**用户影响力**

## 1 相关变量定义

（1）用户被认可度。该指标主要是指其他用户对该用户的评论认可程度，以用户收到的点赞数与其发表评论总数比率进行量化。

= 3-1

其中u代表用户u，代表用户在Yelp社区的受认可程度，代表用户u发表评论中收到其他用户点赞“useful”的总数，代表用户u发表的评论总数（都在user数据中）。

1. 用户活跃度。该指标主要是指用户在一段时间内在Yelp平台的活动频率。

= 3-2

其中u代表用户u，表示用户的活跃度，表示用户u在发表评论的总数，用户u注册年数（都在user数据中）。

1. 用户互动频率。该指标表示用户与Yelp社区内其他用户进行互动的频率。

= 3-3

其中u代表用户u，代表用户互动频率，代表用户u接受了其他用户发送的互动标签总数，为compliment\_hot、compliment\_more、compliment\_profile、compliment\_cute、compliment\_list、compliment\_note、compliment\_plain、compliment\_cool、compliment\_funny、compliment\_writer、compliment\_photos之和，代表用户收到的所有评论的点赞数，为useful、funny、cool之和。

（在user数据中）。

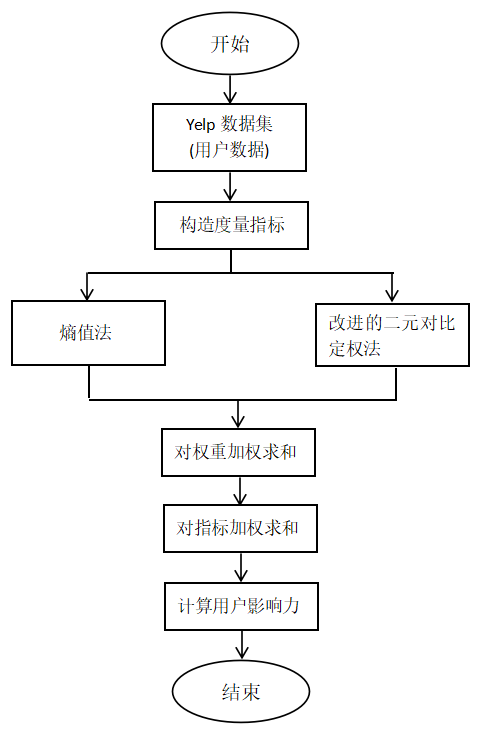
1. 用户被关注度。该指标指用户在Yelp社区的受欢迎程度。主要通过用户收到的点赞数和用户粉丝数来计算。

= a·+ b·fans 3-4

其中u代表用户u，代表用户u的被关注度，代表用户u收到的其他用户给予的点赞总数，包括“useful、funny、cool”，fans代表用户数据中用户u的粉丝数量，a、b则分别代表了点赞总数和粉丝数量的系数，本文中令a、b均为0.5（在user数据中）。

（5）用户粉丝数即用户数据集中“fans”所对应的值，用户注册年数通过用户数据集“since”与2021年计算得出（在user数据中）。

（6）特征集。根据上述构造指标以及用户本身所具有属性指标建立用户特征集，具体包括用户被认可度、用户活跃度、用户互动频率、用户被关注度和用户粉丝数fans、用户注册年数Y。



## 2 改进的二元对比定权法

研究中假设指标数量为m个，且第i个指标用表示。我们首先要对指标进行重要度排序，具体方法为指标两两对比，并且九标度法对各个指标的重要度进行量化。然后构造出重要度排序矩阵，排序矩阵用A表示，表示指标i相对于指标j的重要度。九标度法的判定方法及其标度含义如图所示。

表4-1 九标度法判定方法及其标度含义

|  |  |  |  |
| --- | --- | --- | --- |
| 标度/ | 含义 | 标度/ | 含义 |
| 1 | 指标i与指标j同等重要 | 9 | 指标i与指标j极端重要 |
| 3 | 指标i比指标j稍微重要 | 2、4、6、8 | 重要度介于相邻奇数之间 |
| 5 | 指标i比指标j较为重要 | 1/ | 反比较（重要度和上述相反） |
| 7 | 指标i比指标j非常重要 |  |  |

