(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104598452 A
(43)申请公布日 2015.05.06

(21)申请号 201310526980.4

(22)申请日 2013.10.30

(71)申请人 北京思博途信息技术有限公司
地址 102218 北京市昌平区东小口镇中东路398号中煤建设集团大厦1号楼4层秒针系统

(72)发明人 丁若谷 陈家耀 冯是聪 吴明辉

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 王丹 栗若木

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

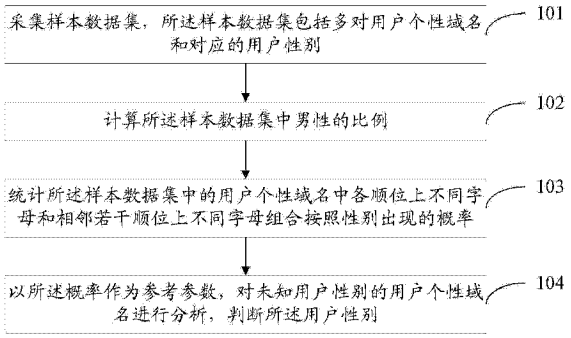
权利要求书3页 说明书14页 附图2页

(54)发明名称

用户性别分析方法和装置

(57)摘要

本发明提供了一种用户性别分析方法和装置。涉及数据分析领域；解决了现有分析方式不适用于个性域名和姓名关联较弱的场合的问题。该方法包括：采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。本发明提供的技术方案适用于数据分析，实现了基于自动化算法的用户性别分析。



CN 104598452 A

1. 一种用户性别分析方法，其特征在于，包括：

采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；

统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；

以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。

2. 根据权利要求 1 所述的用户性别分析方法，其特征在于，所述统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率的步骤之前，还包括：

计算所述样本数据集中男性的比例。

3. 根据权利要求 1 所述的用户性别分析方法，其特征在于，统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上字母组合按照性别出现的概率包括：

步骤 a：取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别；

步骤 b：对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数；

步骤 c：对所述样本数据集中的全部用户个性域名进行如步骤 a 至 b 的处理，直至所述样本数据集遍历完成；

步骤 d：统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

4. 根据权利要求 3 所述的用户性别分析方法，其特征在于，统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率具体为：

根据表达式

$$P(n-gram\text{对应男性})=\frac{n-gram\text{对应男性频率}}{n-gram\text{对应男性频率}+n-gram\text{对应女性频率}},$$

分别计算各顺位上各字母和相邻若干顺位上各字母组合对应为男性的概率；其中，等式左侧的 P(n-gram 对应男性）为长度为 n 的相邻若干顺位上的字母组合对应为男性的概率，n 为 1 时 P(n-gram 对应男性）为单一顺位上的字母对应为男性的概率；等式右侧的 n-gram 对应男性频率为单一顺位上的字母或长度为 n 的相邻若干顺位上的字母组合对应为男性的次数，n-gram 对应女性频率为单一顺位上的字母或长度为 n 的相邻若干顺位上的字母组合对应为女性的次数。

5. 根据权利要求 1 所述的用户性别分析方法，其特征在于，以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别包括：

步骤 a：获取所述未知用户性别的用户个性域名的长度，记为 k；

步骤 b :按照表达式

$$P(url\text{对应用户性别为男性}) = \sum_{h=1,2,\ldots,n} w_h \left(\sum_{i=1,2,\ldots,k} \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性}) \right),$$

计算所述用户的性别为男性的概率，其中，url 表示个性域名中用户指定的部分；substr(url,j,i) 表示 url 中第 j 位字符开始长度为 i 的相邻若干顺位上的字母组合构成的子字符串，i 为 1 时为单一顺位上的字母构成的子字符串 ；n 表示 substr(url,j,i) 的个数 ；w_h 表示该字母或字母组合的权重 ；P(substr(url,j,i) 在样本数据集中对应男性）表示上述子字符串上的字母或字母组合对应的男性概率 ；

步骤 c :比较步骤 b 中的计算结果与所述样本数据集中男性的比例 ；

步骤 d :在步骤 b 中的计算结果大于等于步骤 c 计算得到的比例时，判定所述未知性别用户的性别为男性。

6. 根据权利要求 5 所述的用户性别分析方法，其特征在于，所述步骤 d 之后，还包括 ：

步骤 e :在步骤 b 中的计算结果小于步骤 c 计算得到的比例时，判定所述未知性别用户的性别为女性。

7. 一种用户性别分析装置，其特征在于，包括 ：

采样模块，用于采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别 ；

参考参数计算模块，用于统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率 ；

分析模块，用于以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。

8. 根据权利要求 7 所述的用户性别分析装置，其特征在于，该装置还包括 ：

参考比例计算模块，用于计算所述样本数据集中男性的比例。

9. 根据权利要求 8 所述的用户性别分析装置，其特征在于，所述参考参数计算模块包括 ：

性别提取单元，用于取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别 ；

计数单元，用于对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数 ；

统计单元，用于对所述样本数据集中的全部用户个性域名进行所述计数单元的处理，直至所述样本数据集遍历完成，统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

