用户行为分析

## 综述

学习每个用户的知识和经验，是知道吧最核心的内容，这是形成“人脑云”的基础数据。目前，知道吧学习用户来自三个方面：

1. 用户资料的填写，在注册时要求用户填写公司和学校的信息，对于不在资料库中的资料需要经过人工审核；
2. 用户在网站或APP的行为，如提问、回答、浏览问题、评论问题、赞问题、浏览他人个人资料、向人脉求助等；
3. 用户通过“更多—标示自己”的功能，为自身贴标签。

为了量化每种行为，并转换为用户知识和经验的一种描述，我们规定：用户的每种行为能给用户增加一定的**知识量**，知识量表示该行为触发对部分标签的权值的增加量或减少量的总和。

由于每个标签的权值最高为1，最低为0。每次给用户的知识量增量确定后，再根据标签现有的权值再将知识量分配给标签。

对于和问题相关的知识量增量，和该问题的属性相关，比如有的问题非常详尽，则其知识量相对较多。因此，在确定知识量之前，我们需要先确定该问题的贡献量和消耗量，作为衡量该问题的知识量的参照。

## 贡献量计算

贡献量在问题进入结束状态后的第一次评价时做一次计算，当第二个用户完成了评价时，进行一次更新。（在用户分享后，对用户行为学习时需要使用）

在设计分享率的时候，目标是通过贡献量来统计其贡献值，用来计算用户的分享率和等级。在计算贡献量的时候，主要考虑：1. 回答问题的数量。 2. 回答问题的质量。

**贡献量=贡献量基数\*贡献量权值。**

签约专家绩效也参考本标准，并进行适当调整。

**补充：**

添加操作时间段，鼓励用户在稀缺的时间段进行交流。

主动暂停，则减少贡献量；接受暂停继续，奖励贡献量。

**每天回答问题的数量不同，每个问题的贡献量基数也不一样。**

每天回答的前3个问题，贡献量基数：1

每天回答的第4-5个问题，贡献量基数：1.05

每天回答的第6-7个问题，贡献量基数：1

每天回答的第8-10个问题，贡献量基数：0.9

每天回答的第11-20个问题，贡献量基数：0.8

每天回答的第20-50个问题，贡献量基数：0.5

每天回答超过50个的问题，贡献量基数：0

(说明：此部分暂时没有实现)

**评价对贡献量具有决定性的作用：**

如果对方评价为1星，表示此次对话是属于极差的内容，贡献量=-0.5。

如果对方评价为2星，贡献量权值=总得分\*0.6

如果对方评价为3星，贡献量权值=总得分\*1

总得分由以下几个部分组成：响应新消息速度/主动回答（20%）、聊天中平均响应时间[不包含暂停的的时间]（30%）、聊天过程总字数（40%）、自己是否填写评价内容（5%）、是否进行分享（5%）、用户等级附加分（5%）、暂停附加分（5%）。

**响应新消息速度/主动回答（20%）：**

主动回答得分为：0.2

在新问题推送后0-20s内接受：0.2

在新问题推送后21-60s内接受：0.18

在新问题推送后61-180s内接受：0.15

在新问题推送后181-600s内接受：0.10

在新问题推送后601-1200s内接受：0.05

在新问题推送后1201s后接受：0.01

在1800s后取消专家的推送消息，并将问题放入急需库重新投放。

**聊天中平均响应时间的定义（30%）：**

假设A向B提问，在A和B的聊天过程中，统计在聊天过程中B对A聊天的响应速度，在每个建立聊天的问题中，日志系统每个answer带有两个属性：距离上一次answer（不管是谁的）的时间间隔、本次的answer的字数。对于第一次answer的时间间隔从建立聊天时算起，对于暂停的问题，从暂停的对话刚继续时算起（正常手机打字速度为20-60字/分钟）。

平均响应时间=每次answer的时间间隔总和/每次answer的字数累加

平均响应时间

在0-3s内：0.3

在3-6s内：0.25

在6-9s内：0.20

在9-12s内：0.10

在12s以上：0

**聊天过程总字数（40%）**

回答者总字数的统计，作为一个指标。

字数为0：0

字数1-35个：0.05

字数36-70个：0.1

字数71-210个：0.14

字数211-420个：0.18

字数421个以上：0.2

**自己是否填写评价内容（5%）**评分+评论：0.05

仅评分：0.03

自动评价：0

**是否进行分享（5%）。**

站内分享+2个社交分享：0.05

站内分享+1个社交分享：0.04

站内分享：0.03

2个社交分享：0.02

1个社交分享：0.01

不分享：0

（说明：目前的分享种类只有站内分享，没有社交分享）

**用户等级附加分（5%）**

激励等级高的用户为等级低的用户回答问题，这里给出一定的激励。根据用户的等级差，越大则增加越多。回答者等级 - 提问者等级 = 用户级差，最高为35，最低位0。只有级差为正时才具有附加分。

