基于用户评价改进产品结构——以华为手机为例

冯时（2019012342，fengs19@mails.tsinghua.edu.cn）

摘要：用户评价作为产品重要的市场反馈，对产品结构的设计有着非常高的参考价值。本文将以华为手机为例，简析一种利用用户评价改进产品结构的方法，并介绍该方法可能的改进。在第一部分中，本文将简要介绍华为手机的产品结构。在第二部分中，本文结合华为京东官方旗舰店的用户评论数据，分析华为手机的产品结构并给出了合理的改进建议。在第三部分中，给出了一个基于用户评论改进产品结构设计的一般性框架，并且展望了该框架未来可能的改进。最后一部分在第三部分的基础上进一步介绍了数据挖掘的主要方法与发展方向。

华为手机产品结构简介

截止至 2019 年 11 月 23 日星期六，华为京东官方自营旗舰店[①]共有 27 款产品，涵盖了从 429 元到 12999 元的价格区间，是一个具有代表性的手机产品线。

华为手机主要分为四个系列：Mate 系列、P 系列、Nova 系列和畅享系列，产品定位由高到低。其中 Mate 系列拥有 8 款机型，分别是 Mate30 RS、Mate30 Pro5G、Mate30 5G、Mate30 Pro、Mate30、Mate20 X(5G)、Mate20 Pro 和 Mate20。这里 Mate20 X(5G)、Mate20 和 Mate20 Pro 是 2018 年发布的产品，搭载海思麒麟 980 处理器；而 Mate30 RS、Mate30 Pro5G、Mate30 5G、Mate30 Pro 和 Mate30 则是 2019 年发布的机型，搭载海思麒麟 990 处理器，前三款机型支持 5G 技术。Mate 系列是华为的高端手机产品线，主要由旗舰机型组成，如表一所示。

表一：华为 Mate 系列配置价格表

型号	价格（元）	屏幕参数	内存大小	处理器	相机参数	电池大小	网络	重量
华为 HUAWEI Mate 30 5G	4999	6.62 英寸 /OLED/2340*1080 像素	8GB	麒麟 990	后置三摄 /主摄像素：4000 万像素	4200mAh	5G	196 克

[①] 参见 https://mall.jd.com/index-1000004259.html。

华为 HUAWEI Mate 30 Pro 5G	6899	6.53 英寸 /OLED/2400*1176 像素	8GB	麒麟 990	后置四摄 /主摄像素：4000 万像素	4500mAh	5G	约 198 克
华为 HUAWEI Mate 30	4299	6.62 英寸 /OLED/2340*1080 像素	8GB	麒麟 990	后置三摄 /主摄像素：4000 万像素	4200mAh	4G	196 克
华为 HUAWEI Mate 30 Pro	5799	6.53 英寸 /OLED/2400*1176 像素	8GB	麒麟 990	后置四摄 /主摄像素：4000 万像素	4500mAh	4G	约 198 克
华为 HUAWEI Mate 20 X (5G)	5199	7.2 英寸 /OLED/1080*2244 像素	8GB	麒麟 980	后置三摄 /主摄像素：4000 万像素	4200mAh	5G	约 233 克
华为 HUAWEI Mate 20 Pro	4499	6.39 英寸 /OLED/1440*3120 像素	8GB	麒麟 980	后置三摄 /主摄像素：4000 万像素	4200mAh	4G	约 189 克
华为 HUAWEI Mate 20	3099	6.53 英寸 /LCD/1080*2244 像素	6GB	麒麟 980	后置三摄 /主摄像素：1200 万像素	4000mAh	4G	约 188 克
华为 HUAWEI Mate 30 RS	12999	6.53 英寸 /OLED/2400*1176 像素	12GB	麒麟 990	后置四摄 /主摄像素：4000 万像素	4500mAh	5G	约 198 克

华为 P 系列拥有三款机型，分别是 P30 Pro、P30 和 P20 Pro，都不支持 5G 网络。前两款是 2019 年上半年发布的产品，搭载了海思麒麟 980 处理器，是华为的次旗舰机型。而 P20 Pro 则是 2018 年发布的产品，是上一代的次旗舰机型。和 Mate 系列相比，华为 P 系列手机屏幕较小，但是拥有更为强大的拍照功能，外观上也更加的年轻化。具体配置如表二所示。

