(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 103914494 B
(45)授权公告日 2017.05.17

(21)申请号 201310008156.X

(22)申请日 2013.01.09

(65)同一申请的已公布的文献号
申请公布号 CN 103914494 A

(43)申请公布日 2014.07.09

(73)专利权人 北大方正集团有限公司
地址 100871 北京市海淀区成府路298号方正大厦5层
专利权人 北京大学
北京北大方正电子有限公司

(72)发明人 赵立永 于晓明 杨建武 郑妍

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
代理人 赵婷婷

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 21/31(2013.01)

(56)对比文件
CN 101187920 A,2008.05.28,
US 7716225 B1,2010.05.11,
CN 102289522 A,2011.12.21,
CN 102355664 A,2012.02.15,

审查员 刘申

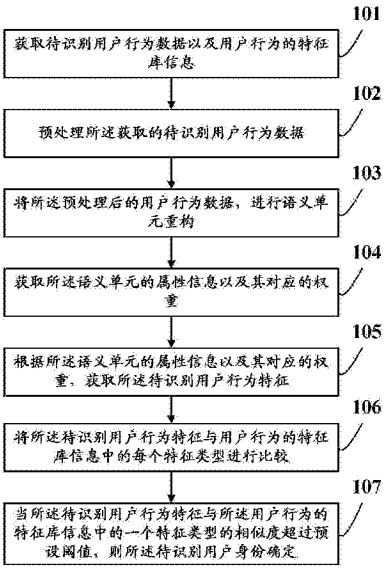
权利要求书2页 说明书6页 附图4页

(54)发明名称

一种微博用户身份识别方法及系统

(57)摘要

本发明提供一种微博用户身份识别方法及系统，所述方法包括：获取待识别用户行为数据以及用户行为的特征库信息；预处理所述获取的待识别用户行为数据；将所述预处理后的用户行为数据，进行语义单元重构；获取所述语义单元的属性信息以及其对应的权重；根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。采用本发明提供的微博用户身份识别方法及系统可以有效提高微薄用户身份识别的准确性及实时性。



CN 103914494 B

1.一种微博用户身份识别方法，其特征在于，包括：

获取待识别用户行为数据以及用户行为的特征库信息；

预处理所述获取的待识别用户行为数据；

将所述预处理后的用户行为数据，进行语义单元重构；

获取所述语义单元的属性信息以及其对应的权重；

根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；

将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；

当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定；

在所述待识别用户身份确定之后，所述方法还包括：

获取所述确定用户身份的待识别用户的至少一个语义单元以及对应所述用户身份的用户类型信息；

比较所述语义单元与所述用户身份的用户类型信息，给出所述各个语义单元与所述用户身份的用户类型信息的相似度；

按照所述相似度由大到小的顺序，对所述语义单元进行排序；

获取相似度前top-n个语义单元作为该类型用户的行为特征；

将所述用户的行为特征添加到所述用户行为的特征库的对应类别中。

2.根据权利要求1所述的微博用户身份识别方法，其特征在于，在获取待识别用户行为数据以及用户行为的特征库信息的步骤之前，该方法还包括：

获取已知用户行为数据；

预处理所述获取已知用户行为数据；

将所述预处理后的用户行为数据，进行语义单元重构；

获取所述语义单元的属性信息以及其对应的权重；

根据所述语义单元的属性信息以及其对应的权重，获取所述已知用户行为特征；

将所述获取所述已知用户行为特征，按照类别存储在所述用户行为的特征库中。

3.根据权利要求1所述的微博用户身份识别方法，其特征在于，所述行为特征至少包括一个语义单元；所述语义单元属性信息至少包括：索引值，字符信息，词性，词频和文档频率；所述语义单元至少包括一个词；所述词的属性信息包括：词的索引，词频，文档频率，IDF值，权值。

4.根据权利要求3所述的微博用户身份识别方法，其特征在于，所述预处理步骤主要包括：行为数据筛选、拼写纠正、分词和词性标注。

5.一种微博用户身份识别系统，其特征在于，包括：

信息获取单元，用于获取待识别用户行为数据以及用户行为的特征库信息；

预处理单元，用于预处理所述获取的待识别用户行为数据；

语义单元重构单元，用于将所述预处理后的用户行为数据，进行语义单元重构；

属性及权重信息获取单元，还用于获取所述语义单元的属性信息以及其对应的权重；

行为特征抽取单元，用于根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；

比较单元，用于将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类

型进行比较；

身份确定单元，用于当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定；

所述系统还包括：信息反馈单元，用于获取所述确定用户身份的待识别用户的至少一个语义单元以及对应所述用户身份的用户类型信息；比较所述语义单元与所述用户身份的用户类型信息，给出所述各个语义单元与所述用户身份的用户类型信息的相似度；按照所述相似度由大到小的顺序，对所述语义单元进行排序；获取相似度前top-n个语义单元作为该类型用户的行为特征；将所述用户的行为特征添加到所述用户行为的特征库的对应类别中。

