

# 用户产生内容的质量评价与排序

第6届R语言会议@2013-05-19

王浩@宝宝树

alexwhu@163.com



- **用户产生内容的形式、特点与属性**
  - UGC – User Generated Content
- **UGC质量评价的难点与挑战**
  - UGC质量评价与排序的相关工作讨论
- **宝宝树在UGC质量评价与排序的工作介绍**
  - 基于内容文字本身做质量评估
  - 基于辅助因素利用Logistic regression做排序
  - ppt里面全是数字，每个工作，都会以数据为支撑，以数据为导向
- **宝宝树其他部分算法工作简介**

## 用户产生内容的形式、特点与属性

# 1.1 用户产生内容的形式

## Web 1.0

网站生成固定内容

静态页面

桌面浏览器

简单 & 同步

## Web 2.0

用户自己生成内容

Mashup和Web服务

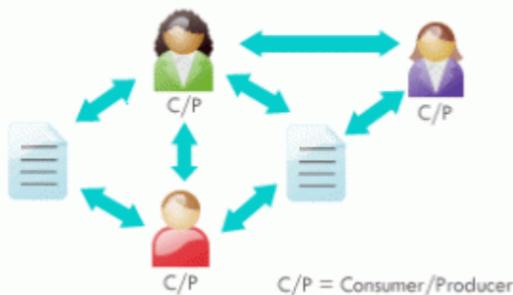
复杂的客户端软件

复杂 & 异步

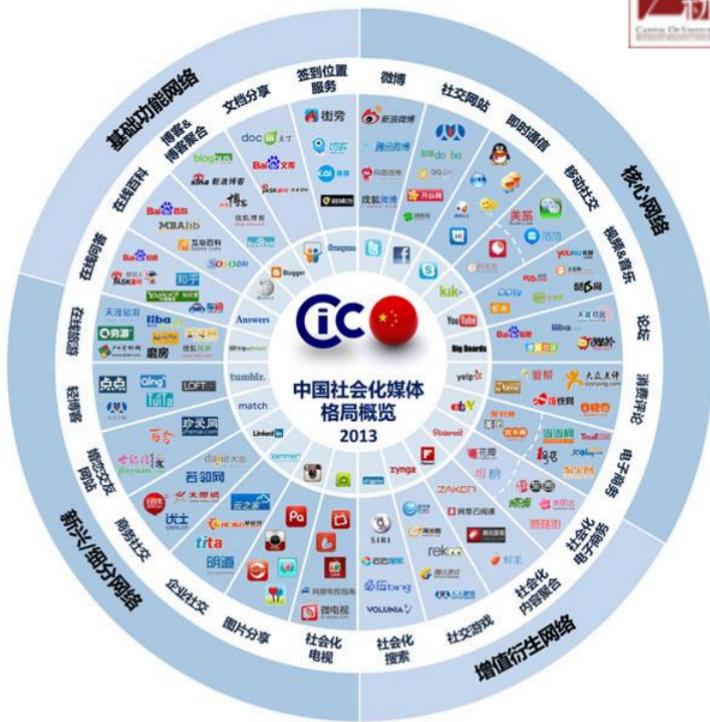
Web 1.0



Web 2.0



# 1.1 用户产生内容的形式



统计之都👑：【转发送书】本条微博若转发超过1000次，关注并转发本条微博的粉丝将有机会获得@华章计算机科学出版的《机器学习:实用案例解析》，这本书是穿着中文马甲的《Machine Learning for Hackers》。目前共5本可送。喜欢R语言的看过来，不喜欢R语言的就不用抢机会啦。



人均¥60

口味4(非常好) 环境4(非常好) 服务4(非常好)

百年龙鱼，味道鲜美，口感细腻，真的很不错，流连忘返啊，。。。



cainend

R in action 这也是目前最权威的了 ★★★★★

还是先有点基础 才能看这本书



生命的平明

宝宝发烧后出疹子怎么办？

👤 0 当时年龄：6个月9天 提问时间：2012-08-07 14:02 关闭时间：2012-08-17

宝宝6个月了，前两天发烧吃了退烧药好了，但是发现宝宝身上起了好多小红点脸上也有，看着很痒，很是着急。听家里老人说不能见风是吗，该怎么办？宝妈们帮帮我！

那应该是幼儿急疹，就是发了三四天高烧之后会全身长红疹子，注意不能吹风哦。



2,086



772

签到



(6)

转发(865)

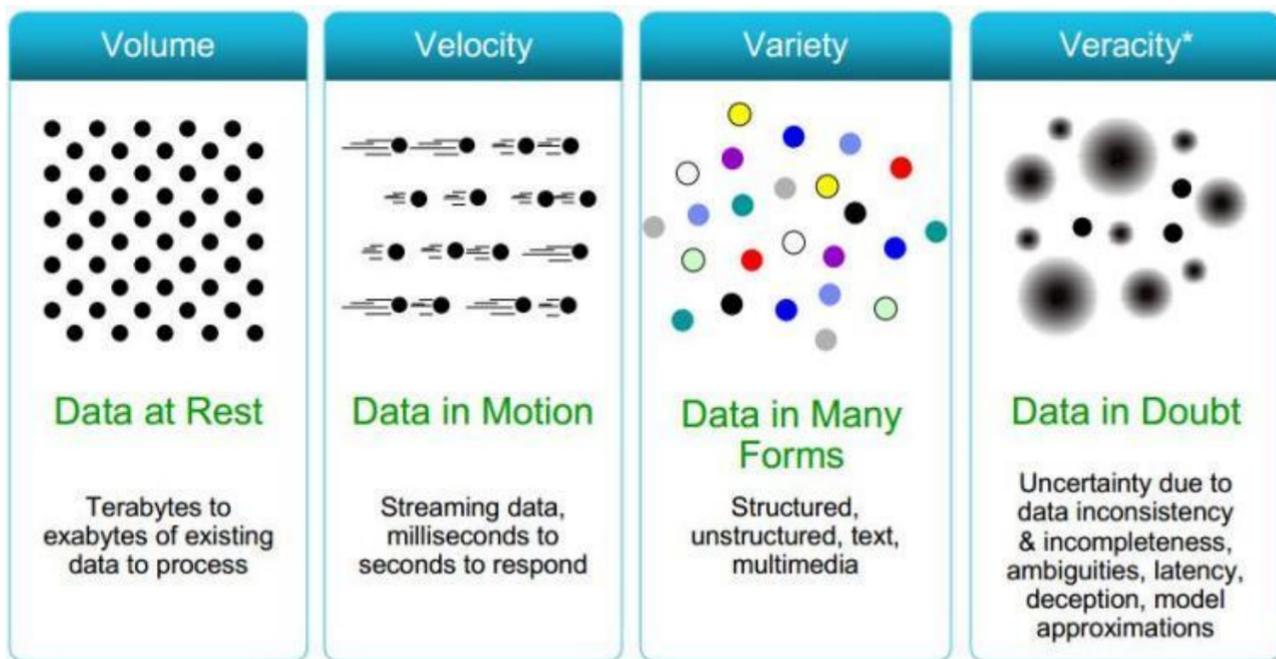
收藏

评论(113)



## 1.2 用户产生内容的特点

- 大数据的4V特点，就是UGC的特点



## 1.3 用户产生内容的属性

- UGC普遍都是短文本 & 用户拼写错误较常见
- UGC之间，基本上没有链接交互性
  - UGC依赖而又不单一存在于单一的web page中
  - 在产品形态上，UGC，作为媒体属性，通常存在于某个web page之中，而不是作为单独的page存在
  - 多条UGC，可能会共存于同一个web page之中
- 对某个评论（UGC）的赞（UGC），算是UGC之间的链接