表二：华为 P 系列配置价格表

型号	价格（元）	屏幕参数	内存大小	处理器	相机参数	电池大小	网络	重量

华为 HUAWEI P30 Pro	4488	6.47 英寸 /OLED/2340*1080 像素	8GB	麒麟 980	后置四摄 /主摄像素：4000 万像素	4200mAh	4G	192 克
华为 HUAWEI P30	3688	6.1 英寸 /OLED/2340*1080 像素	8GB	麒麟 980	后置三摄 /主摄像素：4000 万像素	3650mAh	4G	165 克
华为 HUAWEI P20 Pro	2200	6.1 英寸 /OLED/2240*1080 像素	6GB	麒麟 970	后置三摄 /主摄像素：4000 万像素	4000mAh	4G	约 180 克

华为 Nova 系列是华为手机相对比较年轻一个产品系列。在定位上，华为 Nova 系列比 Mate 系列和 P 系列低一个档次，做工不如 Mate 系列和 P 系列精致，但外观更为时尚，主要面向年轻客户。Nova 系列机型比较有针对性，如 Nova 5i Pro 在屏幕以及处理器上有所欠缺，但以较低的价格提供了非常强大的拍照体验。Nova 系列拥有 7 款机型，分别是 Nova5 Pro、Nova 5z、Nova 5i Pro、Nova 5i、Nova 5、Nova 4 和 Nova 4e。其中 Nova5 Pro 采用了和旗舰机型一样的海思麒麟 980 处理器，而其他机型则使用了较为廉价的解决方案，如海思麒麟 810 等，详情可见表三。

表三：华为 Nova 系列配置价格表

型号	价格（元）	屏幕参数	内存大小	处理器	相机参数	电池大小	网络	重量
华为 HUAWEI nova 5z	1399	6.26 英寸 /LCD/2340*1080 像素	6GB	麒麟 810	后置三摄 /主摄像素：4800 万像素	4000mAh	4G	约 178 克
华为 HUAWEI nova 5 Pro	2599	6.39 英寸 /OLED/2340*1080 像素	8GB	麒麟 980	后置四摄 /主摄像素：4800 万像素	3500mAh	4G	约 171 克
华为 HUAWEI nova 4	1599	6.4 英寸 /LCD/2310*1080 像素	6GB	麒麟 970	后置三摄 /主摄像素：2000 万像素	3750mAh	4G	172 克
华为 HUAWEI nova 4e	1249	6.15 英寸 /LCD/2312*1080 像素	6GB	麒麟 710	后置三摄 /主摄像	3340mAh	4G	约 159 克

					素：2400 万像素			
华为 HUAWEI nova 5i	1999	6.4 英寸 /LCD/2310*1080 像素	8GB	麒麟 710	后置四摄 /主摄像素：2400 万像素	4000mAh	4G	178 克
华为 HUAWEI nova 5i Pro	1799	6.26 英寸 /LCD/2340*1080 像素	6GB	麒麟 810	后置四摄 /主摄像素：4800 万像素	4000mAh	4G	约 178 克
华为 HUAWEI nova 5	2599	6.39 英寸 /OLED/2340*1080 像素	8GB	麒麟 810	后置四摄 /主摄像素：4800 万像素	3500mAh	4G	约 171 克

华为畅享系列是华为手机面向低端客户的产品线。主要由千元机和百元机组成，其中包括 9 款机型，分别是畅享 10S、畅享 10、畅享 10 Plus、畅享 MAX、畅享 9S、畅享 9 PLUS、畅享 9、畅享 9e 和畅享 8e 青春版。这些机型主要搭载华为海思和联发科（Mediatek）的低端处理器，定位较低，受到了老年用户和学生群体的喜爱，在配置上基本可以满足日常使用的要求。具体的配置可参见表四。

表四：华为畅享系列配置价格表

型号	价格（元）	屏幕参数	内存大小	处理器	相机参数	电池大小	网络	重量
华为 HUAWEI 畅享 10S	1549	6.3 英寸 /OLED/2400*1080 像素	6GB	麒麟 710F	后置三摄/ 主摄像素：4800 万像素	4000mAh	4G	约 163 克
华为 HUAWEI 畅享 9S	1099	6.21 英寸 /LCD/2340*1080 像素	6GB	麒麟 710F	后置三摄/ 主摄像素：2400 万像素	3400mAh	4G	约 160 克
华为 HUAWEI 畅享 10 Plus	1449	6.59 英寸 /LCD/2340*1080 像素	4GB	麒麟 710F	后置三摄/ 主摄像素：4800 万像素	4000mAh	4G	约 196.8 克
华为 HUAWEI 畅享 8e 青春版	429	5.45 英寸 /LCD/1440*720 像素	2GB	Mediatek MT6739	后置单摄/ 主摄像素：1300 万像素	3020mAh	4G	约 142 克