基于上述含义及对应标度，则可得到指标重要性赋值公式，如下：

（） = 3-5

其中表示指标相对于指标的重要性，并用表示A重要度排序矩阵，可以表示成如下形式：

 3-6

由于矩阵A表示的是指标重要性排序矩阵，则通过该矩阵即可反映出两个指标相互之间的重要程度。然后要进行的是各指标权重比的计算，权重比的计算则根据公式X-X求得的各指标重要性对比求出，具体公式如下：

= 3-7

其中表示指标相对于指标的重要性，表示指标相对于指标的重要性。

用矩阵R可以表示成m\*m的矩阵，则权重比矩阵R表示成如下公式：

 3-8

由于本文中权重比矩阵是根据对指标重要度进行的主观排序，因此会存在一定误差。所以需要进行一致性检验来修改权重比矩阵，来降低或消除主观误差。权重比矩阵的指标方程如下：

R·W = ·W 3-9

其中表示权重矩阵的最大特征根（存在且唯一），W表示最大特征根所对应的指标向量。

权重比矩阵与指标方程确定后，下一步则需要对权重比矩阵进行一致性检验。一致性检验的指标表示如下：

CI = 3-10

其中m表示指标个数，也代表矩阵阶数。当CI=0时，即最大特征根=m时，此时认为该权重比矩阵具有完全一致性。当CI>0时，随着CI的增大，则权重比矩阵的一致性随之逐渐减弱，此时需要根据随机一致性比例CR来判定权重比矩阵是否具有满意的一致性。CR计算公式如下：

CR = 3-11

式（）中，RI为平均随机一致性指标，不同的指标个数（矩阵阶数）对应不同的RI值，具体如表4-2所示。当CT<0.1时，则可认为权重比矩阵R的一致性比较满意，否则就需要修改权重比矩阵，直到得到满意结果为止。

表4-2 平均随机一致性指标数值表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 矩阵阶数 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| RI值 | 0 | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 |

若权重比矩阵符合一致性检验，然后对指标向量W进行归一化处理后得到主观权重向量，主观权重向量可以表示为w1=，其中是指标的权重估计值。

## 3 熵值法

首先对相关变量进行定义，假设有m个用户，n个指标，表示第i个用户的第j个指标。则在本研究中，n=6。

=

其中表示第i个用户的第j个指标归一化后的值，表示的是未进行归一化之前的值，归一化采用的是最大最小归一化。

在进行了用户指标归一化处理后，再对第i个用户占第j个指标的比重进行计算，用第i个用户在第j个指标中的数值比上第j个指标的数值和（还没太搞懂），具体公式如下：

= ,i= 1,2,...,n,j= 1,2...,m

其中是第i个用户在第j个指标下经过归一化处理后的值。

其次对第j个指标的信息熵进行计算。由于较小的信息熵的不确定度更小，因此包含了更火的信息，对于整个用户影响力度量的计算也具有较大的影响。具体计算公式如下：

= -

接着对信息熵的冗余度进行计算。信息熵和信息熵冗余度之和为1，信息熵越小，则信息熵冗余度越大。而我们要得到包含有效信息多的指标，即信息熵小的指标。因此，也可以通过信息熵冗余度来表达指标包含的信息量及准确度。指标信息熵冗余度越大，则包含的有效信息越多。信息熵冗余度的计算公式如下：

= 1 -

其中表示第j个指标的信息熵冗余度，表示第j个指标的信息熵。

最后要对各项指标的权重进行计算，第j个指标的权重等于第j个指标的信息冗余度与所有指标的信息冗余度之和的比值，代表第j个指标在所有指标中所占的重要程度。具体公式如下：

=

其中代表第j个指标的权重，代表第j个指标的信息熵冗余度。

以上就是通过熵值法对用户指标的权重计算，需保留此时计算得到的所有指标权重，为最后的指标综合权重计算做好准备。

## 4 综合权重赋值法

综合权重赋值法是一种综合的计算方法，它结合了改进的二元对比定权法和熵值法。其中二元对比定权法也可称为主观加权法，熵值法也可称为客观加权法。综合权重赋值法运用了上述两种方法的优点，通过此方法确定指标权值，不仅具有数学意义，而且结合了主观思想。使用综合权重赋值法确定的指标不仅具有数据支持，同时又因为在其中加入了主观法，所以不会出现指标重要度和实际相差过于悬殊的情况，使用此方法确定的指标权重也会更加具有说服力。