UGC质量评价的难点与挑战

UGC质量评价与排序的相关工作讨论



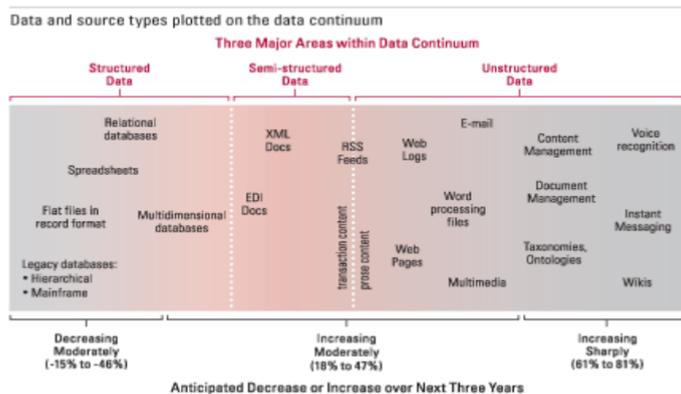
## 2.1 UGC质量评价的难点与挑战 – 场景适用

- **不同应用场景下的UGC质量评价与排序，因诉求不同，难点与挑战也有显著差别**
  - Digg与Reddit等用户推荐文章类服务
  - 新浪微博智能排序、Facebook NewsFeed
  - 社区问答类产品回答质量评价排序
  - 社交搜索、传统搜索结果中的社交媒体UGC评价
  - .....

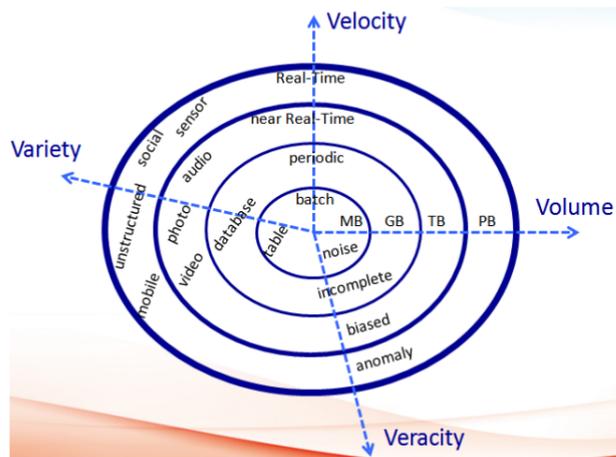


## 2.1 UGC质量评价的难点与挑战 – 大数据4V

- UGC的4V特点，给内容质量评价带来天然的技术挑战
- 除常规大数据处理技术挑战外，UGC内容质量评价难点特别在于：
  - 非结构化数据处理，以及与结构化数据的有机整合
  - UGC数据中“噪音”数据的识别、缺失值处理，及对模型影响的感知

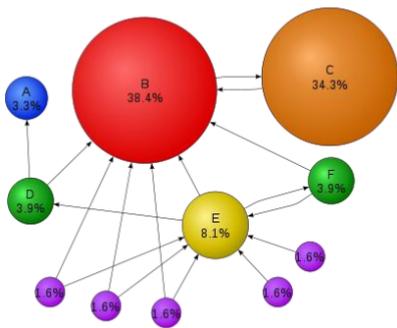


The data clearly signals a shift toward using more semi-structured and—especially—unstructured data sources.



## 2.1 UGC质量评价的难点与挑战 – UGC短小&又无链接

- UGC之间，难以像web pages那样，去描述与建模
  - UGC通常文本很短，传统web page文本较长
  - UGC间无直接的链接，web pages间有链接
    - 多个UGC共同出现在同一个页面中，多多少少算一种非直接的链接



Google PageRank vs. Facebook EdgeRank

已解决 ❤ 我喜欢 0

 怀孕四个月差几天，这几天小肚子有种说不出的感觉。有东西挪动还是什么，感觉好奇怪，是胎动吗？

1450532 👍 0 当时年龄：孕16周3天 提问时间：2013-01-12 06:29 解决时间：2013-01-17

来自iPhone客户端 浏览4001次 回答746个  
[关键词：怀孕四个月小肚子东西好奇胎动]

共有746个其他回答

 是哦！刚开始动静小、过些天宝宝会越来越大、胎动会越来越明显的、幸福的准妈妈、祝您好孕

枫叮头 来自Android客户端 | 2013-01-12 06:51

 是胎动，一般刚开始的时候都比较感觉像是小鱼游来游去的，等6个月以后就会明显了

我的宝宝我的家 来自网页 | 2013-01-12 09:39

<http://www.babytree.com/ask/detail/4889534>

## 2.2 UGC质量排序示例

- **UGC质量评价，主要关注于以文本内容形式存在的UGC**
- **示例包括：**
  - **Question-Answer community中的回答质量，如宝宝树育儿问答**
  - **E-commerce中的商品评论，如淘宝、亚马逊电商**
  - **Social-Graph SNS中的信息流排序，如微博智能排序、FB信息流**
  - **Interest-Graph SNS中的UGC，如豆瓣评论**

## 2.2 UGC质量排序示例：问答服务

- Question-Answer community中的回答质量，如宝宝树育儿问答

已关闭 ❤ 我喜欢 0

 宝宝半个月了，还没长体重~

👤 0 当时年龄：15天 提问时间：2012-07-15 19:11 关闭时间：2012-07-24

萌Mommy

奶水不够，每次吃奶要吸很久，而且有时候几分钟十几分钟又开始哭，混合喂养中，牛奶中偶尔加清火宝，拉的便便很稀，水水样，而且还有些漏尿水，每天要放很多屎屁屁，睡觉也睡的不安稳，动个不停，哼哼唧唧的，有时候一个上午都不睡觉，很担心是不是没吃饱或者是奶水不好，会不会影响宝宝生长啊，这么小应该是需要睡能吃才长的快吧。需不需要补充钙和鱼肝油了。

来自网页 浏览2272次 回答10个  
[关键词：补充 清火宝 便秘 漏尿 屎屁屁 牛奶 宝宝半个月了]

 推荐回答

 半个月一般看不出来，鱼肝油可适量地补充一些~~一个月后再称，一般按月来的！

来自米卡早教专家 | 2012-08-17 17:57

 0



### 共有9个其他回答



wqed

如果总拉稀就不用加去火的了，混合喂养多喝点水就行了，不用专一去火。再有给宝宝吃点乳酶生片就行了。如果奶水不够就多喂点奶粉吧，别饿着宝宝

来自网页 | 2012-07-15 19:18



atu0806

你家便便稀就暂时不要加清火宝哦，宝宝就是要尽量保证睡眠哦，半个月以后就可以补鱼肝油了

来自网页 | 2012-07-15 19:19



豆琪儿

我觉得一般小宝宝吃饱了会两三个小时吃一次奶的啦，你可以注意观察看看啊。小宝宝是不是吃饱啦。

来自网页 | 2012-07-15 19:16

## 2.2 UGC质量排序示例：问答服务

- Question-Answer community中的回答质量，如宝宝树育儿问答

已关闭 ❤ 我喜欢 0

 问下大家苹果手机怎么看黄片？谢谢 换个问题看看>>

 i288547 👤 0 当时年龄：1个月11天 提问时间：2012-08-31 23:33 关闭时间：2012-09-09

来自iPhone客户端 浏览12101次 回答6个  
[关键词：苹果手机怎么看黄片]

