信息获取装置 111 和周边单位名称列表获取装置 112。消息内容和地理位置信息获取装置 111 用于根据社交网络用户的唯一标识获取该社交网络用户的数据,这些数据可以包括该社交网络用户发布的消息内容(例如微博用户发表的博文内容、发言内容等)、社交网络用户的签到信息、社交网络用户发布消息的地理位置信息等。消息内容和地理位置信息获取装置 111 接收社交网络用户的唯一标识作为输入,请求社交网络服务器 2(例如微博服务器),并且从社交网络服务器 2 抓取该社交网络用户发布的消息内容和地理位置信息等。在一个实施例中,在请求社交网络服务器 2 时,还可以配置一定的参数,例如,配置获取指定时间段(如工作时间段)内的社交网络用户数据、配置获取地理位置信息的精度,以及配置容错次数等。

[0105] 在一个实施例中,消息内容和地理位置信息获取装置 111 在成功获取社交网络用户的消息内容和地理位置信息后,统一该地理位置信息的精确度,并按出现次数作为指标降序排序每个地理位置信息,选择前 N 个作为频繁地理位置信息,这里精确度和 N 都可以由用户设置。

[0106] 周边单位名称列表获取装置 112 用于将消息内容和地理位置信息获取装置 111 获得的地理位置信息,或者 N 个频繁地理位置信息逐条发送给社交网络服务器 2,定位该地理位置信息对应的地理位置并获取该地理位置周边的单位名称列表。

[0107] 二、社交网络用户地理信息处理设备 12

[0108] 社交网络用户地理信息处理设备 12 用于对周边单位名称列表中的每个单位名称进行分词,得到该单位名称的分词结果。在一个实施例中,社交网络用户地理信息处理设备 12 可先使用中文分词再使用二元组分词进行单位名称的分词。

[0109] 其中,中文分词过程可参考用户词典 33,以使分词结果更符合客户需要。此外,社交网络用户地理信息处理设备 12 在分词过程中,如果发现得到的某个分词在别名词库 32 中拥有别名,还将该分词的别名也加入分词结果中。

[0110] 三、社交网络用户单位名称计算推送设备 13

[0111] 社交网络用户单位名称计算推送设备 13 包括 3 个部分,分别是单位名称近似度权重计算装置 131,单位名称聚合装置 132 和单位名称推荐装置 133。其中,单位名称近似度权重计算装置 131 用于计算分词后的周边单位名称列表中每个单位名称的近似度权重。单位名称聚合装置 132 用于对单位名称及其近似度权重进行聚合。单位名称推荐装置 133 用于根据单位名称聚合装置 132 的聚合结果按近似度权重对单位名称进行降序排列,选择前 N 个单位名称进行结果推送。

[0112] 在一个实施例中,单位名称近似度权重计算装置 131 用于将单位名称分词结果匹配消息内容,完全匹配的近似度权重高而基本匹配的近似度权重低。如果完全匹配成功,则可以根据公式(1)来计算该单位名称的近似度权重。如果仅是基本匹配成功,则可以将该单位名称的分词结果中除该单位名称的全称外的每个分词与消息内容进行匹配,将每个分词的匹配成功次数之和作为该单位名称的近似度权重。

[0113] 在一个实施例中，单位名称聚合装置 132 用于将周边单位名称列表中具有共同最大前缀、相同或者互为别名的单位名称进行聚合，并合并它们的近似度权重。其中，单位名称聚合装置 132 可采用别名词库 32 来去除单位名称列表中的重复项，并且采用前缀词库 31 来确认聚合获得的共同最大前缀是否满足要求。其中，共同最大前缀必须是前缀词库中所包含的前缀词或者满足用户设定的长度才有效。

[0114] 应该注意到并理解，在不脱离后附的权利要求所要求的本发明的精神和范围的情况下，能够对上述详细描述的本发明做出各种修改和改进。因此，要求保护的技术方案的范围不受所给出的任何特定示范教导的限制。