6.根据权利要求5所述的微博用户身份识别系统，其特征在于，该系统还包括：用户行为的特征库构建单元，用于获取已知用户行为数据；预处理所述获取已知用户行为数据；将所述预处理后的用户行为数据，进行语义单元重构；获取所述语义单元的属性信息以及其对应的权重；根据所述语义单元的属性信息以及其对应的权重，获取所述已知用户行为特征；将所述获取所述已知用户行为特征，按照类别存储在所述用户行为的特征库中。

7.根据权利要求5所述的微博用户身份识别系统，其特征在于，所述行为特征至少包括一个语义单元；所述语义单元属性信息至少包括：索引值，字符信息，词性，词频和文档频率；所述语义单元至少包括一个词；所述词的属性信息包括：词的索引，词频，文档频率，IDF值，权值。

8.根据权利要求7所述的微博用户身份识别系统，其特征在于，所述预处理步骤主要包括：行为数据筛选、拼写纠正、分词和词性标注。

一种微博用户身份识别方法及系统

技术领域

[0001] 本发明涉及计算机信息处理技术领域，尤其涉及一种微博用户身份识别方法及系统。

背景技术

[0002] 随着web技术的发展和微博的出现，越来越多的用户加入到互联网中，成为虚拟社会中的一员，促进了信息传播方式的变革，提高了信息传播的效率。然而，微薄用户身份的识别作为微薄后台维护的重要组成部分，其识别过程主要通过微薄用户在网络注册、存储的数据信息进行用户身份识别。例如：从网站获取待识别用户访问网站的日志、临时信息及注册信息来实现用户身份识别；或者，通过中文文本分类方法进行微薄用户身份识别。

[0003] 但是，在现有的微薄用户身份识别过程中，发明人发现技术至少存在如下问题：

[0004] 现有技术中通过网站获取待识别用户访问网站的日志、临时信息及注册信息来实现用户身份识别的过程，由于用户身份识别过程所依据的数据主要依靠从网站获取用户注册信息以及该用户的日志及临时信息，从而使得数据获取较为困难，且准确性不高。

[0005] 现有技术中采用中文文本分类的方法虽然可以实现微薄用户身份识别，但是，无法满足当前微博用户身份识别的准确性及实时性。

发明内容

[0006] 针对现有技术中存在的缺陷，本发明的目的是提出一种准确性高，实时性强的微博用户身份识别方法及系统。

[0007] 本发明提供一种微博用户身份识别方法，包括：

[0008] 获取待识别用户行为数据以及用户行为的特征库信息；

[0009] 预处理所述获取的待识别用户行为数据；

[0010] 将所述预处理后的用户行为数据，进行语义单元重构；

[0011] 获取所述语义单元的属性信息以及其对应的权重；

[0012] 根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；

[0013] 将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；

[0014] 当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。

[0015] 本发明还提供一种微博用户身份识别系统，包括：

[0016] 信息获取单元，用于获取待识别用户行为数据以及用户行为的特征库信息；

[0017] 预处理单元，用于预处理所述获取的待识别用户行为数据；

[0018] 语义单元重构单元，用于将所述预处理后的用户行为数据，进行语义单元重构；

[0019] 属性及权重信息获取单元，还用于获取所述语义单元的属性信息以及其对应的权

重；

[0020] 行为特征抽取单元，用于根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；

[0021] 比较单元，用于将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；

[0022] 身份确定单元，用于当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。

[0023] 本发明提供的微博用户身份识别方法及系统，通过获取待识别用户行为数据以及用户行为的特征库信息；预处理所述获取的待识别用户行为数据；将所述预处理后的用户行为数据，进行语义单元重构；获取所述语义单元的属性信息以及其对应的权重；根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。采用本发明提供的微博用户身份识别方法及系统可以有效提高微薄用户身份识别的准确性及实时性。