1-5：0.005

6-10：0.015

11-15：0.025

16-20：0.035

21-25：0.04

26-30：0.045

31-35：0.05

(说明：此部分暂时没有实现)

**暂停附加分（5%）**

统计由回答者触发的暂停动作，包括离线、点击暂停、超时未响应等。

暂停次数为零：0.05

暂停次数为1：0.025

暂停次数大于2：0

(说明：此部分暂时没有实现)

## 消耗量计算

在设计消耗量的时候，目标是通过消耗量来统计其获取的帮助，一方面是用于分享率的计算，另一方面控制给用户的权限。在计算消耗量的时候，主要考虑：1. 提问问题的数量。2. 获得回答问题的质量。消耗量必须等贡献量计算完成后才能进行计算。

消耗量=消耗量基数\*消耗量权值。

**每天提问问题的数量不同，每个问题的消耗量基数也不一样。**

每天提问的前3个问题，消耗量基数：1

每天提问的第4-5个问题，消耗量基数：1.1

每天提问的第6-7个问题，消耗量基数：1.2

每天提问的第8-10个问题，消耗量基数：1.25

每天提问的第11-15个问题，消耗量基数：1.5

每天提问的超过15个问题，消耗量基数：2.0

**评价对消耗量具有决定性的作用：**

如果对方评价为1星，消耗量权值=总得分\*2.0

如果对方评价为2星，消耗量权值=总得分\*1.5

如果对方评价为3星，消耗量权值=总得分\*1

总得分 = 2\*（回答者的总得分） **-** （提问者抵扣得分）

注：最低为0

抵扣得分由以下几个部分组成：

聊天中平均响应时间[不包含暂停的的时间]（30%）、聊天过程总字数（40%）、自己是否填写评价内容（10%）、是否进行分享（10%）、关键词分（10%）、附加分（5%）。

**聊天中平均响应时间的定义（30%）：**

假设A向B提问，在A和B的聊天过程中，统计在聊天过程中A对B聊天的响应速度，在每个建立聊天的问题中，日志系统每个answer带有两个属性：距离上一次answer（不管是谁的）的时间间隔、本次的answer的字数。对于第一次answer的时间间隔从建立聊天时算起，对于暂停的问题，从暂停的对话刚继续时算起（正常手机打字速度为20-60字/分钟）。

平均响应时间=每次answer的时间间隔总和/每次answer的字数累加

平均响应时间

在0-3s内：0.3

在3-6s内：0.25

在6-9s内：0.20

在9-12s内：0.10

在12s以上：0

**聊天过程总字数（40%）**

提问者总字数的统计，作为一个指标。

字数为0：0

字数1-35个：0.05

字数36-70个：0.1

字数71-210个：0.14

字数211-420个：0.18

字数421个以上：0.2

**自己是否填写评价内容（10%）**评分+评论：0.1

仅评分：0.06

自动评价：0

**是否进行分享（10%）。**

站内分享+2个社交分享：0.1

站内分享+1个社交分享：0.08

站内分享：0.06

2个社交分享：0.04

1个社交分享：0.02

不分享：0

(说明：此部分暂时没有实现)

**关键词分（10%）**

关键词填的不好要扣分，鼓励用户更谨慎、更认真的填写关键词。比如，标准化过程关键词有忽略、空值的情况。

关键词标准化前后数量一致：0.1

关键词标准化后数量变成零：0

关键词标准化后数量减少1个：0.8

关键词标准化后数量减少2个：0.5

关键词标准化后数量减少3个或以上：0.2

(说明：此部分暂时没有实现)

**暂停附加分（5%）**

统计由提问者触发的暂停动作，包括离线、点击暂停、超时未响应等。

暂停次数为零：0.05

暂停次数为1：0.025

暂停次数大于2：0

(说明：此部分暂时没有实现)

## 用户行为对于的知识量增量

每次用户行为增加的知识量增量是一定的，如下表所示。

|  |  |  |
| --- | --- | --- |
| 动作类型 | 说明 | 知识量增量 |
| 填写资料 | 审核通过后，每项 | 20 |
| 双方评价结束给提问者 | 双方评价结束后 | （回答者）贡献量\*0.5 |
| 双方评价结束给回答者 | 双方评价结束后 | （回答者）贡献量权值\*2 |
| 查看问题 | 即时 | 0.2 |
| 赞 | 即时 | 0.5 |
| 评论 | 即时 | 0.8 |
| 填写标签 | 审核通过后 | 10 |