 推荐回答

 看孩子要紧，看片就算了，才一个多月你也不适有什么性生活啊 来自米卡早教专家 | 2012-09-01 09:58

 0 @耗资小王  
weibo.com/alexwhu

安心儿妈妈

## 2.2 UGC质量排序示例：电商评论

### • E-commerce中的商品评论，如淘宝、亚马逊电商

#### 按有用程度排序

9/9 人认为此评论有用

☆☆☆☆☆ 送货速度不错 2012年12月20日

评论者 harypotter

购买过此商品

刷卡机该换了，刷个卡没信号走了半天。  
还有手机电源键有点问题。周末还要去苹果店修一下。

回应 | 这条评论对您有用吗？  是  否

140/159 人认为此评论有用

☆☆☆☆☆ 我在亚马逊的经历 2013年2月19日

评论者 风向

购买过此商品

2月17日下单，2月18日晚收到手机。物流很快，但是商品却很烂

在网上看了些日子，最终选在亚马逊购买是因为亚马逊这三个字

IPHONE5收到后，拆封不到十二个小时，就发现屏幕上方有整片

#### 按发表时间排序

☆☆☆☆☆ 不错 还可以  
包装很好 是正品 速度可以 还可以。用的还不错！  
陈龙 在2天前发表 吴志勇 在2天前发表

☆☆☆☆☆ 还不错！开箱后边框有点瑕疵！！  
懒得换了！！  
还不错！开箱后边框有点瑕疵！！懒得换了！！  
喜洋洋 在4天前发表

☆☆☆☆☆ 喜欢才买  
最终还是买了iphone5，喜欢它的纤细，轻薄。  
秋日的私语 在4天前发表

☆☆☆☆☆ 挺好的  
挺好的，用了一段时间没有什么其他的不好反应，  
就是比安卓的系统要稳定！！  
李金鑫 在5天前发表

亚马逊  
amazon.cn

## 2.2 UGC质量排序示例：电商评论

- E-commerce中的商品评论，如淘宝、亚马逊电商

商品详情 | 累计评价 46286 | 月成交记录 4025件 | 给我推荐

与描述相符  
**4.8**  
★★★★★

1 非常不满 | 2 不满意 | 3 一般 | 4 满意 | 5 非常满意 (4.8)

查看追加 (1080)  有内容评价

按时间 ↓ | 按信用 ↓ | 按推荐 ↓

按买家信用等级从高到低进行排序

按评价内容丰富程度进行推荐排序

物流很给力，收到后马上试用，非常OK，性能方面，还有待观察，先给好评，全五星  
05.10

挺好的  
05.10

很好，很强大。穿透能力挺不错的  
05.10

天猫 TMALL.COM

## 2.2 UGC质量排序示例：微博排序

### • Social-Graph SNS中的信息流排序：微博智能排序

#### 什么是智能排序

根据关注、标签和微博内容等相关信息，帮助用户梳理微博内容，对同类微博进行合并、对可能感兴趣的微博内容进行优先展示的排序的功能。



#### 智能排序能给大家带来什么？

高效



提升阅读效率，第一时间找到你感兴趣的微博，最关心的内容不容错过。

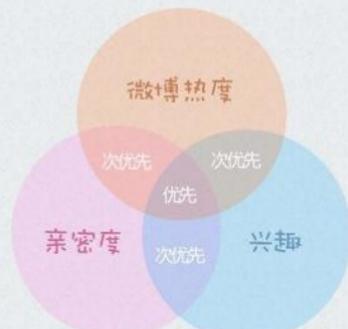
降噪



将重复信息合并，剔除垃圾信息，你的微博内容也可以小清新。

#### 智能排序如何计算的

智能排序主要计算元素：微博热度、亲密度、兴趣



## 2.2 UGC质量排序示例：NewsFeed

- Social-Graph SNS中的信息流排序：FB信息流
- From Will Cathcart, Facebook NewsFeed PM :
  - Can we better predict what people are looking at?
  - Can we better predict what people won' want to see or are less likely to interact with?

# WHAT IS EDGERANK?

*EdgeRank is an algorithm developed by Facebook to govern what is displayed—and how high—on the News Feed.*

### WHAT IS THE EDGERANK ALGORITHM?

EDGERANK

$$\sum_{\text{edges } e} u_e w_e d_e$$

**u** Affinity score between viewing user and edge creator  
**w** Weight for this edge type (status, comment, like, tag, etc.)  
**d** Time Decay factor based on how long the edge was created

facebook.

[www.facebook.com/EdgeRankAlgorithm](http://www.facebook.com/EdgeRankAlgorithm)

## 2.2 UGC质量排序示例：产品评论

### • Interest-Graph SNS中的UGC质量，如豆瓣评论

#### 推荐系统实践



作者: 项 高  
出版社: 人民邮电出版社  
出版年: 2012-6  
页数: 197  
定价: 49.00元  
装帧: 平装  
丛书: 图灵原创  
ISBN: 9787115281586

读书笔记 ····· (共67篇)

按有用程度 按页码先后 最新笔记

热门评论 最新评论

附上Reference 中的干货 (Paper,Blog等资料的链接) ▼

wacow (别太暴躁 再暴躁就可以报复社会) ★★★★★

这只是一本197页的书 我想你未必过瘾 但作者附上了诸多好资料 无论是paper, blog文章, wikipedia 词条, 数据集还是开源项目等 你可以选择拥有 附上我收集的资料链接, 格式基本按照 URL+资料名称+出现在书中的页数, 某些链接可能需要你翻过一道墙, 某些重复引用的我就没重复贴上链接了  
[http://en.wikipedia.org/wiki/Information\\_overload](http://en.wikipedia.org/wiki/Information_overload) P1  
[http://www.readwriteweb.com/archives/recommen.....](http://www.readwriteweb.com/archives/recommen...) (32回应)

2012-07-21 2:30 75/77有用

自己来冒个泡 ▼

xivector ★★★★★

这本书大约写了10个月的时间, 如果一定要自己评价一下这本书, 只能说还行。这本书基本达到了写作目标: 1. 帮助刚毕业的学生迅速了解如何将他们学到的理论用于实际 2. 帮助程序员迅速将他们的编程能力应用到推荐系统中来 3. 强调数据分析的重要性, 淡化算法 4. 运用多种评测方法, 强调全面评测的重要性 不过本书也有一些遗憾, 如果将来会再版这本书, 可以修正这些遗憾: 1. 推荐系统和搜索引擎不同, 他还没有一个统..... (12回应)

2012-06-24 8:02 37/38有用

最有用的好评

附上Reference 中的干... 75/77有用

★★★★★ wacow 2012-07-21

这只是一本197页的书 我想你..... (查看原文)

更多 5星(2条), 4星(5条)的评论

最有用的中差评

比较失望, 远远没有宣... 9/15有用

★★★★☆ 晓东 2012-07-05

本来书还不到200页, 大部分章节..... (查看原文)

更多 3星(4条), 2星(1条), 1星(3条)的评论

豆瓣 douban

## 2.2 UGC质量排序示例：社会化问答&新闻投票

### • Interest-Graph SNS中的UGC质量：知乎回答

互联网公司的**数据科学家**（Data Scientist）职责和日常工作内容什么？

26 票 王浩：个人从事**数据**算法相关工作，一些个人见解如下：**数据科学家**工作可以包括3个方面：1、对历史**数据**的处理平台搭建：具体就是公司的基础**数据**平台建... [查看](#) »

3 个回答 227 人关注 · 取消关注

<http://www.zhihu.com/question/20935226>

按票数排序

按时间排序

### • 投票机制缺陷，曾经毁掉一家优秀公司 - Digg



**digg** I will digg your 40 links by my digg account with 500 plus follower for \$5  
I will digg your post by my digg account. It means... (by seodirectory)

Read more Collect Share order now!