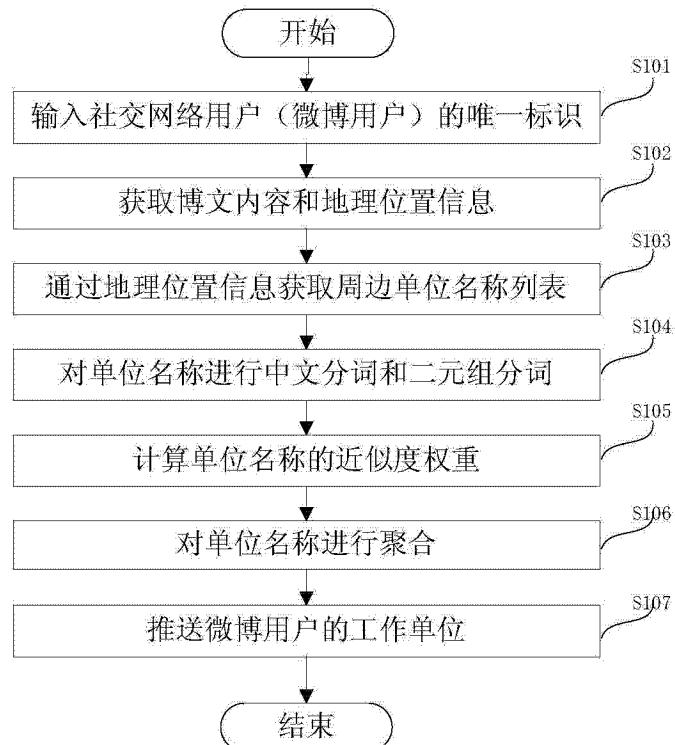


图 1

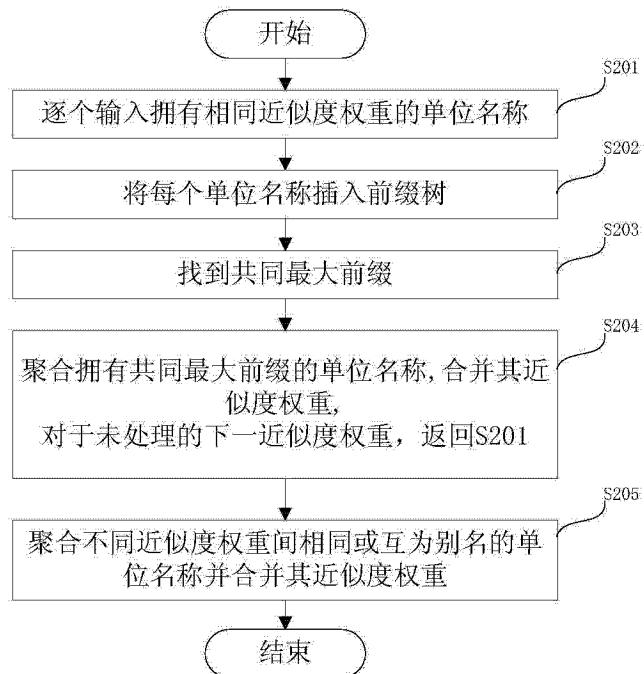


图 2