能触发学习动作的资料目前包括：公司、职位、学校、专业。

由于对于比较资深的人士，活跃度普遍不高，但其知识面相对高，因此，根据资料等级会对用户的知识量在这个基准之上进行提升，乘以一个等级系数作为最终的知识量增量。需要乘以等级系数的动作包括：提问、回答、查看、赞、评论。下面为资料等级相应的等级系数。

35-31：5

30-26：4

25-21：3

20-16：2

15以下：1

(说明：资料等级体系的基准增加暂时没有实现)

## 知识量增量对于到用户标签权值的变化

用户的知识量增量确定后，接下来将确定给用户的哪些标签添加多少的权值。第一步是确定初始标签的初始知识量。第二步是对初始标签和关联标签的初始知识量进行再分配。

**第一步：**

针对不同类型的动作，初始标签及相应的初始知识量的分配方式有一定的差异：

1. 资料

对于资料通过审核后，会预先分配好树状结构中的目录，将初始标签定为：定义好的初始标签以及所有的子节点。初始知识量定义为：总的知识量/初始标签的数量。

1. 动作

始标签定为：问题对应的标准化关键词。

初始知识量定义为：总的知识量/初始标签的数量。

1. 添加标签

用户所添加的标签经过标准化后。

初始知识量定义为：填写标签所规定的知识量增量。

**第二步：**

按顺序完成每个初始标签和初始知识量的分配。假设初始标签为A，与A关联度值从高到底的排序，分别为B, C, D…，初始知识量为S0, 在第1次分配结束后剩余S1，第n词分配结束后剩余Sn。

1. 假设用户对于关键词A，权值为α（用户无标签则为0），增加权值余量的β（根据动作不同，β取值不同，如取值为10%）（如果1<β时，则增量为α），第1次分配结束后知识量剩余S1=S0 - (（1-α）\*β)。

2. 如果S1>0，则对关键词B重复第一步。接着进行C，D。。。

3. 如果Sn=0，完成对知识量的分配。

其中β的取值为：

|  |  |
| --- | --- |
| 动作类型 | β的取值 |
| 填写资料 | 10% |
| 提问 | 10% |
| 回答 | 8% |
| 查看问题 | 2% |
| 赞 | 3% |
| 评论 | 5% |
| 填写标签 | 20% |

## 用户活跃时间学习（暂时没有实现）

学习用户的活跃时间，主要是通过日志文件，统计用户在哪些时间段对APP进行了哪些操作，进而分析出，哪个时间段更合适推送、展示问题给用户，为推荐系统提供更好的效果。

用户通过设置的长期时间习惯，优先于系统学习的习惯。

统计的时间规则：按照一周为周期进行统计，一天则按每个小时分段，因此总共有7×24=168个统计时间段。分析用户周一到周日，在每个时间段内的行为，最后得出用户的哪些时间段更加活跃，给出一个得分，每天凌晨进行一次更新。

在每个时间段根据动作类型统计行为的次数（针对每个用户，有168个表），分三个统计周期。可以在全部的行为中记录动作发生的时间，每天通过脚本更新最近一个月、最近三个月的次数数据。

表格1单个时间段内用户行为次数的统计

|  |  |  |  |
| --- | --- | --- | --- |
| 行为类型\统计周期 | 最近一个月 | 最近三个月 | 全部 |
| 活跃 | 次数 | … | … |
| 积极 | 2 | … | … |
| 不活跃 | … | … | … |

通过上述行为的次数，可以分析两个数据：1. 用户在其个人习惯中哪个时间段更加活跃；2，用户在整体用户中的活跃度。目前我们只分析第1种情况。

统计用户在168个时间段的总数据，为上面168个数据表中次数的累加。

表格 2全部168个时间段内用户行为次数的总和

|  |  |  |  |
| --- | --- | --- | --- |
| 行为类型\统计周期 | 最近一个月 | 最近三个月 | 全部 |
| 活跃 | 次数 | … | … |
| 积极 | 53 | … | … |
| 不活跃 | … | … | … |

下表中用户每个时间段内行为百分比是根据前面两个表的数据计算出来(单个时间段的行为次数/总的时间次数)，在计算之前以一个负初始值（如-1）代替，若总的行为次数为0时，每个时间段该百分比为0。