附图说明

[0024] 图1为本发明实施例提供的一种微博用户身份识别方法的流程图；

[0025] 图2为本发明提供的一种微博用户身份识别方法中用户行为的特征库的构建流程图；

[0026] 图3为本发明提供的一种微博用户身份识别方法中更新用户行为的特征库的流程图；

[0027] 图4为本发明实施例提供的一种微博用户身份识别系统结构示意图；

[0028] 图5为本发明实施例提供的另一种微博用户身份识别系统结构示意图；

[0029] 图6为本发明实施例提供的一种微博用户身份识别方法中语义单元属性信息数据结构示意图。

具体实施方式

[0030] 下面结合附图对本发明实施例提供的一种微博用户身份识别方法及系统进行详细描述。

[0031] 如图1所示，为本发明实施例子提供的一种微博用户身份识别方法，该方法包括：

[0032] 101：获取待识别用户行为数据以及用户行为的特征库信息；

[0033] 102：预处理所述获取的待识别用户行为数据；所述预处理主要包括行为数据筛选、拼写纠正、分词和词性标注。

[0034] 103：将所述预处理后的用户行为数据，进行语义单元重构；所述语义单元重构是在预处理的基础上应用词性信息进行词粘连的方法，通过合并特定的词，来构建包含更丰富语义的语义单元(词串)。

[0035] 104：获取所述语义单元的属性信息以及其对应的权重；其中，所述语义单元的属性信息是指统计每个语义单元的词频和文档频率；所述语义单元的权重则采用TFIDF函数

来实现用户行为特征的权值计算，实现用户行为特征的数值化。

[0036] 105：根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；所述待识别用户行为特征是指抽取最能代表用户行为的特征，并且特征项(即语义单元)具有很好的区分度，对于单个待识别用户主要采用词权重、词频、词性相结合的方法，根据词权重和词频进行关键词排序；根据停用词表过滤掉停用词或非停用词(满足词长大于最大长度或小于最小长度)；选取词性为“a”，“cw”，“v”，“j”，“ns”，“nr”，“nt”，“nz”或者包含“不”的词。

[0037] 106：将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；所述比较的过程进行用户分类，主要可以采用KNN算法，K值选取方法采用概率分布的方法，即相似的特征向量和特征向量空间之比。具体分类思路为：比较待识别用户和用户行为特征库信息中每个用户类别的相似度sim(u,C)，比较用户和每个类别中包含用户的相似度sim(u,Cui)，如果sim(u,C)大于经验阈值，或者多数sim(u,Cui)大于经验阈值，则认为用户和该类别存在相关性，选取相似度最大的用户类别来确定用户身份。

[0038] 采用调整余弦相似度的测量方法计算特征向量之间的相似度，具体步骤如下：

[0039] (1)对于特征向量库中每一个特征向量，计算与该用户特征向量的相似度；

[0040] (2)进行向量对齐操作，对于向量v1和v2，求其所有特征项的并集C(v1,v2)，然后将v1和v2映射到C上，得到新的向量v1’和v2’；

[0041] (3)采用调整余弦相似度计算公式计算v1’和v2’的相似度。

[0042] 107：当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。

[0043] 如图2所示，为本发明实施例子提供的一种微博用户身份识别方法中构建用户行为的特征库流程，该构建方法包括：

[0044] 201：获取已知用户行为数据；具体的讲，就是获取已知用户行为数据，即训练数据；该训练数据用于构建用户行为的特征库。

[0045] 202：预处理所述获取已知用户行为数据；具体的讲，就是按照已知用户的不同身份，对训练数据(即已知用户数据)进行标注，对相同身份的每个用户的微博消息进行过滤，过滤的方法是比较消息的长度和观测值θ(通过对大量微博消息统计分析，10个字符以内的微博消息包含较少或没有语义信息，因此本系统中θ＝10III)之间的大小关系，如果长度小于观测值，则将微博作为噪声过滤掉。拼写检查主要根据拼写常见错误对照表进行拼写错误校正。利用分词和词性标注工具进行分词及词性标注，处理后每个词都包含词字符串信息和词性，分词和词性标注的工具均来自已知技术，此处不再赘述。

[0046] 203：将所述预处理后的用户行为数据，进行语义单元重构；所述语义单元重构具体为：由于长词串相对于短词串包含更多语义信息，具有更强的表达能力，所以语义单元重构就是在步骤201处理结果的基础上，通过特定的规则对相邻的特定词进行词粘连，进而产生更长的语义串。该步骤要处理的相邻词包括“ns”地名，“nr”人名，“nt”机构名，“nz”专有名词和“j”简称等，处理的规则是组合第一次出现该类型词和最后一次出现该类型词之间的所有词。标注粘连后的词串词性为“cw”，在特征选择和权值计算时，该类词更重要。