10. 根据权利要求 9 所述的用户性别分析装置，其特征在于，所述统计单元计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率具体为 ：

根据表达式

$$P(n-gram对应男性)=\frac{n-gram对应男性频率}{n-gram对应男性频率+n-gram对应女性频率},$$

分别计算各顺位上各字母和相邻若干顺位上各字母组合对应的男性概率，其中，P(n-gram对应男性）为长度为n的相邻若干顺位上一字母组合对应为男性的概率，n为1时P(n-gram对应男性）为单一顺位上一字母对应为男性的概率，n-gram对应男性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为男性的次数，n-gram对应女性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为女性的次数。

11. 根据权利要求8所述的用户性别分析装置，其特征在于，所述分析模块包括：

域名长度获取单元，用于获取所述未知用户性别的用户个性域名的长度，记为k；

概率计算单元，用于按照表达式

$$P(url对应用户性别为男性)$$

$$=\sum_{h=1,2,\ldots,n} w_h\Big(\sum_{i=1,2,\ldots,k}\ \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)在样本数据集中对应男性)\Big)$$

计算所述用户的性别为男性的概率，其中，url表示个性域名中用户指定的部分，substr(url，j，i）表示url中第j位字符开始长度为i的相邻字符构成的子字符串，n表示substr(url，j，i）的个数，w_h表示该字母或字母组合的权重，P(substr(url,j,i）在样本数据集中对应男性）表示url中第j位字符或第j位字符开始长度为i的相邻字符构成的子字符串上的字母或字母组合对应的男性概率；

比较单元，用于比较概率计算单元的计算结果与参考比例计算模块计算得到的比例；

判定单元，用于在所述比较单元比较的结果为比较概率计算单元的计算结果大于等于参考比例计算模块计算得到的比例时，判定所述未知性别用户的性别为男性。

12. 根据权利要求11所述的用户性别分析装置，其特征在于，

所述判定单元，还用于在所述比较单元比较的结果为比较概率计算单元的计算结果小于所述参考比例计算模块计算得到的比例时，判定所述未知性别用户的性别为女性。

用户性别分析方法和装置

技术领域

[0001] 本发明涉及数据分析领域，尤其涉及一种用户性别分析方法和装置。

背景技术

[0002] 在互联网环境下，用户的性别是一项十分重要的信息。根据用户的性别，互联网内容提供者可以向不同用户展现不同的内容。例如，男性用户可能相比女性用户对电子竞技更感兴趣，而女性用户可能相比男性用户对时尚服饰更感兴趣。在这种情况下，如果用户的性别得到识别，互联网广告提供商就可以为男性用户展示电子竞技的广告，为女性用户展示时尚服饰的广告，从而使得广告更有针对性，取得更好的广告效果。

[0003] 对于注册博客、微博或其他社交网站的用户来说，很多服务提供商都会在用户完成必要的注册信息后，建议用户填写一些用户本身的属性，例如性别，年龄，工作状态，为自己设置个性域名等，而往往这些属性中在涉及到用户隐私的信息注册事项通常都是选择性填写事项，而非必须填写事项，这样，就导致了相当一部分用户选择不填写此类信息，例如用户为保护自己的信息不外漏，会选择不填写年龄，性别等，那么，对于数据分析机构或供应商本身来说，也就无法直接获取用户的性别信息。但对于不涉及隐私的选择性填写事项来说，被填写的成功率往往很高。例如，个性域名，服务提供商为了增加用户体验和亲和力，往往允许用户为自己的微博或个人空间主页设置具有代表用户本身性质的虚拟 url。用户可以将这些域名格式设置为自己的名字，或任意自己喜欢的数字，或字母组合，即时尚又方便。然而，出于人类自身的性别差异，在对个性域名的设置上，男性和女性往往本能的去设置一些代表自身属性的域名。例如，某用户可能注册一个个性域名 :http://weibo.com/basketballfans，其中 weibo.com 是微博服务提供商的域名，basketballfans 部分即用户选择的个性域名。那么，通过具有用户代表性的个性域名来推算出用户的性别信息，即不侵犯用户又可收集用户信息。

[0004] 在现有的技术中，最相似的技术是美国专利 7,447,996[1]。这一专利提出了一种软件模块，用于在即时通讯系统中根据不同的用户名推断用户的性别，根据不同的性别展示不同的虚拟形象。依赖于特定的人类行为学数据，即特定语言中的人名和性别之间的关系。例如，这一专利中提及，针对中文姓名，通过人类行为学数据库的检索，“xiuxiu”和“lili”更可能是女性的名字。

[0005] 人类行为学数据库并不适用于多种网络应用场景，尤其不适用于个性域名和姓名关联较弱的场合。个性域名的组成通常包括了超出常见姓名范畴的大量成分，这些成分很难通过人类行为学数据分析。例如，个性域名中可能包括“basketball”，即篮球 ；而可能将篮球放入个性域名的篮球爱好者中，男性可能占主导地位。如果将“篮球对应男性”这类数据加入数据库，所需的工作将极大增加，并且很难完备。