华为 HUAWEI 畅享 9 Plus	898	6.5 英寸 /LCD/1080*2340 像素	4GB	麒麟 710F	后置双摄/主摄像素：1300 万像素	4000mAh	4G	约 173 克
华为 HUAWEI 畅享 MAX	899	7.12 英寸 /LCD/1080*2244 像素	4GB	高通骁龙 660	后置双摄/主摄像素：1600 万像素	5000mAh	4G	约 210 克
华为 HUAWEI 畅享 10	1349	6.39 英寸 /LCD/1560*720 像素	6GB	麒麟 710F	后置双摄/主摄像素：4800 万像素	4000mAh	4G	约 176 克
华为 HUAWEI 畅享 9	999	6.26 英寸 /LCD/1520*720 像素	4GB	高通骁龙 450	后置双摄/主摄像素：1300 万像素	4000mAh	4G	约 168 克
华为 HUAWEI 畅享 9e	849	6.088 英寸 /LCD/1560*720 像素	3GB	Mediatek MT6765	后置单摄/主摄像素：1300 万像素	3020mAh	4G	约 150 克

总而言之，华为手机的四个系列涵盖了低端手机、中端手机、大屏高端手机和小屏高端手机，产品结构较为清晰，面向用户的分类也比较清晰。应该说华为手机的产品结构是比较合理的。但事实上，根据京东华为官方旗舰店的用户评价来看，其产品结构仍然存在一些问题，下面我们将对此进行分析。

基于用户评价的改进

对于一款手机，我们应当从多个维度来进行评价，包括处理器性能和屏幕素质等关键参数。在用户评论中少有提及的部分大多是产品的不足之处，这是因为对于购买产品的用户而言，当然不会为了一款产品的缺陷而购买。而被提到多次的则是产品的优势或是与众不同的特征，正是这些特征或优点吸引了用户的关注，从而使得这些特征会在用户评论中被重点提到。基于这样的分析，我们可以得到用户眼中产品的不足之处和优势，再结合前文已有的产品结构，帮助我们分析产品结构的不足之处。

图一：华为 Mate 系列评论词云

图二：华为 P 系列用户评论词云

首先通过网络爬虫技术（程序源代码见附录），我们得到了华为京东官方旗舰店的用户

评价关键词词云，我们将在文章第三部分详细描述产品特征挖掘的技术。其中华为 Mate 系列如图一所示，华为 P 系列如图二所示，华为 Nova 系列如图三所示，而华为畅享系列则如图四所示。可以看到，在高端手机方面，无论是 P 系列还是 Mate 系列，购买华为手机的用户最看重的是 5G、屏幕和拍照，却少有人提到处理器性能、游戏和系统流畅度等关键词。同时品牌力的体现也有些不足，在评论中大量出现了“苹果”这样的关键词，这说明华为高端手机常被用户与苹果手机进行对比，而相对应的，在苹果手机的用户评价中却并未多次出现华为[②]。在表一与表二中，华为的高端手机产品都采用了麒麟 990 或是麒麟 980 处理器，在性能上不及苹果最新款手机搭载的 A13 处理器[③]，而在屏幕分辨率上也不及苹果 iphone 11 Pro Max。可以看出，在高端市场，华为仍然缺少一款有强大处理器性能的产品，同时华为手机在系统体验上相比苹果 ios 操作系统也有所欠缺。理想的解决方案是推出一款在处理器性能上能与竞品抗衡的产品，提升屏幕素质，并且保持其他配置上的优势，这样不仅仅完善了产品线，给追求极致性能的用户一个选择，对品牌影响力的提升也有着至关重要的作用。

图三：华为 Nova 系列用户评论词云

[②] 以同样的方法爬取苹果京东官方自营旗舰店用户评论得到的关键词词云中未出现“华为”、“Mate”等关键词。

[③] 参见汐元.麒麟 990 vs. 苹果 A13 全面解析：谁是 2019 最强芯？[EB/OL]. https://www.ithome.com/0/444/644.htm, 2019-09-11.