[0047] 204：获取所述语义单元的属性信息以及其对应的权重；

[0048] 其中，所述获取语义单元的属性信息，是基于步骤201和步骤202，为所述语义单元

进行统一编号，建立微博一语义单元索引向量，按用户统计语义单元的属性信息，包括词频和文档频率，为单个用户行为特征提取做准备，按照相同身份用户进行词频和文档频率统计，为相同身份类别的类别行为特征提取做准备，处理结果信息保存到如图6所示的数据结构中。

[0049] 所述获取所述语义单元的权重的具体过程为：

[0050] 首先，根据自然语言处理领域中常用的停用词表过滤掉停用词，并过滤掉词频小于经验阈值且词性为非包含“n”、“cw”的语义单元。其次，采用基于TF-IDF权值计算方法，计算每个语义单元的权值，对于特定类型的语义单元赋予更高的权值，具体方法为，对于词性为“nr”人名，如是式(2)所示，加权系数α＝2.0，对于词性为“cw”粘连词，如是式(3)所示，加权系数为β＝1.5，具体权值计算公式为：

[0051] $weight1 = TF \mid \log_2 IDF$ (1)

[0052] $weight2 = 2.0 \mid TF \mid \log_2 IDF$ (2)

[0053] $weight3 = 1.5 \mid TF \mid \log_2 IDF$ (3)

[0054] 205：根据所述语义单元的属性信息以及其对应的权重，获取所述已知用户行为特征；具体获取过程为：

[0055] 对于所述获取的已知用户身份的训练数据主要采用卡方统计、词性、词频相结合的方法；首先计算每个语义单元相当于用户类别的卡方值，按照卡方值对语义单元进行排序；过滤掉长度等于1，且词性为非nr的词；根据停用词表过滤掉停用词或非停用词(满足词长大于最大长度或小于最小长度)；选取词性为“a”，“cw”，“v”，“j”，“ns”，“nr”，“nt”，“nz”或者包含“不”的词；上述信息均不能区分时，选择词频较大的语义单元。

[0056] 为了控制分类过程中特征的维数，设定选取语义单元的上限值θ＝200。

[0057] 206：将所述获取所述已知用户行为特征，按照类别存储在所述用户行为的特征库中。

[0058] 如图3所示，为本发明实施例子提供的一种微博用户身份识别方法中更新用户行为的特征库的流程；该流程包括：

[0059] 301：获取所述确定用户身份的待识别用户的至少一个语义单元以及对应所述用户身份的用户类型信息；

[0060] 302：比较所述语义单元与所述用户身份的用户类型信息，给出所述各个语义单元与所述用户身份的用户类型信息的相似度；该步骤可以采用卡方统计方法，计算语义单元与用户类别的卡方值，通过所述获取的卡方值来评价相关性。

[0061] 303：按照所述相似度由大到小的顺序，对所述语义单元进行排序；

[0062] 304：获取相似度前top-n个语义单元作为该类型用户的行为特征；

[0063] 305：将所述用户的行为特征添加到所述用户行为的特征库的对应类别中。

[0064] 需要说明的是，以上所述的实施例子中所述行为特征至少包括一个语义单元；所述语义单元属性信息如图6所示，至少包括：索引值，字符信息，词性，词频和文档频率；所述语义单元至少包括一个词；所述词的属性信息包括：词的索引，词频，文档频率，IDF值，权值。

[0065] 所述预处理步骤主要包括：行为数据筛选、拼写纠正、分词和词性标注。

[0066] 如图4所示，为本发明实施例子提供的一种微博用户身份识别系统，该系统包括：

[0067] 信息获取单元401，用于获取待识别用户行为数据以及用户行为的特征库信息；

[0068] 预处理单元402，用于预处理所述获取的待识别用户行为数据；

[0069] 语义单元重构单元403，用于将所述预处理后的用户行为数据，进行语义单元重构；

[0070] 属性及权重信息获取单元404，还用于获取所述语义单元的属性信息以及其对应的权重；

[0071] 行为特征抽取单元405，用于根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；

[0072] 比较单元406，用于将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；

[0073] 身份确定单元407，用于当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。

[0074] 需要说明的是，如图5所示，该系统还包括：用户行为的特征库构建单元501和/或信息反馈单元502；

[0075] 所述用户行为的特征库构建单元501，用于获取已知用户行为数据；预处理所述获取已知用户行为数据；将所述预处理后的用户行为数据，进行语义单元重构；获取所述语义单元的属性信息以及其对应的权重；根据所述语义单元的属性信息以及其对应的权重，获取所述已知用户行为特征；将所述获取所述已知用户行为特征，按照类别存储在所述用户行为的特征库中。