**Digg Votes** I will get you 103+ real and Legit digg votes for \$5  
Digg is Power Social News Media... All Search engines... (by seo\_lee)

103+ Diggings Read more Collect Share order now!

更多投票机制的优缺点和分析，可参看阮一峰的《基于用户投票的排名算法》系列博客：

- 2012.10.16：贝叶斯推断及其互联网应用（三）：拼写检查（26条评论）
- 2012.03.28：基于用户投票的排名算法（六）：贝叶斯平均（17条评论）
- 2012.03.20：基于用户投票的排名算法（五）：威尔逊区间（32条评论）
- 2012.03.16：基于用户投票的排名算法（四）：牛鞭冷却定律（18条评论）
- 2012.03.11：基于用户投票的排名算法（三）：Stack Overflow（15条评论）
- 2012.03.07：基于用户投票的排名算法（二）：Reddit（26条评论）
- 2012.02.24：基于用户投票的排名算法（一）：Delicious和Hacker News（31条评论）



## 宝宝树在UGC质量评价与排序的工作介绍

- 推荐系统协同过滤思想在内容质量评估中的应用
- Logistic regression做排序
- 每个想法，都会以数据为支撑，以数据为导向

## 3.1 育儿问答 – 产品介绍

- **宝宝树是全球访问量最大、最受欢迎的垂直母婴网站**
  - 每个月超过5500万独立用户访问，覆盖80%的婴幼儿妈妈群体
- **宝宝树育儿问答：**
  - <http://www.babytree.com/ask/>
- **产品理念：**
  - 过来人妈妈帮助新手妈妈，互助问答
  - 第一时间、随时随地，让千万妈妈帮您排忧解难
- **产品特色：**
  - 问答集中在母婴领域，以孕育年龄段区分问题与用户
  - 与宝宝树众多其他产品打通，用户间交流互动性强



## 3.1 育儿问答 – 用户属性

- 妈妈们的兴趣图谱，随时间而快速迁移
  - “妈妈们”的双重用户身份，自己 & 婴幼儿（胎儿）
  - 整个孕、育过程中，“妈妈们”的话题关注点，会随着宝宝（胎儿）的自然年龄的增长，而发生自然而又快速的迁移

对比关键词“叶酸 唐氏筛查 见红 湿疹 辅食 断奶”在 ask 讨论的热度曲线



## 3.2 技术需求

- **产品基本数据：**
  - 提问数量 - 每天2本10万个（育儿）为什么
  - 回答数量 - 1:10的平均回答率
  - 回答速度 - 10分钟必有回答
- **回答的质量，如何提升和追踪？**
  - 妈妈们带着需求来到网站，期望迅速得到精炼的靠谱答案
  - 回答量很大，纯文字信息，需要仔细阅读后才能分辨靠谱与否
- **如何利用算法自动判断内容质量，并优先呈现优质内容？**



## 3.3 现状与数据

- 如果将对回答内容的质量判断任务，留给用户
  - 无为而治：浏览用户自行判断和决定内容，是否靠谱
  - 中心决策：提问者主动在获得的回答中，选出“自认为”的“最佳回答”
  - 集体智慧：其他用户为答案投票，根据票数选出“最佳回答”
  - 集体智慧：浏览用户针对“最佳回答”进行“赞”的操作

 最佳回答

  
女儿璐璐

亲。宝宝拉肚子，大多数是和喂养不当和饮食有关。亲平时要注意自己的饮食，不要吃上火的东西，和凉的东西，这样都会直接影响到宝宝的。

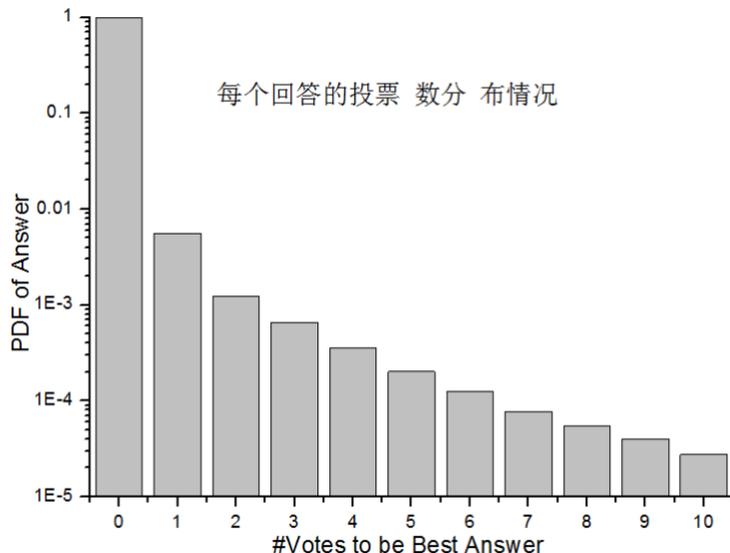
亲多注意给宝宝补水，不要让宝宝缺水。

来自网页 | 2010-04-26 14:48

 27

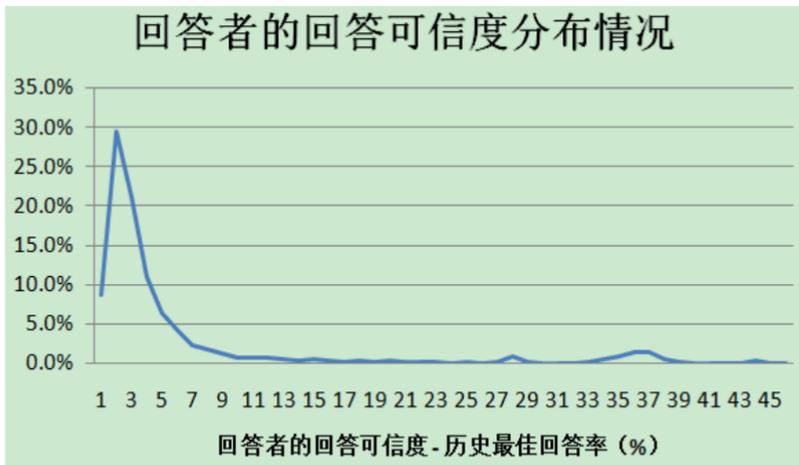
## 3.3 现状与数据

- “懒”用户的数据
  - 中心决策数据：有最佳回答的问题比例，约占25%
    - 我们期望能够对全部问题都置顶展示一个“最佳回答” – 100%
  - 集体智慧：99.2%的回答，0投票，0.5%的回答，仅有1个投票“赞”