发明内容

[0006] 本发明提供了一种用户性别分析方法和装置，解决了现有分析方式不适用于个性

域名和姓名关联较弱的场合的问题。

[0007] 一种用户性别分析方法，包括：

[0008] 采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；

[0009] 统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；

[0010] 以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。

[0011] 优选的，所述统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率的步骤之前，还包括：

[0012] 计算所述样本数据集中男性的比例。

[0013] 优选的，统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上字母组合按照性别出现的概率包括：

[0014] 步骤 a：取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别；

[0015] 步骤 b：对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数；

[0016] 步骤 c：对所述样本数据集中的全部用户个性域名进行如步骤 a 至 b 的处理，直至所述样本数据集遍历完成；

[0017] 步骤 d：统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

[0018] 优选的，统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率具体为：

[0019] 根据表达式

[0020]

$$P(n-gram\text{对应男性}) = \frac{n-gram\text{对应男性频率}}{n-gram\text{对应男性频率} + n-gram\text{对应女性频率}},$$

[0021] 分别计算各顺位上各字母和相邻若干顺位上各字母组合对应为男性的概率。其中，等式左侧的 P(n-gram 对应男性）为长度为 n 的相邻若干顺位上的字母组合对应为男性的概率，n 为 1 时 P(n-gram 对应男性）为单一顺位上的字母对应为男性的概率；等式右侧的 n-gram 对应男性频率为单一顺位上的字母或长度为 n 的相邻若干顺位上的字母组合对应为男性的次数，n-gram 对应女性频率为单一顺位上的字母或长度为 n 的相邻若干顺位上的字母组合对应为女性的次数。

[0022] 优选的，以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别包括：

[0023] 步骤 a：获取所述未知用户性别的用户个性域名的长度，记为 k；

[0024] 步骤 b：按照表达式

[0025]

$$P(url\text{对应用户性别为男性})$$

$$=\sum_{h=1,2,\ldots,n} w_h\Big(\sum_{i=1,2,\ldots,k}\ \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性})\Big)$$

[0026] 计算所述用户的性别为男性的概率，其中，url 表示个性域名中用户指定的部分；substr（url，j，i）表示 url 中第 j 位字符开始长度为 i 的相邻若干顺位上的字母组合构成的子字符串，i 为 1 时为单一顺位上的字母构成的子字符串；n 表示 substr(url,j,i) 的个数；w_h 表示该字母或字母组合的权重；P(substr(url,j,i) 在样本数据集中对应男性）表示上述子字符串上的字母或字母组合对应的男性概率；

[0027] 步骤 c：比较步骤 b 中的计算结果与所述样本数据集中男性的比例；

[0028] 步骤 d：在步骤 b 中的计算结果大于等于步骤 c 计算得到的比例时，判定所述未知性别用户的性别为男性。

[0029] 优选的，所述步骤 d 之后，还包括：

[0030] 步骤 e：在步骤 b 中的计算结果小于步骤 c 计算得到的比例时，判定所述未知性别用户的性别为女性。

[0031] 本发明还提供了一种用户性别分析装置，包括：

[0032] 采样模块，用于采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；

[0033] 参考参数计算模块，用于统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；

[0034] 分析模块，用于以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。

[0035] 优选的，该装置还包括：

[0036] 参考比例计算模块，用于计算所述样本数据集中男性的比例。

[0037] 优选的，所述参考参数计算模块包括：

[0038] 性别提取单元，用于取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别；

[0039] 计数单元，用于对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数；

[0040] 统计单元，用于对所述样本数据集中的全部用户个性域名进行所述计数单元的处理，直至所述样本数据集遍历完成，统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

[0041] 优选的，所述统计单元计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率具体为：

[0042] 根据表达式

[0043]

$$P(n-gram\text{对应男性})=\frac{n-gram\text{对应男性频率}}{n-gram\text{对应男性频率}+n-gram\text{对应女性频率}}$$

[0044] 分别计算各顺位上各字母和相邻若干顺位上各字母组合对应的男性概率，其中，P(n-gram对应男性）为长度为n的相邻若干顺位上一字母组合对应为男性的概率，n为1时P(n-gram对应男性）为单一顺位上一字母对应为男性的概率，n-gram对应男性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为男性的次数，n-gram对应女性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为女性的次数。

[0045] 优选的，所述分析模块包括：

[0046] 域名长度获取单元，用于获取所述未知用户性别的用户个性域名的长度，记为k；

[0047] 概率计算单元，用于按照表达式

[0048]

$$P(url\text{对应用户性别为男性}) = \sum_{h=1,2,\ldots,n} w_h \left(\sum_{i=1,2,\ldots,k} \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性}) \right)$$

[0049] 计算所述用户的性别为男性的概率，其中，url表示个性域名中用户指定的部分，substr（url，j，i）表示url中第j位字符开始长度为i的相邻字符构成的子字符串，n表示substr(url，j，i）的个数，w_h表示该字母或字母组合的权重，P(substr(url,j,i）在样本数据集中对应男性）表示url中第j位字符或第j位字符开始长度为i的相邻字符构成的子字符串上的字母或字母组合对应的男性概率；