[0076] 所述信息反馈单元502，用于获取所述确定用户身份的待识别用户的至少一个语义单元以及对应所述用户身份的用户类型信息；比较所述语义单元与所述用户身份的用户类型信息，给出所述各个语义单元与所述用户身份的用户类型信息的相似度；按照所述相似度由大到小的顺序，对所述语义单元进行排序；获取相似度前top-n个语义单元作为该类型用户的行为特征；将所述用户的行为特征添加到所述用户行为的特征库的对应类别中。

[0077] 以上所述行为特征至少包括一个语义单元；所述语义单元属性信息至少包括：索引值，字符信息，词性，词频和文档频率；所述语义单元至少包括一个词；所述词的属性信息包括：词的索引，词频，文档频率，IDF值，权值。

[0078] 所述预处理步骤主要包括：行为数据筛选、拼写纠正、分词和词性标注。

[0079] 本发明提供的微博用户身份识别方法及系统，通过获取待识别用户行为数据以及用户行为的特征库信息；预处理所述获取的待识别用户行为数据；将所述预处理后的用户行为数据，进行语义单元重构；获取所述语义单元的属性信息以及其对应的权重；根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征；将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较；当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定。采用本发明提供的微博用户身份识别方法及系统可以有效提高微薄用户身份识别的准确性及实时性。

[0080] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，所述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，包括如下步骤：(方法的步骤)，所述的存储介质，如：ROM/RAM、磁碟、光盘