表格 3 单个时间段内用户行为的百分比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 行为类型\统计周期 | 最近一个月 | 最近三个月 | 全部 | 粗加权百分比 |
| 活跃 | 百分比 | … | … | 粗加权百分比，排名 |
| 积极 | 3% | … | … | … |
| 不活跃 | … | … | … | … |

粗加权百分比 = 最近一个月百分比\*0.5 +最近三个月百分比\*0.4 + 全部\*0.1

精加权百分比 = （活跃的粗加权百分比\*2 +积极的粗加权百分比-不活跃的粗加权百分比）/2

通过用户的精权百分比和排名判断用户的活跃时间段。通过同时满足以下条件判断用户在这个时间段是否活跃：

|  |  |  |
| --- | --- | --- |
| 粗加权百分比 | 排名 | 用户活跃度 |
| 大于1/168\*2（平均值的2倍） | 在1-56之间 | 积极 |
| 大于1/168（平均值的1倍） |  | 活跃 |
| 小于1/168/2（平均值的1/2倍） | 在112-168之间 | 不活跃 |

# 附

## 资料等级

**用户资料等级**

通过资料等级，来判断用户的资深程度。因为资深的用户参与度不高，但其价值巨大。按重要程度将影响因素排列如下，每个影响因素都有相应的等级，资料等级由所有因素计分的总和。

1. 年纪/工作年龄：（二选一，优先“毕业时长”）

|  |  |  |
| --- | --- | --- |
| 年龄 | 毕业时长 | 级别 |
| 60+ | 40+ | 10 |
| 40~60 | 18~40 | 9 |
| 35~40 | 13~18 | 7 |
| 30~35 | 8~13 | 6 |
| 27~30 | 3~8 | 4 |
| 25~27 | 1~3 | 3 |
| 18~25 | 0~1 | 1 |
| 18以下 |  | 0 |

1. 单位（根据规模调整，最多上下移动一级）：

|  |  |  |
| --- | --- | --- |
| 单位级别 | 单位类型 | 级别 |
| 一级 | 政府 | 2 |
| 二级 | 国企 | 1.8 |
| 三级 | 事业单位 | 1.5 |
| 四级 | 外企 | 1.3 |
| 五级 | 私企 | 1.2 |
| 六级 | 其他 | 1 |

职位（根据实际情况适当调整，最多上下移动一级）：

|  |  |
| --- | --- |
| 处级/总经理/董事长 | 10 |
| 副处/总监/部长 | 8 |
| 科员/中层领导/经理 | 5 |
| 副科/技术研发/市场销售 | 3 |
| 一线工作人员 | 1 |

社会地位得分为：单位分\*职位分。

1. 学校/学历（择高选择）：

|  |  |
| --- | --- |
| 学校类型 | 得分 |
| 清北/海龟/博士 | 5 |
| 前9所/MBA/硕士 | 4 |
| 985 | 3 |
| 本科 | 1 |
| 其他 | 0 |

**资料等级的变化**

如果用户在同一等级受到的评价低，则资料等级会根据评价进行快速下调，并同时降低该用户的知识量的百分比。

如果用户向同一等级、更高等级的回答中，受到好评，将提升用户等级。

|  |  |  |
| --- | --- | --- |
|  | 提问 | 回答 |
| 评价为1星 | 降低1级 |  |
| 最近10次有3次1星 | 降低为最近10次回答者中最低的等级，再扣2级。 | 降低2级 |
| 最近5次有2次2星 | 降低1级 |  |
| 最近10次有5次2星以下 | 降低为最近10次回答者中最低的等级，再扣2级。 | 降低2级 |

## 用户标签结构

**用户标签结构**分五组：

1. 用户显示标签

用来显示在用户的资料中的，包括所有选择的、添加的标签，按分类有标准标签、非标标签、未审核标签。

2. 用户标准库（标准标签+权值）

用来描述用户的专业程度，将用户添加的非标标签映射到标准标签后，进行权值的更新。

3. 用户未审库（未审标签+权值）

用户自己提交的标签、自动添加的标签若不存在于标准库、非标准库和空值库中，则需要审核后才加入标准库中，待人工审核后会加入标准库或非标库。若人工审核认为该标签无意义，则加入空值库。

4. 用户衍生未推荐标签库

用户主动添加的标签、回答问题的标签、提问的标签所衍生出来，实际已经加入标准库，但可以推荐给用户的标签选择，从而加重权重。

5. 用户违规标签和资料处理

违规则删除标签，放入空值库，并通过系统通知提醒用户。