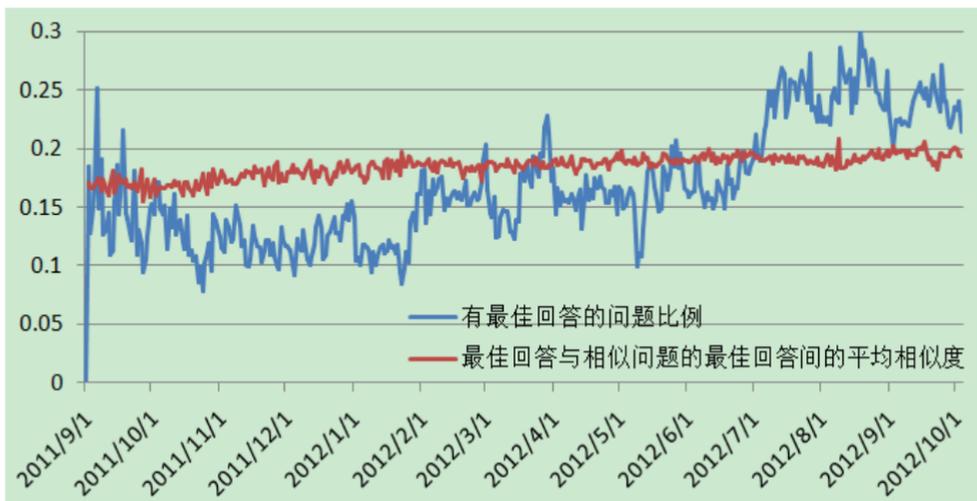
## 3.3 现状与数据

- 完全依赖于用户对育儿回答的质量进行判断，难度较大
  - 育儿问答是快问快答型产品（快餐店）
  - 区别于知乎、Quora等问答平台（西餐厅）



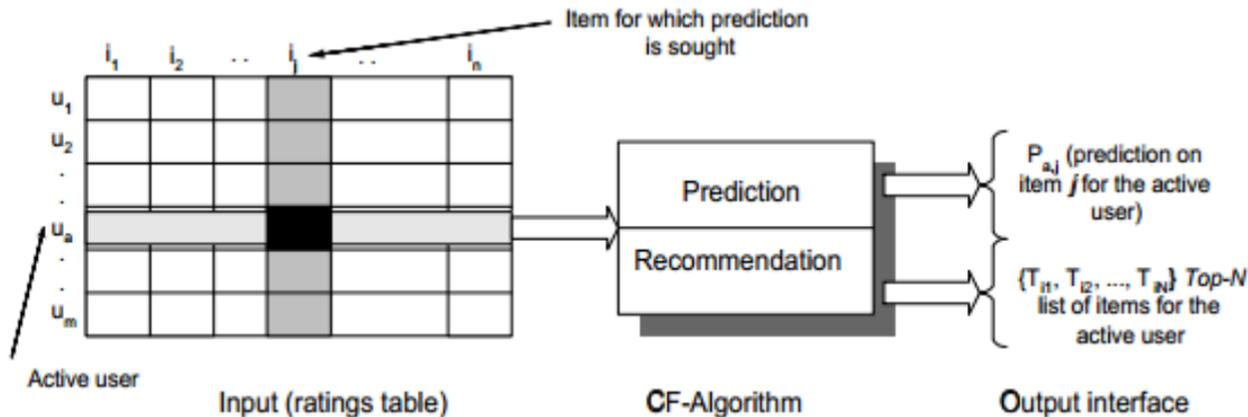
## 3.4 技术方案 - 算法自动判断内容质量，并优先呈现优质内容

- 最原始的想法：相似的问题，有相似的回答
- 进一步想法：相似问题的最佳回答，也相似
  - 适合的场景：居家、生活、育儿类等，无标准答案、讨论型的问答产品
  - 不适合场景：技术、科普、历史类等，有确切答案、结果型的问答产品



## 3.4 技术方案 - 算法自动判断内容质量，并优先呈现优质内容

- 最原始的想法的背后来由：推荐系统中的协同过滤思想
- 把每个问题，看作常规推荐系统中的User
- 把每个回答，看作常规推荐系统中的Item



[http://www.grouplens.org/papers/pdf/www10\\_sarwar.pdf](http://www.grouplens.org/papers/pdf/www10_sarwar.pdf)

## 3.4 技术方案 - 算法自动判断内容质量，并优先呈现优质内容

- 最原始的想法的背后由来：推荐系统中的协同过滤思想，但是：
  - 问题 $Q_i$ 的回答列表，与其他问题 $Q_j$ 的回答列表，完全不重合，所以不存在协同
  - “协同过滤” -> “内容相似”：基于内容相似度来评估回答质量

|       | $A_1$ | $A_2$ | ... | ... | $A_j$ | ... | ... | $A_n$ |
|-------|-------|-------|-----|-----|-------|-----|-----|-------|
| $Q_1$ |       |       |     |     |       |     |     |       |
| $Q_2$ | 黄色    | 红色    |     |     |       |     |     |       |
| ...   |       |       |     |     |       |     |     |       |
| ...   |       |       | 黄色  | 红色  |       |     |     |       |
| ...   |       |       |     |     |       |     |     |       |
| $Q_i$ |       |       |     |     | 蓝色    | 蓝色  | 紫色  | 蓝色    |
| ...   |       |       |     |     |       |     |     |       |
| ...   |       |       |     |     |       |     |     |       |
| $Q_m$ |       |       |     |     |       |     |     |       |

$Q(1-m)$ : 问题       $A(1-n)$ : 回答

- 红色: 已有问题的提问者选出的最佳答案
- 黄色: 已有问题的其他回答(非最佳答案)
- 蓝色: 当前问题的所有回答
- 紫色: 当前问题的潜在“最佳答案”

$m \times n$

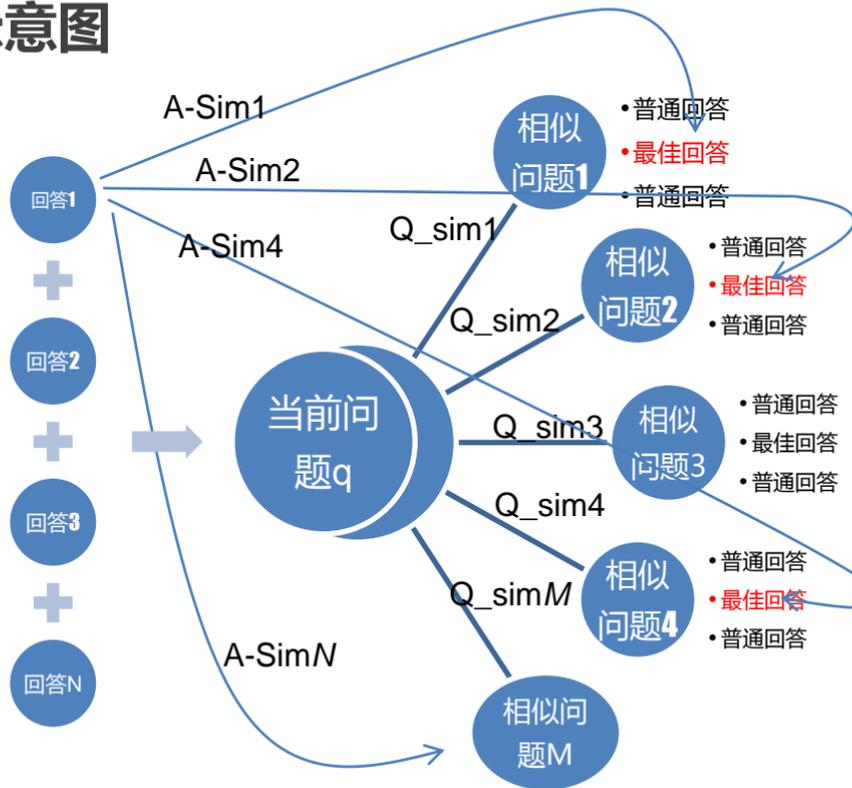
- 相似的问题，有相似的回答
- 相似问题的最佳回答，也相似

1. 找到相似问题 - 相关问题推荐  
✓找问题邻居
2. 计算回答与相似问题最佳回答之间的相似度  
✓找回答邻居
3. 评估得分后，糅合其他数据并排序

- 回答机器人，迅速回答新问题

## 3.4 技术方案 - 算法自动判断内容质量，并优先呈现优质内容

### 示意图



$$Score(q,i) =$$

$$\sum_{j=1, 2, \dots, M} Q\_sim(q, j) * A\_sim(i, j),$$

where

$$j = 1, 2, \dots, M,$$

for

the  $i$ -th Answer of Question  $q$ ,

$$i = 1, 2, \dots, N,$$

$Q\_sim(q, j)$ : similarity between Question  $q$  and Question  $j$

$A\_sim(i, j)$ : similarity between Answer  $i$  of Question  $q$  and the best Answer of Question  $j$