等。

[0081] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

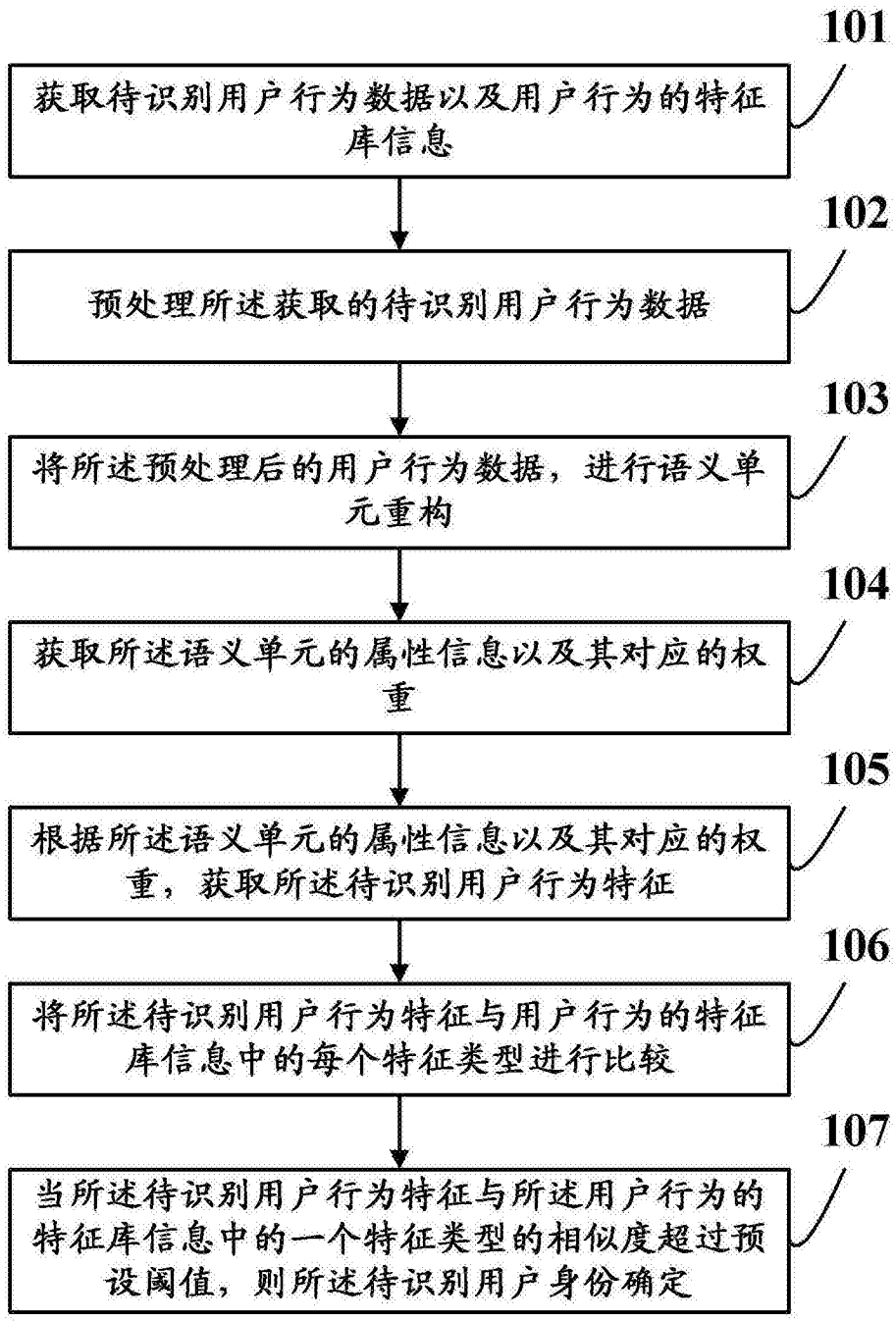

101
获取待识别用户行为数据以及用户行为的特征库信息
102
预处理所述获取的待识别用户行为数据
103
将所述预处理后的用户行为数据，进行语义单元重构
104
获取所述语义单元的属性信息以及其对应的权重
105
根据所述语义单元的属性信息以及其对应的权重，获取所述待识别用户行为特征
106
将所述待识别用户行为特征与用户行为的特征库信息中的每个特征类型进行比较
107
当所述待识别用户行为特征与所述用户行为的特征库信息中的一个特征类型的相似度超过预设阈值，则所述待识别用户身份确定