[0050] 比较单元，用于比较概率计算单元的计算结果与参考比例计算模块计算得到的比例；

[0051] 判定单元，用于在所述比较单元比较的结果为比较概率计算单元的计算结果大于等于参考比例计算模块计算得到的比例时，判定所述未知性别用户的性别为男性。

[0052] 优选的，所述判定单元，还用于在所述比较单元比较的结果为比较概率计算单元的计算结果小于所述参考比例计算模块计算得到的比例时，判定所述未知性别用户的性别为女性。

[0053] 本发明提供了一种用户性别分析方法和装置，采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别，然后统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率，再以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别，实现了基于自动化算法的用户性别分析，更加灵活和准确，解决了现有分析方式不适用于个性域名和姓名关联较弱的场合的问题。

附图说明

[0054] 图1为本发明的实施例一提供的一种用户性别分析方法的流程图；

[0055] 图2为本发明的实施例二提供的一种用户性别分析装置的结构示意图；

[0056] 图3为图2中参考参数计算模块202的结构示意图；

[0057] 图4为图2中分析模块203的结构示意图。

具体实施方式

[0058] 本发明的实施例提供了一种用户性别分析方法和装置，通过一种自动化的算法，

避免了对人类行为学数据库的依赖。

[0059] 下文中将结合附图对本发明的实施例进行详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

[0060] 首先结合附图，对本发明的实施例一进行说明。

[0061] 本发明实施例提供了一种用户性别分析方法，使用该方法完成用户性别分析的流程如图 1 所示，包括：

[0062] 步骤 101、采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；

[0063] 步骤 102、计算所述样本数据集中男性的比例；

[0064] 步骤 103、统计所述样本数据集中男性所占比例及用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；

[0065] 本步骤具体包括：

[0066] 步骤 a：取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别；

[0067] 步骤 b：计算所述样本数据集中男性的比例；步骤 c：对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数；

[0068] 对于单一字符位上字母出现的次数进行计数的方法如下：

[0069] 将所述用户个性域名第一位上的字母的出现次数加 1，然后再将所述用户个性域名第二位构成的字符串的出现次数加 1，依次统计至该用户个性域名的最后一位。

[0070] 对于相邻若干顺位上不同字母组合出现的次数进行计数的方法如下：

[0071] 首先确定相邻若干顺位构成的字符串的长度 n，然后以所述用户个性域名第一位为起始取 n 顺位构成字符串，将该字符串上的字母组合出现的次数加 1；然后以该用户个性域名第二位为起始取 n 顺位构成字符串，将该字符串上的字母组合出现的次数加 1。依此类推，直致字符串的末位为用户个性域名的最后一位为止。n 的取值由 2 至所述用户个性域名的长度。

[0072] 步骤 c：对所述样本数据集中的全部用户个性域名进行如步骤 a 至 b 的处理，直至所述样本数据集遍历完成；

[0073] 步骤 d：统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

[0074] 本步骤中，根据表达式

[0075]

$$P(n-gram\text{对应男性})=\frac{n-gram\text{对应男性频率}}{n-gram\text{对应男性频率}+n-gram\text{对应女性频率}},$$

[0076] 分别计算各顺位上各字母和相邻若干顺位上各字母组合对应为男性的概率。其中，等式左侧的 P(n-gram 对应男性）为长度为 n 的相邻若干顺位上一字母组合对应为男性的概率，n 为 1 时 P(n-gram 对应男性）为单一顺位上一字母对应为男性的概率；等式右侧的 n-gram 对应男性频率为单一顺位上一字母或长度为 n 的相邻若干顺位上一字母组合对应为男性的次数，n-gram 对应女性频率为单一顺位上一字母或长度为 n 的相邻若干顺位上

一字母组合对应为女性的次数。

[0077] 步骤 104、以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别；

[0078] 本步骤具体包括：

[0079] 步骤 a：获取所述未知用户性别的用户个性域名的长度，记为 k；

[0080] 步骤 b：按照表达式

[0081]

$$P(url\text{对应用户性别为男性}) = \sum_{h=1,2,\ldots,n} w_h \left(\sum_{i=1,2,\ldots,k} \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性}) \right)$$

[0082] 计算所述用户的性别为男性的概率，其中，url 表示个性域名中用户指定的部分，substr（url，j，i）表示 url 中第 j 位字符开始长度为 i 的相邻字符构成的子字符串，n 表示 substr(url，j，i) 的个数，w_h 表示该字母或字母组合的权重，P(substr(url,j,i) 在样本数据集中对应男性）表示 url 中第 j 位字符或第 j 位字符开始长度为 i 的相邻字符构成的子字符串上的字母或字母组合对应的男性概率；

[0083] 步骤 c：比较步骤 b 中的计算结果与步骤 102 计算得到的男性的比例；

[0084] 步骤 d：在步骤 b 中的计算结果大于等于步骤 102 计算得到的比例时，判定所述未知性别用户的性别为男性；

[0085] 步骤 e：在步骤 b 中的计算结果小于步骤 102 计算得到的比例时，判定所述未知性别用户的性别为女性。

[0086] 下面结合附图，对本发明的实施例二进行说明。

[0087] 本发明实施例提供了一种用户性别分析装置，其结构如图 2 所示，包括：

[0088] 采样模块 201，用于采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别；

[0089] 参考参数计算模块 202，用于统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率；

[0090] 分析模块 203，用于以所述样本数据集中男性的比例和所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别。