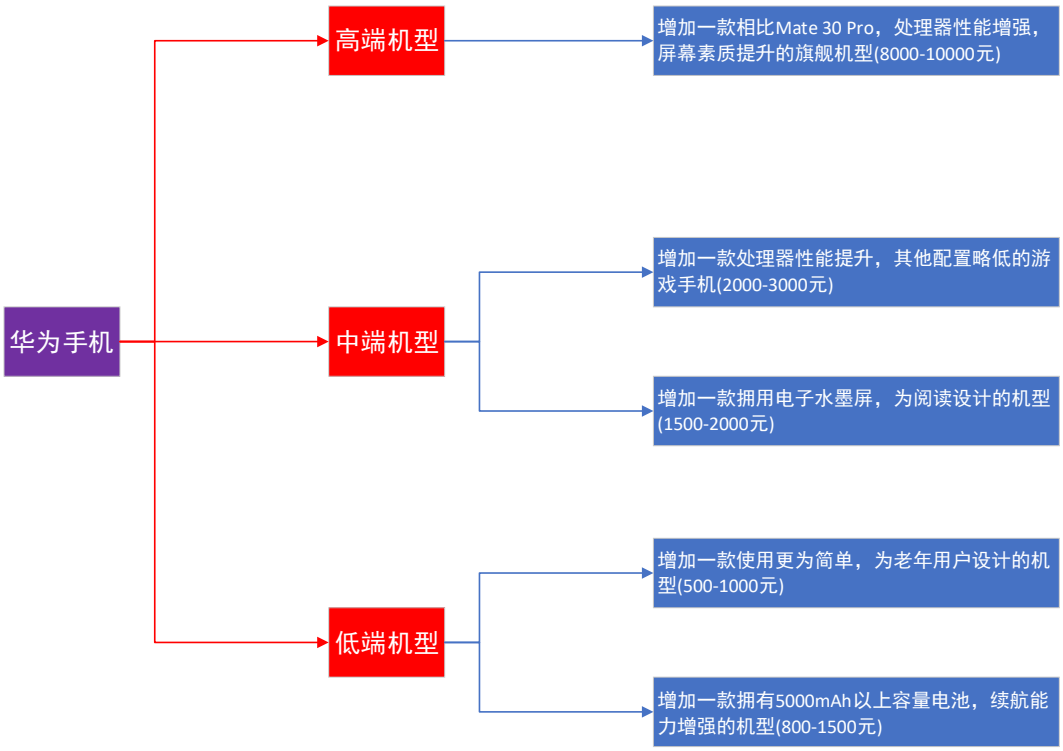
华为 Nova 系列面向的用户群体主要是年轻用户。在词云中，大量出现了“拍照”、“像素”这样的关键词，这说明用户对这个系列手机最为满意的就是拍照性能。从表三的配置表中同样可以看出，Nova 系列的摄像头配置非常接近于旗舰机型。然而，能打动年轻用户的并不仅仅只有拍照，Nova 系列大量机型重合度较高，如 Nova 5 Pro 和华为 P30 Pro 参数上极为接近。通过市场上其他产品的对比，一款注重游戏性能和一款注重阅读体验的手机或许能给 Nova 系列单一的产品特点带来改变。这两款手机可以分别与小米黑鲨游戏手机以及海信 A6 对标竞争。

图四：华为畅享系列用户评论词云

华为畅享系列的词云体现了华为低端产品特点上的不足。低端产品的用户大多是学生老人，以及希望用第二台手机作为备用机的用户。而从配置表表四中可以看出，华为并没有对产品配置进行有针对性的调整，词云中被提到最多的不是“续航”、“易于操作”等关键词，反而是“外形”与“外观”。事实上，华为畅享系列作为低端产品线，一些有针对性的机型是必不可少的。如一款与小米多亲 AI 手机竞争的低端手机可以大大提升老年用户的体验，而一款拥有大容量电池的机型则可以成为备用机用户的首选之一。低端手机并不代表着所有

配置的缩水，低端手机不是旗舰手机的全面缩水版，鲜明的特点对于每一款产品而言都是非常重要的。

总而言之，华为手机的产品结构仍然存在较多不足，通过上面的分析，下面我们给出根据文章第一部分的配置表以及词云分析得出的理想产品结构，如图五所示。



图五：华为手机理想的产品结构

基于用户评价改进产品结构的基本流程

用户拥有对一款产品成功与否的最终发言权，而评论则是最真实、最完整的用户反馈。用户对一个产品线产品的评论与产品结构的合理性密切相关。“ 产品结构是指社会产品各个组成部分所占的比重和相互关系的总和。它可以反映社会生产的性质和发展水平，资源的利用状况，以及满足社会需要的程度。”[④]产品结构应当时刻根据社会的需求进行调整，作为一家企业，为了让自己的利益最大化，必然需要满足用户的需求，而产品结构则是迎合用户需求非常重要的一环。而为了明白社会的需求，企业必然需要得到用户的反馈，这就是为什么产品结构和用户的评论是密不可分的。只有设计出了市场以及用户认可的产品结构，才能带来利益的最大化。

[④] 徐鑫.产品结构优化建模及应用研究[D].同济大学,2008.