图1

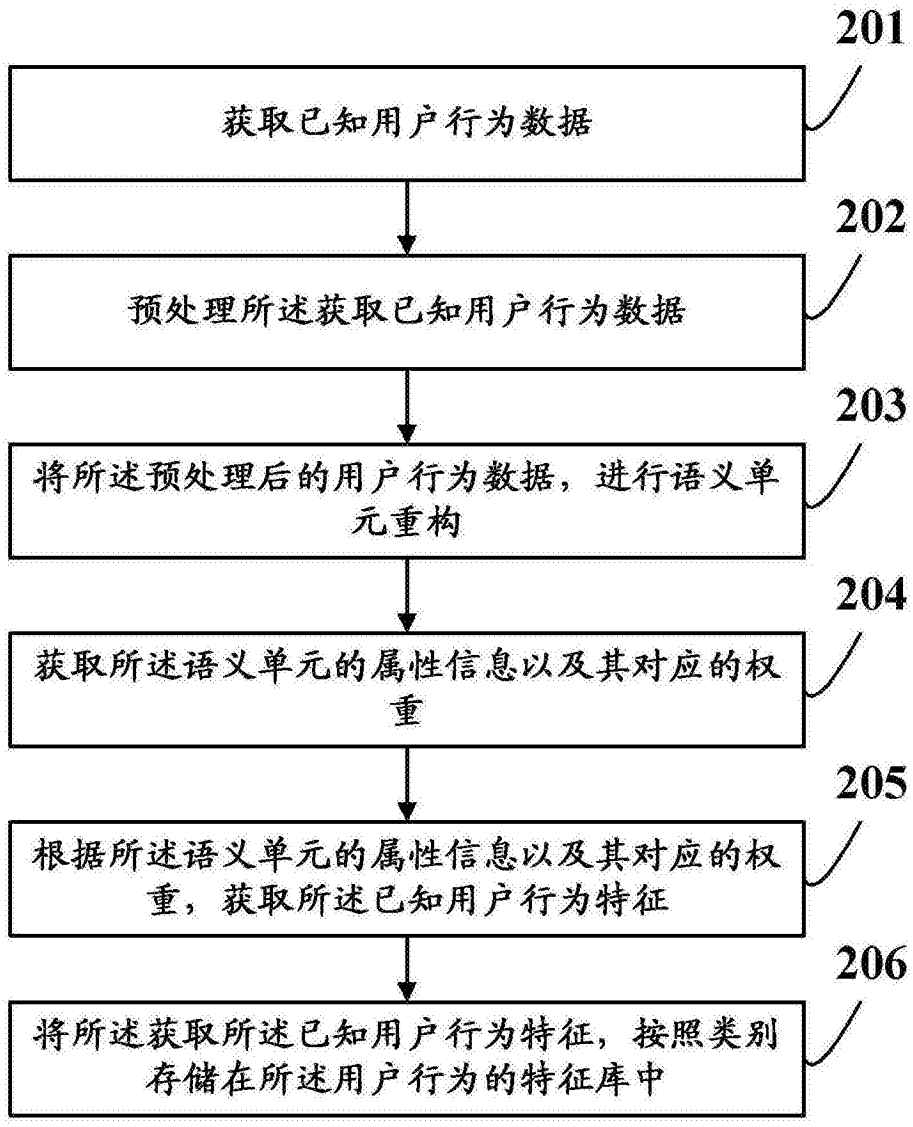

201
获取已知用户行为数据
202
预处理所述获取已知用户行为数据
203
将所述预处理后的用户行为数据，进行语义单元重构
204
获取所述语义单元的属性信息以及其对应的权重
205
根据所述语义单元的属性信息以及其对应的权重，获取所述已知用户行为特征
206
将所述获取所述已知用户行为特征，按照类别存储在所述用户行为的特征库中

图2

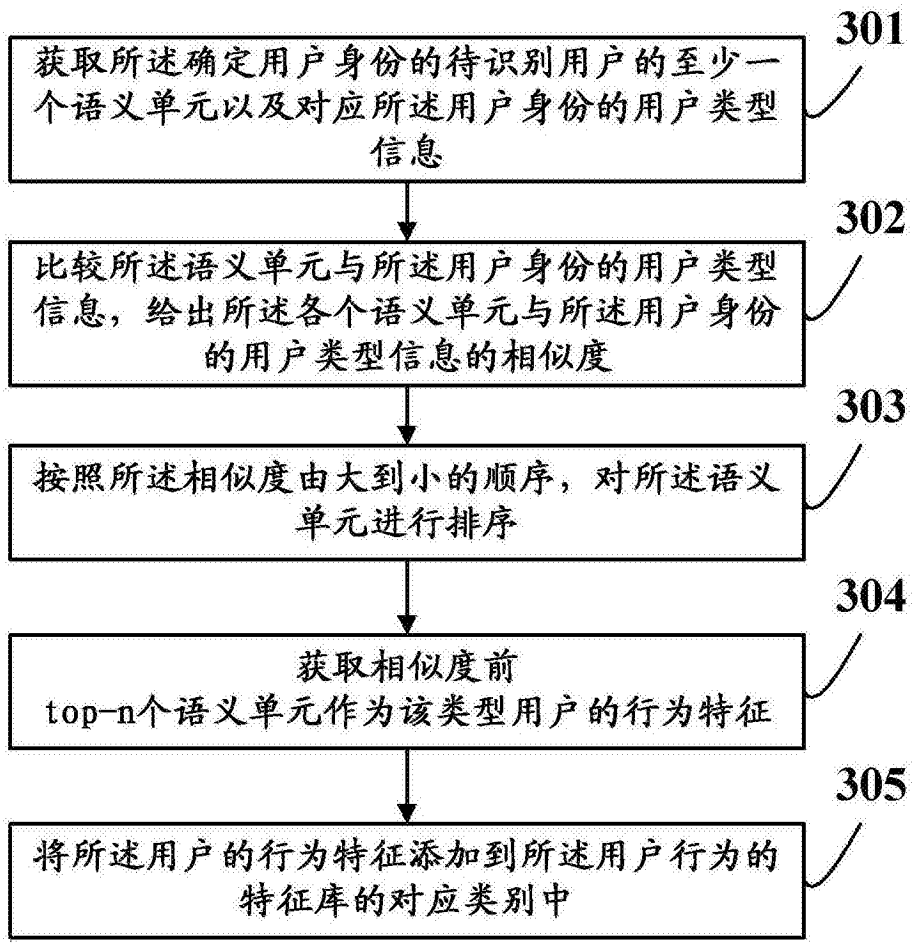

301
获取所述确定用户身份的待识别用户的至少一个语义单元以及对应所述用户身份的用户类型信息
302
比较所述语义单元与所述用户身份的用户类型信息，给出所述各个语义单元与所述用户身份的用户类型信息的相似度
303
按照所述相似度由大到小的顺序，对所述语义单元进行排序
304
获取相似度前top-n个语义单元作为该类型用户的行为特征
305
将所述用户的行为特征添加到所述用户行为的特征库的对应类别中