假设由改进的二元对比定权法确定的权重向量为，由熵权法确定的权重向量为。为了保证使用不同计算方法计算得出的权重指标不存在指标之间差别悬殊或者是互相冲突的情况，所以在进行综合权重赋值前，要对不同的赋权方法得到的赋权结果进行一致性检验。根据文献【72】的研究结果表明，当使用由两种赋权方法确定的权重进行组合赋权时，可以用距离函数来刻画综合赋权法的一致性。具体的计算公式如下：

d（X,Y）=

其中，需要满足0≤d（X,Y）≤1，并且若d（X,Y）的值越小，则说明两种赋权结果越接近。综合权重赋值法采用的是加法集成法。具体公式如下：

W=αX+βY （4-18）

其中w表示通过综合权重赋值法得到的组合权重，α和β分别是两种方法得到的权重的权重系数，也可称为分配系数，且满足α≥0，β≥0，α+β = 1。对于第i个指标的组合权重可以表示成公式如下：

=α+β

对于分配系数α和β值的计算，使用了距离计算公式。因为距离可以反映出权重之间的差异程度，本文使用的是欧氏距离。具体计算公式如下所示：、

d（）=

为了保证α、β的差异程度和距离函数相等，需要满足下述公式：

= （4-21）

根据α+β = 1，然后把主观加权法和客观赋权求出的权重代入公式4-21，即可求出α和β的值。根据α和β的计算公式，结合公式4-18，得出每个指标的综合权重。

上述计算综合权重赋值法的方法又可以称之为线性加权组合赋权方法。然后通过将主观法和客观法计算得到的权重与用户指标进行加权求和，可以计算得出最终的用户影响力。具体计算公式如下：

=

其中i∈[1,n]，表示最终计算得到的第i个用户的影响力具体数值，代表第j个特征的组合权重，表示进行归一化处理后的第i个用户的第j个指标数值。

## 5实验

### 5.1 实验数据

User.jason

### 3.3.2 数据预处理

在开始进行用户影响力度量之前，我们首先需要对数据进行预处理。本研究中数据的预处理可以分为三部分，分别是数据处理、度量指标构造以及指标清洗。

（1）数据处理。由于本文最终选择展示是用户影响力排名靠前的用户，因此首先对用户进行筛选，本文分别选择用户评论数大于1000的Yelp共1835名用户和（这个数据是我已经确认过的）用户评论数少于50条的yelp用户共XX名进行研究（这个有多少我还没有筛选过）。

（2）度量指标构造。本文中构造的指标包括用户被认可度、用户活跃度、用户互动频率以及用户被关注度，还对用户粉丝数和用户注册年限进行计算。上述用户被认可度、用户活跃度、用户互动频率以及用户被关注度指标分别通过公式3-1、3-2、3-3、3-4进行构造。用户粉丝数通过用户数据集中“fans”属性进行提取，用户注册年限通过用户数据集中“yelping\_since”进行计算得出。

（3）指标清洗。最后对指标进行选择，保留的指标有用户ID、用户被认可度、用户活跃度、用户互动频率、用户被关注度、粉丝数、用户注册年限等7个指标。而其中的用户ID是作为用户的唯一标识符，不参与权重计算。其中，用户被认可度、用户活跃度、用户互动频率保留3位小数，用户被关注度和注册年限保留1位小数，粉丝数为整数。下面对数据处理完成后的部分数据进行展示，如表3-3所示：

表 3-3 部分用户度量指标展示(这几条数据是我处理的评论数大于1000里的)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 用户被认可度 | 用户活跃度 | 用户互动频率 | 用户被关注度 | 粉丝数 | 注册年限 |
| 1 | 12.326 | 75.776 | 0.432 | 18858 | 1357 | 16.1 |
| 2 | 9.959 | 157.059 | 0.238 | 25316 | 1025 | 13.6 |
| 3 | 7.067 | 141.500 | 0.407 | 15326 | 313 | 14 |
| 4 | 2.098 | 95.804 | 0.217 | 3208 | 166 | 14.3 |
| 5 | 21.059 | 216.855 | 0.163 | 77535.5 | 392 | 12.4 |