事实上，通过产品结构优化取得巨大进步的企业比比皆是，最好的例子就是国内传统老牌药企丽珠集团。近年来，中药注射剂饱受患者质疑，其安全性和药效并没有得到广泛地认可。而“丽珠集团最大单品是中药注射剂参芪扶正，销售额占总营收接近 20%，该单品的销售受阻，2018 年销售额剩下 10 亿元，占比仅为 11.3%，这导致中药制剂同比快速下降 25%至 15.32 亿元。”[⑤]正是得到了这样的市场反馈，迫使丽珠集团对自己的产品结构做出了改变，将产品重心转移到亮丙瑞林和艾普拉唑上。除此之外，丽珠集团更是布局精准医疗领域，进一步调整产品结构的重心。“2018 年亮丙瑞林、艾普拉唑两个产品的单品销售规模已经与参芪扶正旗鼓相当。2019 年上半年则分别实现销售收入 4.56 亿元和 4.57 亿，同比增长 62.52%和 27.3%，继续保持高增长之势。相反，上半年参芪扶正下降至 4.5 亿元的规模，全年或跌出 10 亿的规模，而鼠神经生长因子的规模上半年则为 2.25 亿元，全年预计小幅下滑，两者上半年的营收占比已经下降至 15%以下，未来还会持续减小。公司大单品结构转换将带来新的成长动能。”[⑥]由此可见，产品结构对一个公司的长期利益、短期利益都是有着巨大影响的，产品结构的优化，对于每个企业来讲都是必须要考虑的问题。

在这个数据爆炸的时代，很多企业家并没有关注用户评论这样的细微却同样重要的数据，过多地注重营业收入等整体性数据，导致不能很好地了解用户的真实需求。事实上，改善产品结构，用户反馈应当是一个很重要的切入点。那么，如何通过用户评论改进产品结构呢？上面已经给出了一个对华为手机产品线进行改进的简单例子，接下来我们结合上面的例子，给出一个更为精确且更为系统的利用用户评论改进产品结构的基本流程。

第一步是选择合适的用户反馈平台。有大量无关信息的反馈平台如论坛、社区等都不能作为数据集，而主观性较强的平台也不能作为良好的数据集。在我们的例子中，选择了华为京东官方旗舰店的用户评论作为数据集，具有较高的客观性，评论也大多与产品特性有关。事实上，进一步，我们可以利用贝叶斯分类可以对用户评论的主观性进行评估，从而得到更为精确的数据集。比如叶强等人提出的中文的主观性自动判别方法[⑦]，接近英文同类研究的成果，可以应用到数据集的选择问题上。

第二步是获取数据集。在我们的例子中，利用了网络爬虫技术获取数据，除了最基本的基于“requests”包的网络爬虫程序以外，“Scrapy”框架等网络爬虫框架也是可以考虑的。孙立伟等人给出了有关网络爬虫技术的研究综述，具体可以参见《网络爬虫技术的研究》一

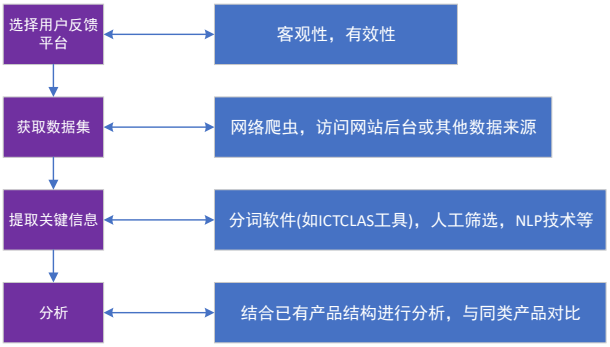
[⑤] 周少鹏.丽珠集团：逐渐走出阴影的医药龙头[J].股市动态分析,2019(34):27-28.
[⑥] 周少鹏.丽珠集团：逐渐走出阴影的医药龙头[J].股市动态分析,2019(34):27-28.
[⑦] 参见叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究[J].信息系统学报,2007(01):79-91.