[0091] 优选的，该装置还包括：

[0092] 参考比例计算模块 204，用于计算所述样本数据集中男性的比例。

[0093] 优选的，所述参考参数计算模块 202 的结构如图 3 所示，包括：

[0094] 性别提取单元 2021，用于取一个用户个性域名中用户指定的部分，同时记录该用户个性域名对应的用户性别；

[0095] 计数单元 2022，用于对所述指定的部分的各顺位上字母出现的次数和 / 或相邻若干顺位上不同字母组合出现的次数进行计数；

[0096] 统计单元 2023，用于对所述样本数据集中的全部用户个性域名进行所述计数单元的处理，直至所述样本数据集遍历完成，统计所述用户个性域名各顺位上字母对于不同性别出现的次数和 / 或相邻若干顺位上的字母组合对于不同性别出现的次数，并计算各顺位上字母和 / 或相邻若干顺位上字母组合对于不同性别出现的概率。

[0097] 优选的，所述统计单元2023计算各顺位上字母和/或相邻若干顺位上字母组合对于不同性别出现的概率具体为：

[0098] 根据表达式

[0099]

$$P(n-gram对应男性)=\frac{n-gram对应男性频率}{n-gram对应男性频率+n-gram对应女性频率},$$

[0100] 分别计算各顺位上各字母和相邻若干顺位上各字母组合对应的男性概率，其中，P(n-gram对应男性)为长度为n的相邻若干顺位上一字母组合对应为男性的概率，n为1时P(n-gram对应男性)为单一顺位上一字母对应为男性的概率，n-gram对应男性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为男性的次数，n-gram对应女性频率为单一顺位上一字母或长度为n的相邻若干顺位上一字母组合对应为女性的次数。

[0101] 优选的，所述分析模块203的结构如图4所示，包括：

[0102] 域名长度获取单元2031，用于获取所述未知用户性别的用户个性域名的长度，记为k；

[0103]

$$P(url对应用户性别为男性)\text{概率计算单元2032，用于按照表达式}=\sum_{h=1,2,\ldots,n} w_h\Big(\sum_{i=1,2,\ldots,k}\ \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)在样本数据集中对应男性)\Big)$$

[0104] 计算所述用户的性别为男性的概率，其中，url表示个性域名中用户指定的部分，substr（url，j，i）表示url中第j位字符开始长度为i的相邻字符构成的子字符串，n表示substr(url，j，i)的个数，w_h表示该字母或字母组合的权重，P(substr(url,j,i)在样本数据集中对应男性）表示url中第j位字符或第j位字符开始长度为i的相邻字符构成的子字符串上的字母或字母组合对应的男性概率；

[0105] 比较单元2033，用于比较概率计算单元2032的计算结果与参考比例计算模块204计算得到的比例；

[0106] 判定单元2034，用于在所述比较单元2033比较的结果为比较概率计算单元2032的计算结果大于等于参考比例计算模块204计算得到的比例时，判定所述未知性别用户的性别为男性。

[0107] 优选的，所述判定单元2034，还用于在所述比较单元2033比较的结果为比较概率计算单元的计算结果小于所述参考比例计算模块计算得到的比例时，判定所述未知性别用户的性别为女性。

[0108] 下面结合附图，对本发明的实施例三进行说明。

[0109] 本发明实施例公开了一种用户性别分析系统，根据用户申请、拥有或使用的个性域名，自动对用户的性别进行分类。本发明实施例首先通过用户数据统计、商业合作等手段取得个性域名和用户性别的对应关系的样本数据集，然后分析个性域名中用户指定的部分，使用机器学习的方法，训练出使用个性域名对用户性别进行分类的分类器。当需要对未知用户性别的个性域名进行分类时，使用这一分类器，即可输出预测的用户性别。

[0110] 具体步骤如下。

[0111] 步骤一：采集个性域名和用户性别的对应关系的样本数据集，分析个性域名中用

户指定的部分。

[0112] 步骤二：计算所述样本数据集中男性所占的比例。

[0113] 步骤三：取一个个性域名中用户指定的部分，记为字符串一，同时记录对应用户性别。

[0114] 步骤四：将字符串一的长度记为 k，统计字符串一中所有 1-gram 的出现频率、2-gram 的出现频率、3-gram 的出现频率直至 k-gram 的出现频率（k 代表字符串的长度，可以为 1 或 1 以上的整数，k 的取值上限），将相应 n-gram（n 代表字符串的长度，可以为 1 至 k）的出现频率按对应用户性别累加。

[0115] 步骤五：重复步骤三，直至步骤一中采集的样本数据集遍历完成。

[0116] 步骤六：计算所有出现过的 n-gram 所对应用户性别的概率和该 n-gram 的出现次数，同时统计样本数据集中不同性别出现的概率，共同作为分类器的参数。