图3

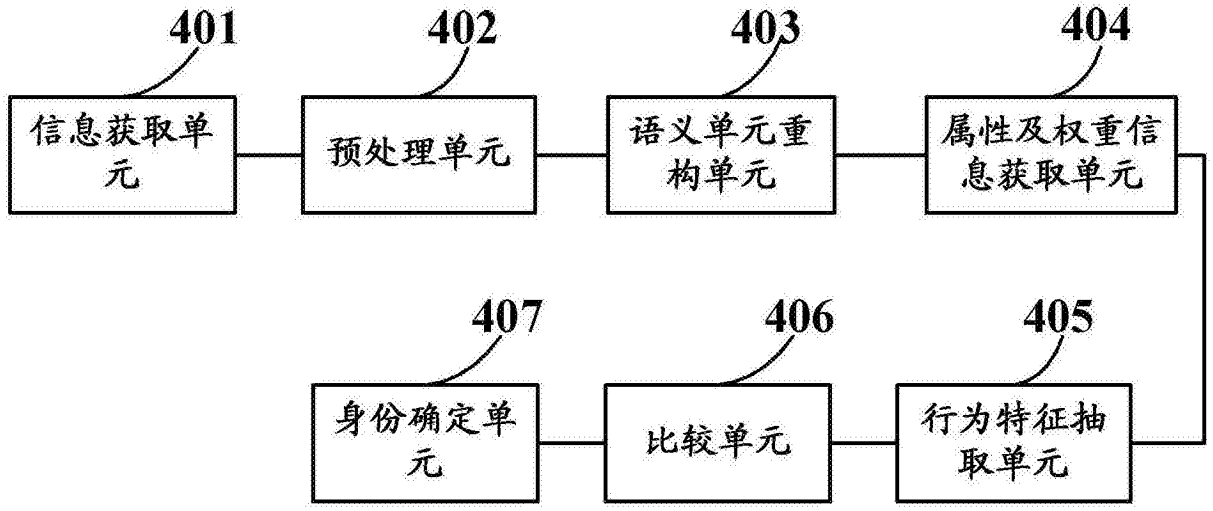

401
信息获取单元
402
预处理单元
403
语义单元重构单元
404
属性及权重信息获取单元
405
行为特征抽取单元
406
比较单元
407
身份确定单元

图4

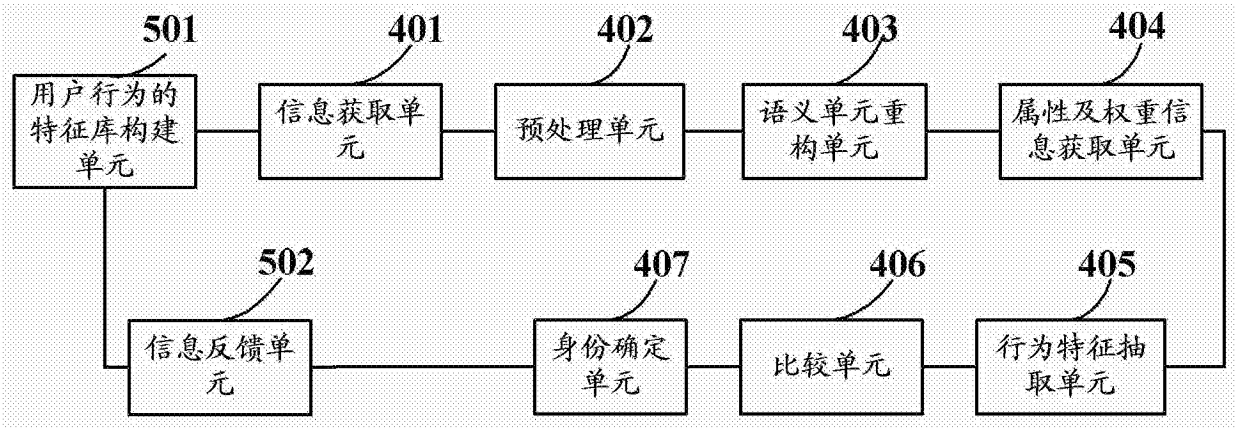

501
用户行为的特征库构建单元
401
信息获取单元
402
预处理单元
403
语义单元重构单元
404
属性及权重信息获取单元
502
信息反馈单元
407
身份确定单元
406
比较单元
405
行为特征抽取单元

图5


索引值
字符信息
词性
词频
文档频率

图6