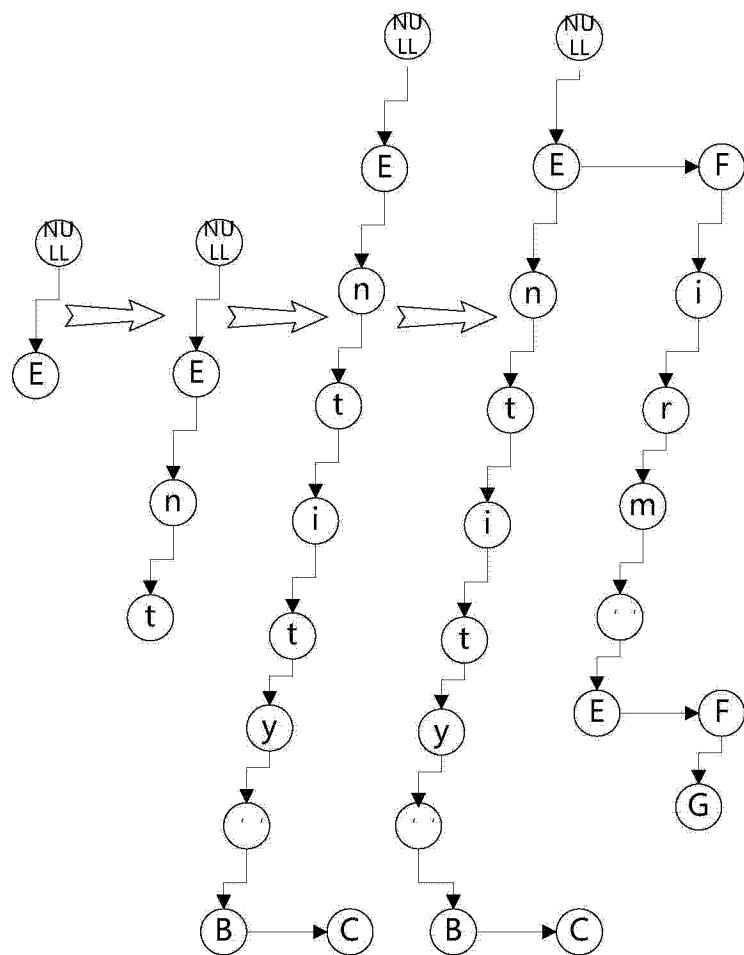


图 3

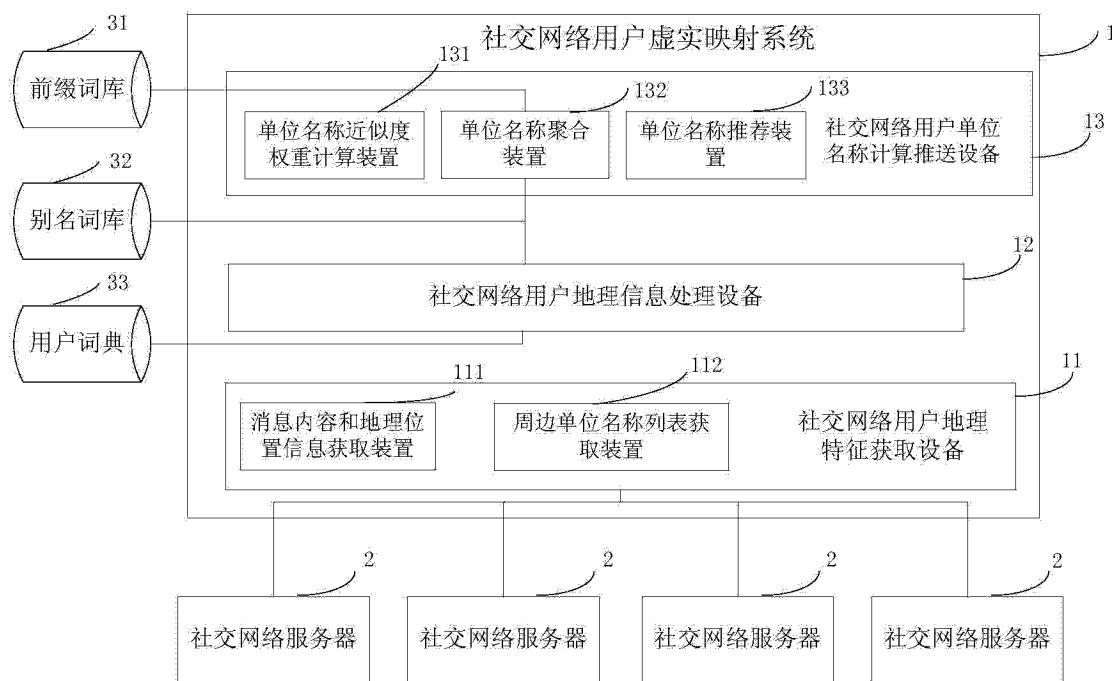


图 4