[0117] 步骤六：使用分类器时，对未知用户性别的个性域名，分析其中用户指定的部分，将长度记为 k，取得其 1-gram 直至 k-gram，按下面的公式计算其性别为男性的概率：

[0118]

$$P(url\text{对应用户性别为男性}) = \sum_{h=1,2,\ldots,n} w_h \left(\sum_{i=1,2,\ldots,k} \sum_{j=1,2,\ldots,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性}) \right)$$

[0119] 计算所述用户的性别为男性的概率，其中，url 表示个性域名中用户指定的部分，substr（url，j，i）表示 url 中第 j 位字符开始长度为 i 的子字符串，n 表示 substr(url，j，i) 的个数，w_h 表示该字母组合的权重。

[0120] 步骤七：若步骤六中计算得到的概率大于步骤二中计算得到的样本数据集中男性出现的比例，则可将该个性域名分类为对应男性用户，反之对应女性用户。

[0121] 下面结合附图，对本发明的实施例四进行说明。

[0122] 本发明实施例提供了一种用户性别分析方法，具体流程如下：

[0123] 步骤一：采集到下面三个个性域名：http://weibo.com/nickleave，http://weibo.com/inferpku，http://t.qq.com/bankofdota，其中用户指定的部分分别为 nickleave、inferpku、bankofdota。本例中，本发明涉及的系统通过商业合作的手段从 weibo.com 和 t.qq.com 的服务提供商处取得信息，得知 nickleave 所对应的用户性别为女，inferpku 和 bankofdota 所对应的用户性别为男。

[0124] 步骤二：计算所有样本中男性的概率。在步骤一中，我们总共收集了三个样本，性别分别为 nickleave（女）、inferpku（男）以及 bankofdota（男）。由此可见，在三个样本中，男性比例占 2/3。

[0125] 步骤三：取 nickleave，女。

[0126] 步骤四：nickleave 对应的 1-gram、2-gram、3-gram、4-gram、5-gram、6-gram、7-gram、8-gram、9-gram，累加至对应女性中，统计结果如表 1 所示。为了表示方便，下表中只列出 1-gram、2-gram、3-gram 三种情况。

[0127] 表 1

[0128]

1-gram	男性频率	女性频率	2-gram	男性频率	女性频率	3-gram	男性频率	女性频率
n	0	1	ni	0	1	nic	0	1
i	0	1	ic	0	1	ick	0	1
c	0	1	ck	0	1	ckl	0	1
k	0	1	kl	0	1	kle	0	1
l	0	1	le	0	1	lea	0	1
e	0	2	ea	0	1	eav	0	1
a	0	1	av	0	1	ave	0	1
v	0	1	ve	0	1			

[0129] 步骤五：对 inferpku 重复上述过程。累加后，表 1 中的结果更新为表 2：

[0130] 表 2

[0131]

[0132]

1-gram	男性	女性	2-gram	男性	女性	3-gram	男性	女性
	频率	频率		频率	频率		频率	频率
n	1	1	ni	0	1	nic	0	1
i	1	1	ic	0	1	ick	0	1
c	0	1	ck	0	1	ckl	0	1
k	1	1	kl	0	1	kle	0	1
l	0	1	le	0	1	lea	0	1
e	1	2	ea	0	1	eav	0	1
a	0	1	av	0	1	ave	0	1
v	0	1	ve	0	1	inf	1	0
f	1	0	in	1	0	nfe	1	0

r	1	0	nf	1	0	fer	1	0
p	1	0	fe	1	0	erp	1	0
u	1	0	er	1	0	rpk	1	0
			rp	1	0	pku	1	0
			pk	1	0			
			ku	1	0			

[0133] 再对 bankofdota 重复上述过程。累加后，表 2 更新为表 3：

[0134] 表 3

[0135]

1-gram	男性频率	女性频率	2-gram	男性频率	女性频率	3-gram	男性频率	女性频率
n	2	1	ni	0	1	nic	0	1
i	1	1	ic	0	1	ick	0	1

[0136]

c	0	1	ck	0	1	ckl	0	1
k	2	1	kl	0	1	kle	0	1
l	0	1	le	0	1	lea	0	1
e	1	2	ea	0	1	eav	0	1
a	2	1	av	0	1	ave	0	1
v	0	1	ve	0	1	inf	1	0
f	2	0	in	1	0	nfe	1	0
r	1	0	nf	1	0	fer	1	0
p	1	0	fe	1	0	erp	1	0
u	1	0	er	1	0	rpk	1	0
b	1	0	rp	1	0	pku	1	0
o	2	0	pk	1	0	ban	1	0
d	1	0	ku	1	0	ank	1	0
t	1	0	ba	1	0	nko	1	0
			an	1	0	kof	1	0
			nk	1	0	ofd	1	0
			ko	1	0	fdo	1	0
			of	1	0	dot	1	0
			fd	1	0	ota	1	0
			do	1	0			
			ot	1	0			
			ta	1	0			

[0137] 步骤六：计算所有出现过的 n-gram 所对应用户性别的概率和该 n-gram 的出现次数。其中，n-gram 对应用户性别为男性的概率(即下表中的男性概率)的计算方法为：