Then recommend the answer with  $\max\{ Score(q,i), i = 1, 2, \dots, M \}$  and sort answers based on  $Score(q,i)$

#### 推荐回答



怀孕后比较明显的特征就是爱睡觉，这个是正常的，不过宝妈不要睡太长时间啊，也要多走动走动，有助于消化。

朱Anggu

来自米卡早教专家 | 2012-06-25 15:17



## 3.5 技术实现 – 内容质量评估

- 找到相似问题 – 即相关问题推荐 – 主要基于内容 – 未来考虑点击数据&用户信息（兴趣点、当前年龄）做全面个性化推荐



为什么小孩经常感冒？

0 当时年龄：还没有宝宝 提问时间：2011-08-13 15:55 解决时间：2011-08-16

酷哥冰山王子

为什么小孩经常感冒？

### 相关问答

- 我家小孩为什么经常感冒 7人回答
- 小孩经常感冒怎么办 3人回答
- 宝宝经常感冒 7人回答
- 4到5岁小孩经常发烧感冒是怎么回事 13人回答
- 孩子为什么老是感冒 15人回答
- 小孩经常感冒 6人回答
- 宝宝怎么经常好感冒啊 14人回答
- 我家宝宝经常感冒怎么办？ 1人回答
- 小孩感冒 6人回答
- 小孩经常生病是什么原因 8人回答

TF-IDF空间向量模型  
LDA topic模型  
WAND算法加速

|                       | 平均需要计算相似度的doc数目 | 计算时间 (s) | 平均每秒计算文档数 | 160w文档共需天数  |
|-----------------------|-----------------|----------|-----------|-------------|
| 两两计算相似度的方法 (baseline) | 1,600,000       | 18567    | 0.108     | 171.9166667 |
| 直接使用倒排索引              | 323,675         | 3700     | 0.541     | 34.25925926 |
| WAND TOP-1000         | 47,524          | 463      | 4.320     | 4.287037037 |
| WAND TOP-500          | 33,204          | 325      | 6.154     | 3.009259259 |
| Wand Top-100          | 13,900          | 200      | 10.000    | 1.851851852 |
| Wand TOP-50           | 9,623           | 183      | 10.929    | 1.694444444 |
| Wand TOP-30           | 7,438           | 174      | 11.494    | 1.611111111 |

## 3.5 技术实现 - 内容质量评估

- 找到相似问题 – 即相关问题推荐
  - 育儿问答，提问全是短文本，比微博还短...
    - 提问标题加权 + 提问内容 + (最佳回答)
  - 类似重复提问较多，如何解决相关性好多样性差
    - Topic model的一些初步工作
  - 计算耗时，WAND算法加速原理：

$$\text{sim}(d_1, d_i) = \frac{d_1 \cdot d_i}{\|d_1\| \|d_i\|} = \frac{\sum_{t \in d_1 \cap d_i} w_{1t} \times w_{it}}{\sqrt{\sum_{t \in d_1} w_{1t}^2} \cdot \sqrt{\sum_{t \in d_i} w_{it}^2}} = \sum_{t \in d_1 \cap d_i} \alpha_{1t} w'_{it}$$

$$\alpha_{1t} = \frac{w_{1t}}{\sqrt{\sum_{t \in d_1} w_{1t}^2}} = \frac{w_{1t}}{\|d_1\|}$$

$$w'_{it} = \frac{w_{it}}{\sqrt{\sum_{t \in d_i} w_{it}^2}} = \frac{w_{it}}{\|d_i\|}$$

在查找候选的Doc过程中做一个近似的评估，跳过那些理论上不需要再考虑的文档，只对进候选的文档进行相关性计算

- 首先预估待评估文档的相关性上界
- 当上界超过当前结果集中相关性最小的才进行全面评估

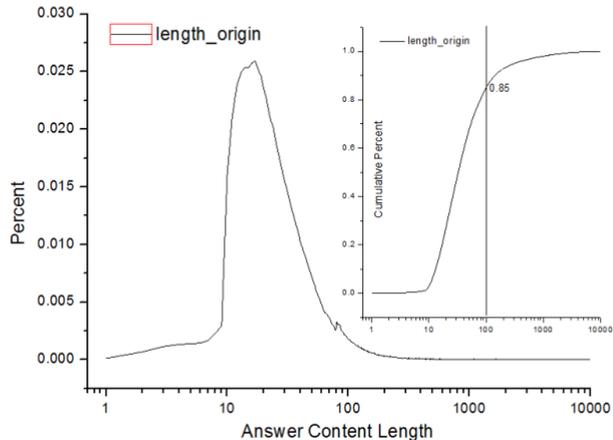
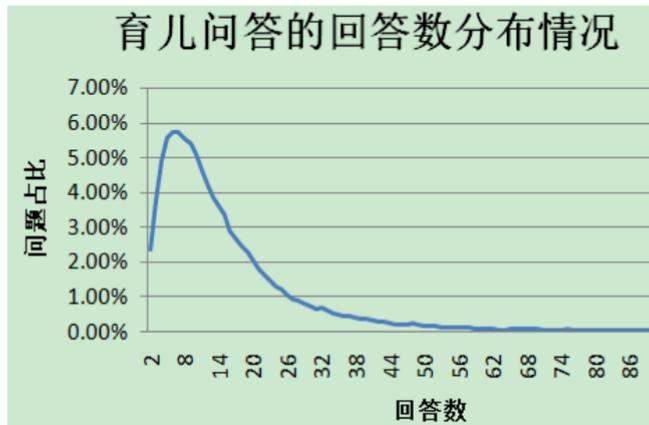
一次计算过程中， $\alpha_{1t}$  固定，如果知道  $w'_{it}$  的上界  $UB(w'_{it})$ ，那么  $\text{sim}(d_1, d_i)$  的上界将确定

$$UB(w'_{it}) = \max\left\{\frac{w_{it}}{\|d_i\|}, t \in D\right\}$$

$$UB(\text{sim}(d_1, d_i)) = \sum_{t \in d_1 \cap d_i} \alpha_{1t} UB(w'_{it})$$

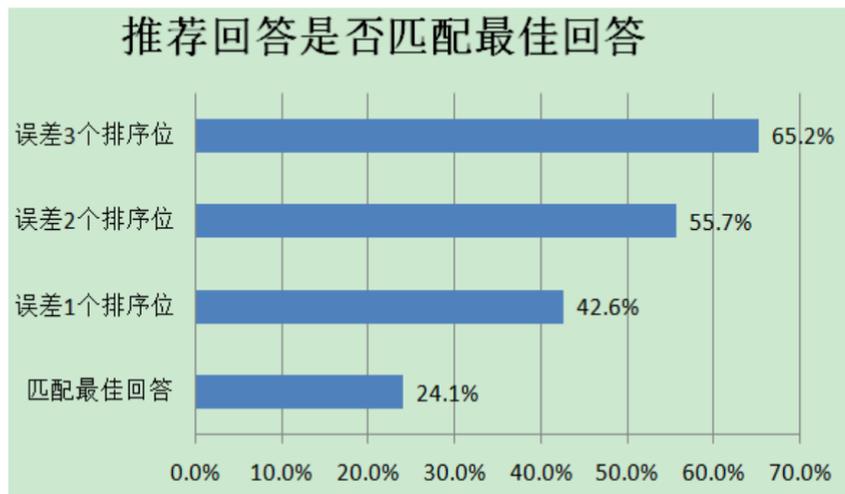
## 3.5 技术实现 - 内容质量评估

- 计算回答与相似问题最佳回答之间的相似度
  - 内容上的相似度
- 计算量相对较小
  - 只与相似问题中的Top  $N$  的问题的最佳回答计算相似度
  - 每个问题的回答数呈正态分布，平均回答数为10
  - 每个回答长度值呈log-normal分布， $Q1=19$ ， $Q2=33$ ， $Q3=65$ ， $Mean=120$