### 3.3.3 实验结果

本章首先使用改进的二元对比定权法对用户数据集进行处理，该算法包含了更多的主观因素，得出的权重和真实情况更为相似。用户被认可度和用户活跃度相比，用户被认可度更能体现用户的影响力，目标用户越被其他用户认可，说明该用户的行为对其他用户的影响越大。所以主观认为用户被认可度的重要性大于用户活跃度。用户粉丝数与用户注册年限相比，粉丝数越多则说明该用户越受欢迎。而用户注册年限只能反映用户加入Yelp网站的时长，因此粉丝数比用户注册年限更重要。用户活跃度与用户互动频率相比，用户互动频率比用户活跃度更能说明影响力大小。用户活跃度只是说明用户在该平台自身的一个活跃情况，与其他用户的行为互相之间影响不大。而用户互动频率则比用户活跃度更能反映用户与其他用户之间的联系。用户互动频率度量的是用户本身发出动态后，其他用户是否给予该用户互动，因此用户互动频率比用户活跃度重要性要大一些。

主观对各个指标进行重要性的两两对比后，可以通过改进的二元对比定权法公式得到指标权重比矩阵，并对权重比矩阵进行一致性检验。

下边这部分需要输出的数据

本实验首先根据公式3-9求出权重比矩阵对应的最大特征值的值为**XX**，指标个数为6个，因此m=6，代入到公式3-10中得到CI=XX。然后根据矩阵阶数为6，对照表3-2可得，RI=XX，再根据公式3-11，得CR=XX，符合CR值小于0.1的情况，因此认为当最大特征值为XX时，权重比矩阵符合一致性检验。此时将=XX时对应的指标向量进行归一化，得到最终的权重向量。最终获得的权重向量可以表示为：=（）。

以上为通过改进的二元对比定权法得到的用户影响力度量指标权重，下面将对熵值法确定权重的过程以及结果进行解释说明。

熵值法则是一种客观的权重赋值方法，该方法是通过一系列计算来确定各个指标的权重。首先还是对指标进行归一化处理，然后计算出指标的信息熵，再根据信息熵的值计算出信息熵冗余度。最后则是根据信息熵冗余度的比值计算出具体每个指标的权重。最终通过熵值法获得的权重向量可以表示为：=（）。

以上则是通过熵值法计算得到的指标权重，接下来要根据公式3-15先对最终的赋权结果进行一致性检验，得到的结果是XX，说明两种赋权方法具有一致性。

进行上述两步操作后，最后通过运用主客观综合权重赋值法确定指标的最终权重w。w是通过线性加权求和的方式获得，根据改进的二元对比定权法求出的和熵值法求出的，再结合公式3-20和3-21分别求出的值，，。最后根据公式3-18，求出综合权重w如下：

w = （）

在实验最后，我们将通过改进的二元对比定权法得到的和熵值法求出的以及主客观综合权重赋值法得到的w通过绘制对比图来进行比较。以此来比较上述三个权重分配方法之间的差别。对比图如图3-2所示：

图3-2 权重对比图

通过权重对比图可知，

然后再就是影响力结果排名表格分析（展示数据选取使用三种方法分别排名前20的用户即可）。

二元对比定权法排名前20的用户

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 用户被认可度 | 用户活跃度 | 用户互动频率 | 用户被关注度 | 粉丝数 | 注册年限 |
|  |  |  |  |  |  |  |

熵值法排名前20的用户

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 用户被认可度 | 用户活跃度 | 用户互动频率 | 用户被关注度 | 粉丝数 | 注册年限 |
|  |  |  |  |  |  |  |

综合赋权排名前20的用户

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 用户被认可度 | 用户活跃度 | 用户互动频率 | 用户被关注度 | 粉丝数 | 注册年限 |
|  |  |  |  |  |  |  |