[0138]

$$P(n-gram对应男性)=\frac{n-gram对应男性频率}{n-gram对应男性频率+n-gram对应女性频率}$$

[0139] 例如，表 4 中的 1-gram n 对应男性概率是用上表中 1-gram n 对应男性频率 2 和女性频率 1 计算得到的，即 2/(2+1)=0.666667。

[0140] 为了表示方便，表 4 中只列出 1-gram、2-gram、3-gram 三种情况。

[0141] 表 4

[0142]

1-gram	男性概率	2-gram	男性概率	3-gram	男性概率
n	0.666667	ni	0	nic	0
i	0.5	ic	0	ick	0
c	0	ck	0	ckl	0
k	0.666667	kl	0	kle	0
l	0	le	0	lea	0
e	0.333333	ea	0	eav	0
a	0.666667	av	0	ave	0
v	0	ve	0	inf	1
f	1	in	1	nfe	1
r	1	nf	1	fer	1
p	1	fe	1	erp	1
u	1	er	1	rpk	1
b	1	rp	1	pku	1
o	1	pk	1	ban	1
d	1	ku	1	ank	1
t	1	ba	1	nko	1
		an	1	kof	1
		nk	1	ofd	1

		ko	1	fdo	1
		of	1	dot	1
		fd	1	ota	1
		do	1		
		ot	1		
		ta	1		

[0143]

[0144] 步骤七：假设需要进行分类的个性域名是 www.renren.com/eleven，其中用户指定的部分是 eleven，在 eleven 中出现的 n-gram 包括 e（三次）、l（一次）、v（一次）、n（一次）、le（一次）、ve（一次），而在上述第三个性别频率表格中出现的 n-gram 包括 e（三次）、l（一次）、v（一次）、n（三次）、le（一次）、ve（一次）。用户名 eleven 在性别频率表格中出现的 n-gram 总次数为 10。由此计算出字母或字母组合的权重 w_h。根据上文中公式，带入以上数值，可得：

[0145]

$$P(url\text{对应用户性别为男性})$$

$$=\sum_{h=1,2,..,n} w_h\Big(\sum_{i=1,2,..,k}\ \sum_{j=1,2,..,k-i+1} P(substr(url,j,i)\text{在样本数据集中对应男性})\Big)$$

$$=\quad (p(e\ \text{男性概率})*3+p(l\ \text{男性概率})*1+p(v\ \text{男性概率})*1+p(n\ \text{男性概率})*1+p(le\ \text{男性概率})*1+p(ve\ \text{男性概率})*1)$$

[0146] $= (0.33*3+0*1+0*1+0.67*1+0*1+0*1)$

[0147] $= (1.66)$

[0148] $=0.166$

[0149] 步骤七：

[0150] 步骤八：由于上步骤二中计算得到的 eleven 对应用户性别为男性的概率 0.166 小于样本中男性所占比例 0.67，因此可将 www.renren.com/eleven 分类为对应女性用户。

[0151] 本发明的实施例提供了一种用户性别分析方法和装置，采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别，统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率，再以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别，，实现了基于自动化算法的用户性别分析，更加灵活和准确，解决了现有分析方式不适用于个性域名和姓名关联较弱的场合的问题。

[0152] 本发明实施例提供的技术方案，通过一种自动化的算法，避免了对人类行为学数

据库的依赖。现有的分析方式对姓名的依赖，其并不适用于个性域名等和姓名关联较弱的场合，而本发明实施例提供的技术方案不存在这一问题。此外，本发明的实施例通过对个性域名的分析，可用于展示广告优化等更广阔的应用中。

[0153] 本领域普通技术人员可以理解上述实施例的全部或部分步骤可以使用计算机程序流程来实现，所述计算机程序可以存储于一计算机可读存储介质中，所述计算机程序在相应的硬件平台上（如系统、设备、装置、器件等）执行，在执行时，包括方法实施例的步骤之一或其组合。

[0154] 可选地，上述实施例的全部或部分步骤也可以使用集成电路来实现，这些步骤可以被分别制作成一个个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件结合。

[0155] 上述实施例中的各装置／功能模块／功能单元可以采用通用的计算装置来实现，它们可以集中在单个的计算装置上，也可以分布在多个计算装置所组成的网络上。

[0156] 上述实施例中的各装置／功能模块／功能单元以软件功能模块的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。上述提到的计算机可读取存储介质可以是只读存储器，磁盘或光盘等。