## 3.6 内容质量评估的效果数据

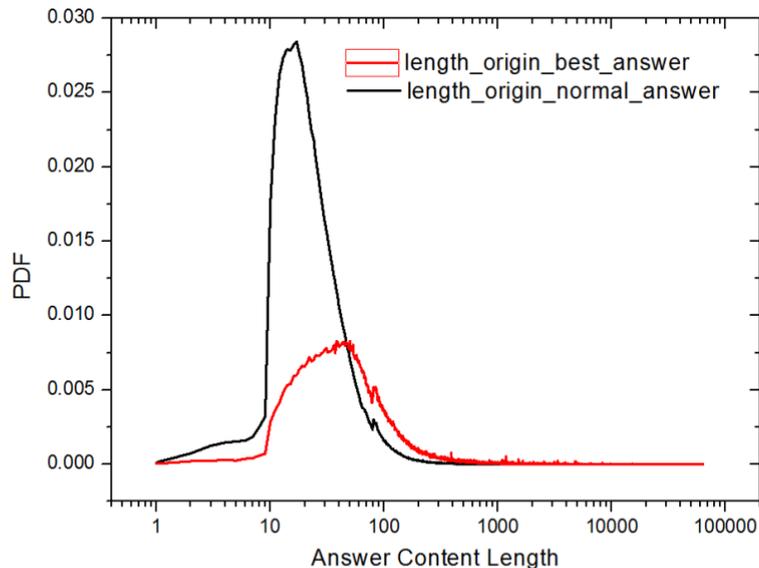
- 以提问者自主选择的最佳回答作为参考，计算推荐回答的匹配程度
  - 对每个回答，以 $S(q,i)$  得分倒序排序
  - 排在第1位，称作“推荐回答”
  - 考察最佳回答，落在回答倒序表中的位置，并计算对最佳回答的覆盖比例



## 3.7 技术实现扩展 – 内容质量排序

- 考虑因素1：回答长度
- 最佳回答，倾向于有较长的内容 ( log-normal )

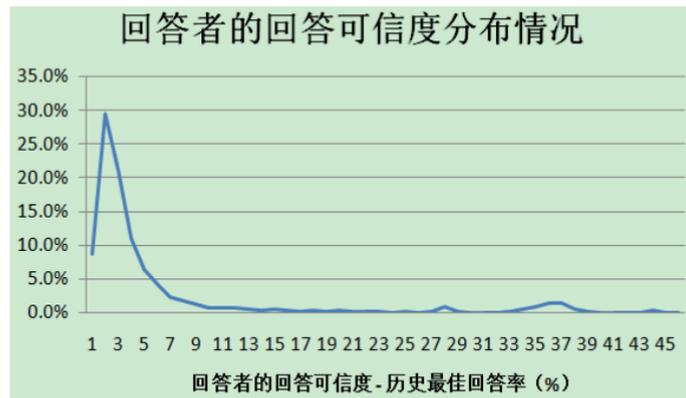
|         | Normal Answer | Best Answer |
|---------|---------------|-------------|
| Min.    | 1             | 1           |
| 1st Qu. | 19            | 45          |
| Median  | 30            | 86          |
| Mean    | 92            | 340         |
| 3rd Qu. | 55            | 207         |
| Max.    | 58029         | 63300       |



## 3.7 技术实现扩展 – 内容质量排序

- 考虑因素2：回答者的信誉度
- 最佳回答者，倾向于有较高的回答可信度
  - 可信度是指，回答者历史回答被选为最佳回答的比例

|        | 回答者的回答可信度均值 |
|--------|-------------|
| 最佳回答者  | 11.59%      |
| 非最佳回答者 | 3.98%       |
| 回答者    | 4.81%       |



## 3.7 技术实现扩展 – 内容质量排序

- 考虑因素3：回答内容与提问内容相似度
- 没什么信息量的回答，倾向于问什么、答什么
  - 不太好举例，就是一种产品感觉~~靠不靠谱需要数据验证

 碱性食品有那些  
👤 0 当时年龄：还没有宝宝 提问时间：2011-05-02 07:43

幸福的方圆 碱性食品包括什么

 碱性食品很多啊，蔬菜和水果碱性的多些

不辣小姜妈



|       | 回答内容与提问内容平均相似度 |
|-------|----------------|
| 最佳回答  | 19.31          |
| 非最佳回答 | 18.05          |

 宝宝湿疹长什么样？要去医院看吗？  
👤 0 当时年龄：15天 提问时间：2012-11-17 10:33 关闭时间：2013-01-12

four

 最好去医院检查一下吧

星星属于猪



## 3.8 内容质量排序方法

- 回答排序时，除回答质量本身，还考虑以下3个因素

| 排序因素         | 度量          |
|--------------|-------------|
| 回答长度         | 汉字个数 / 10   |
| 回答者可信度       | 最佳回答率 * 100 |
| 回答内容与提问内容相似度 | 相似度 * 100   |

- 用logistic regression对回答内容按质量排序 – 转换成分类问题
  - 预测一个回答，是否能够成为最佳回答，类似于广告点击预测中的CTR预估
  - 或者说，把所有回答分为2类：最佳回答（类仅有1个回答），其他回答（剩余）

```
logisitic_data_rec_answer <- glm(match~ feature_answer_content_similar + feature_answerer_credit +  
feature_answer_question_content_similar + feature_length_origin,  
family=binomial(link='logit'),  
data = data_rec_answer)
```

## 3.8 内容质量排序方法

- 用logistic regression结果

Deviance Residuals:

| Min     | 1Q      | Median  | 3Q      | Max    |
|---------|---------|---------|---------|--------|
| -4.7989 | -0.4325 | -0.4065 | -0.3895 | 2.3652 |

Coefficients:

|   | Estimate          | Std. Error | z value  | Pr(> z )     |
|---|-------------------|------------|----------|--------------|
| (Intercept)                             | -2.665e+00        | 5.971e-03  | -446.373 | < 2e-16 ***  |
| feature_answer_content_similar          | <b>3.884e-02</b>  | 7.598e-04  | 51.116   | < 2e-16 ***  |
| feature_answerer_credit                 | <b>5.475e-02</b>  | 1.570e-04  | 348.666  | < 2e-16 ***  |
| feature_answer_question_content_similar | <b>-2.206e-03</b> | 2.857e-04  | -7.722   | 1.14e-14 *** |
| Feature_length_origin                   | <b>2.438e-03</b>  | 3.145e-05  | 77.525   | < 2e-16 ***  |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 3.8 内容质量排序方法

### • 用logistic regression结果

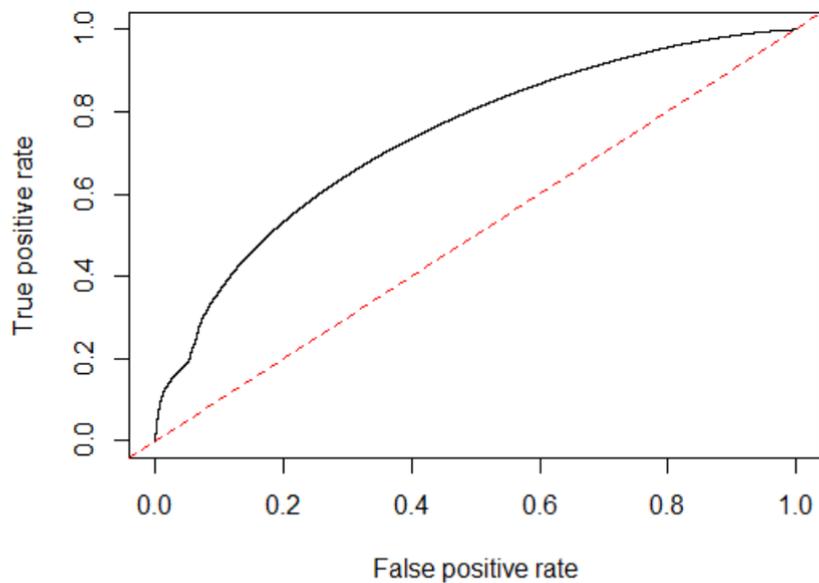
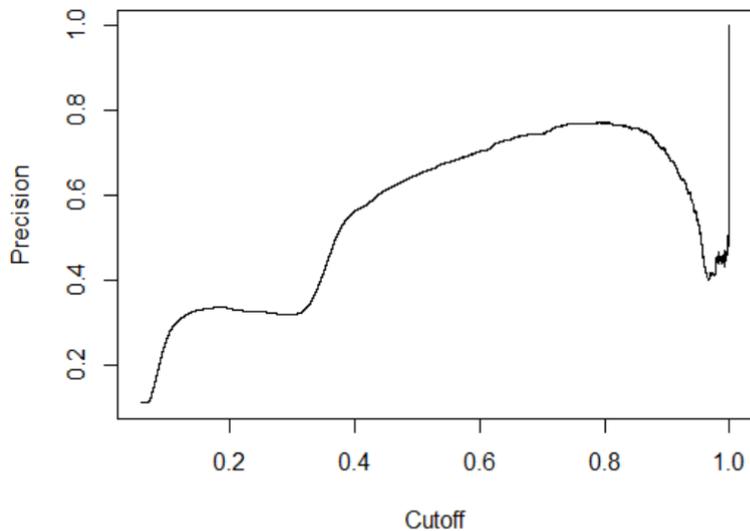
| Feature  | Coefficients          | 0.05水平显著与否    | 增加一个单位时发生比增加      | 标准差                | 增加一个标准差单位时发生比增加              |
|--|-----------------------|---------------|-------------------|--------------------|------------------------------|
| feature_answer_content_similar                     | <b>3.884e-02</b>      | 显著            | 3.96%             | 2.67316            | 10.94%                       |
| feature_answerer_credit                            | <b>5.475e-02</b>      | 显著            | 5.63%             | 9.79699            | 70.98%                       |
| <del>feature_answer_question_content_similar</del> | <del>-2.206e-03</del> | <del>显著</del> | <del>-0.22%</del> | <del>0.08072</del> | <del>0.02% (数据验证可忽略)</del>   |
| feature_length_origin                              | <b>2.438e-03</b>      | 显著            | 0.24%             | 50                 | 12.96%<br>(需进一步降权, antispan) |

$$\text{logit } p = \log o = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

$$\text{发生比: } o = \frac{p}{1-p} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \dots e^{\beta_k x_k}$$

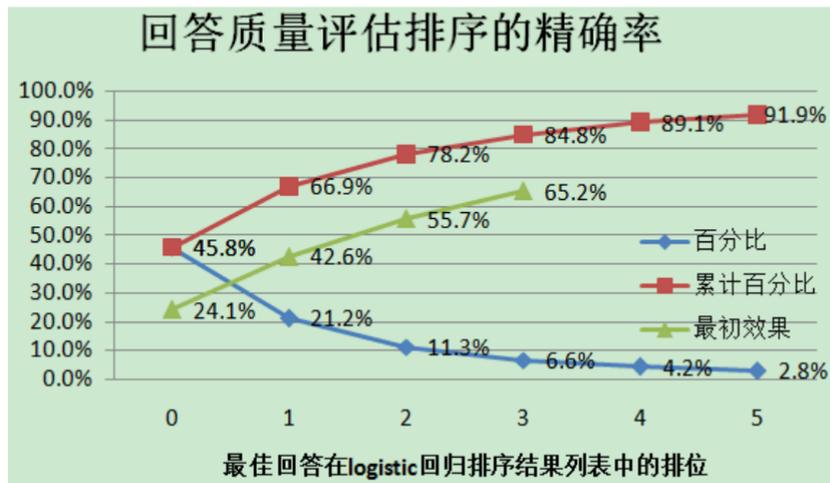
## 3.8 内容质量排序方法

- logistic regression结果



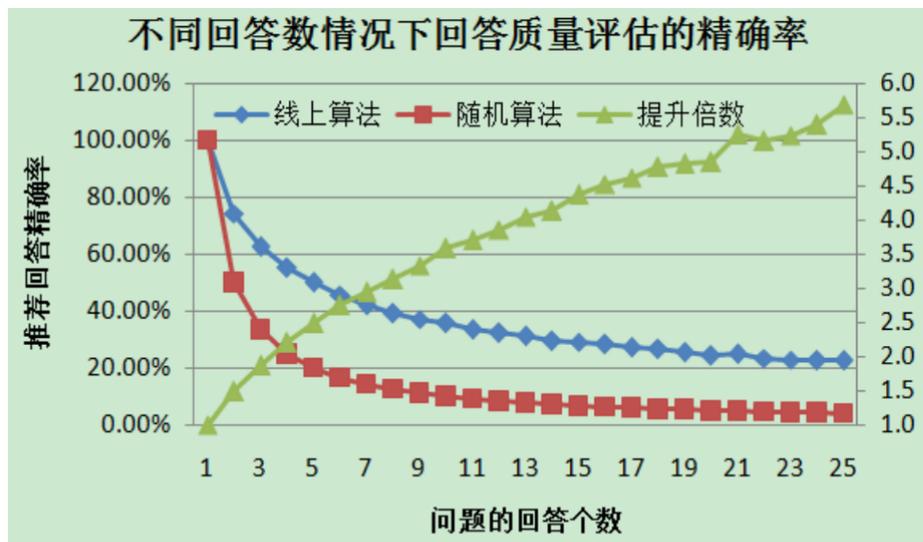
## 3.9 线上算法结果

- 线上产品环境结果：对最佳回答的预测覆盖情况
  - 根据logistic回归结果对每个问题的所有回答进行排序显示
  - 考察占25%的问题的最佳回答，在logistic回归&排序列表中的位置
  - 0，表示预测出来的“推荐回答（排在第1位）”就是提问者认为的最佳回答



## 3.9 线上算法结果

- 线上产品环境结果：不同回答数时，“推荐回答（排在第1位）”对最佳回答的匹配情况



## 4 宝宝树部分算法工作简介 – 欢迎加入宝宝树

- 宝宝年龄预测：细分到月份
- 反作弊反垃圾：内容层次、用户层次、行为层次
- 用户留存预测：及时挽回流失用户，提高用户活跃度
- 个性化推荐算法：由当前页面内容、用户兴趣、宝宝年龄做个性化
- 欢迎关注宝宝树，加入我们，和我们一起工作、学习、成长
  - 算法工程师、数据挖掘工程师；应届生有户口（2个名额）
  - 联系：[wanghao@babytree-inc.com](mailto:wanghao@babytree-inc.com) 或微博



耗资小王



LV 4



lvhl1983

互相关注

未分组

简介：关注推荐系统，数据挖掘，



守望的距离有多远

互相关注

未分组



龙\_2

互相关注

未分组



王瑞珩

互相关注

未分组

- 谢谢大家关注
- 吃饭去.....