[0157] 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权利要求所述的保护范围为准。

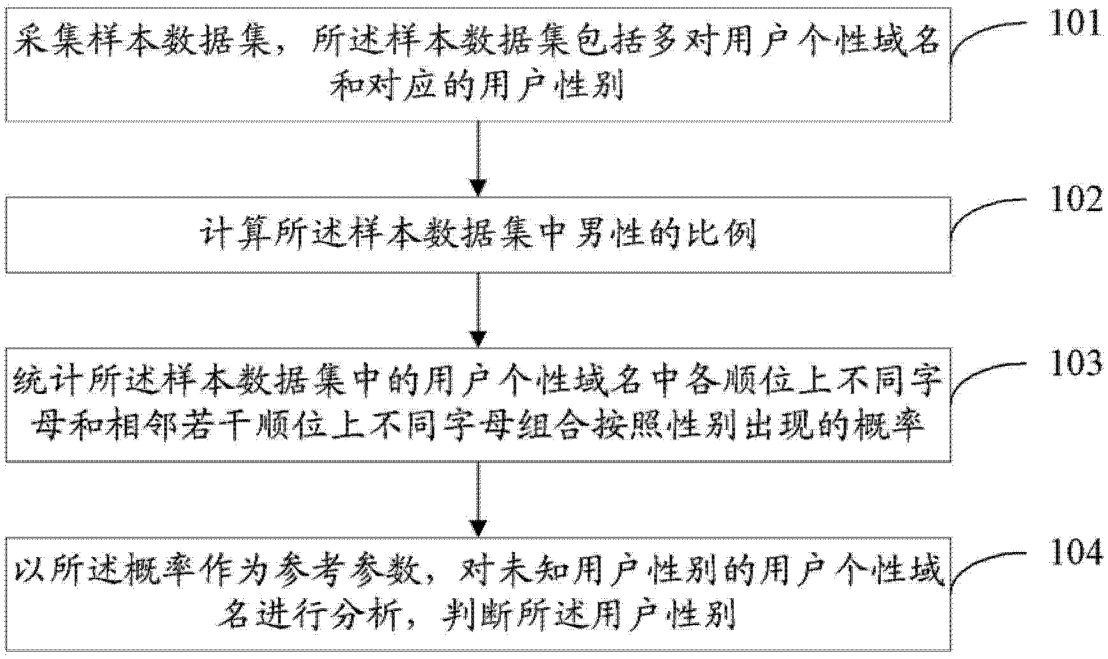

采集样本数据集，所述样本数据集包括多对用户个性域名和对应的用户性别
101
计算所述样本数据集中男性的比例
102
统计所述样本数据集中的用户个性域名中各顺位上不同字母和相邻若干顺位上不同字母组合按照性别出现的概率
103
以所述概率作为参考参数，对未知用户性别的用户个性域名进行分析，判断所述用户性别
104

图 1

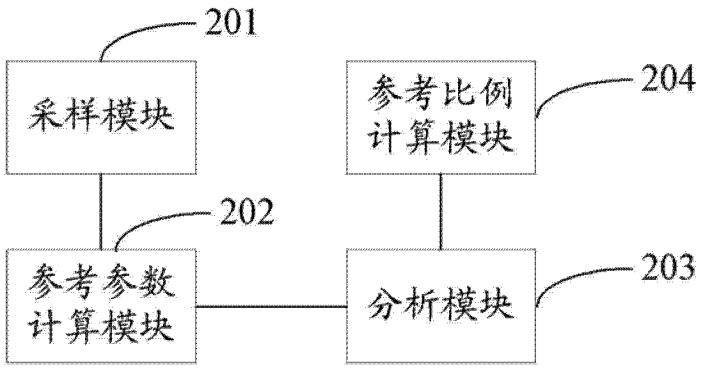

201
采样模块
202
参考参数计算模块
204
参考比例计算模块
203
分析模块

图 2

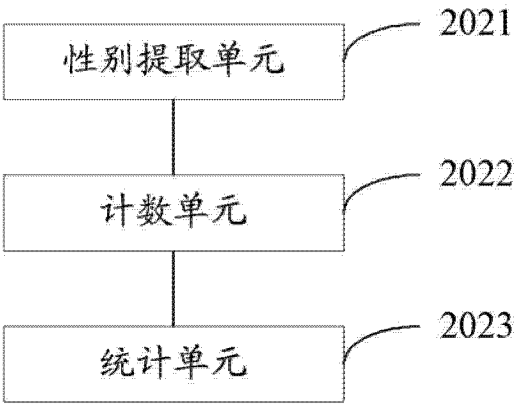

性别提取单元
2021
计数单元
2022
统计单元
2023

图 3

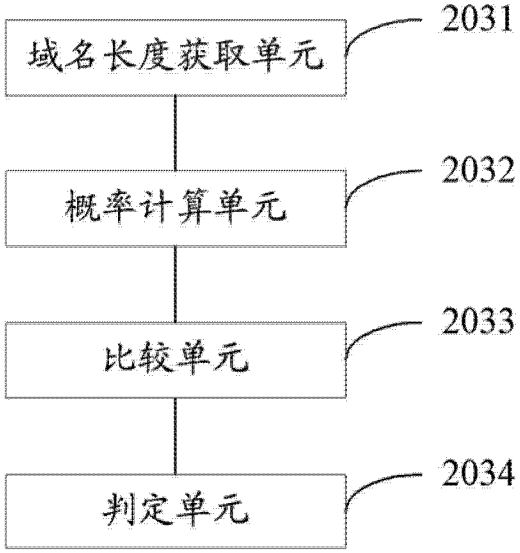

域名长度获取单元
2031
概率计算单元
2032
比较单元
2033
判定单元
2034

图 4