文[8]。事实上，对于企业而言，直接从电商平台使用的数据库获取评论数据也是可行的解决方案。

第三步是在庞大的数据集中提取出有价值的关键信息，这是整个框架最为关键的部分。这个步骤将看似杂乱无章的数据集变为可以加以利用的信息，在本质上提升了数据的价值。在上面的例子中，“jieba”中文分词 python 组件发挥了巨大的作用。这是一个用于文本分析的 python 组件，使用极为简单，可以将数据集切分为众多关键词，从而为进一步的数据统计以及语义分析带来了方便。如果需要更为精确的结果，人工标注将会是首选。叶强等人给出了评论产品特征挖掘方法的七个步骤，主要基于频繁特征项的处理方法[9]。而高威等人则利用用户评论挖掘了 2017 年市面上热门机型的特征，有很高的参考价值[10]。其实第三步，就是数据挖掘技术的应用。我们将在文章的第四部分详细讲述这个部分。

最后一步基于得到的关键信息，结合已有的产品结构进行分析。在上面的例子中，我们通过配置表与用户关注的产品优势与劣势进行分析，给出了对华为手机产品结构的改进的建议。而在实际应用中，更是可以对其他同类产品进行同样的分析，与自身产品进行对比，分析用户在使用其他厂商产品时遇到的问题以及看重的优势，从而改善自身的产品结构。更进一步的，还可以利用一些情感分析方法，更深入的了解用户的需求，从而发现并解决用户的痛点。从某种意义上讲，我们的框架其实是一个商业智能系统的实例。在商业智能的理论中，这一步是决策者应当完成的工作。“商业智能系统从企业运作的日常数据中开发出结论性的、基于事实的和具有可实施性的信息，使企业能够更快更容易的做出更好的商业决策。使企业管理者和决策者以一种更清晰的角度看待业务数据，提高企业运转效率、增加利润并建立良好的客户关系，使企业以最短的时间发现商业机会捕捉商业机遇。”[11]有了这样的商业智能系统，就可以让企业的管理者可以更准确的了解到企业的真实状况，得到更多关于企业运营情况的细节，从而做出正确的抉择，比如产品结构的优化策略。

[8] 参见孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.
[9] 李实,叶强,李一军,Rob Law.中文网络客户评论的产品特征挖掘方法研究[J].管理科学学报,2009,12(02):142-152.
[10] 参见高威,傅湘玲.基于用户评论的手机特征挖掘应用研究[J]. 计算机科学与应用, 2017, 07(08): 738-746.
[11] 余长慧,潘和平.商业智能及其核心技术[J].计算机应用研究,2002(09):14-16+26.

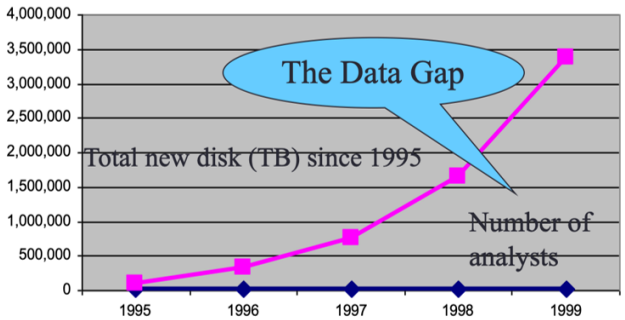


图六：基于用户评价改进产品结构的流程图

完整的流程图如图六所示。总共分为四个步骤，每一个步骤都有较为简便的方法和成本较高，但是更为精确的方法。基于华为手机的例子只是抛砖引玉，更多的是为产品结构优化这一课题带来新的灵感与想法。希望在今后的产品结构设计中，企业能更多的关注用户评论带来的大量信息，而不是仅仅专注于销量等数字。

数据挖掘的基本方法

在前文中提到，基于用户评论的产品结构改进模型中，第三个步骤也就是数据挖掘是非常关键的一环，第三步把庞大的数据库变成了直观的图表或是大多数人容易理解的因果关系，从而变成决策者可以加以利用的工具。从图七中我们可以看到，相比数据量的增长速度，数据挖掘技术的发展几乎可以忽略不计，所以数据挖掘技术可以说是前文模型中改进空间最大的一个环节。

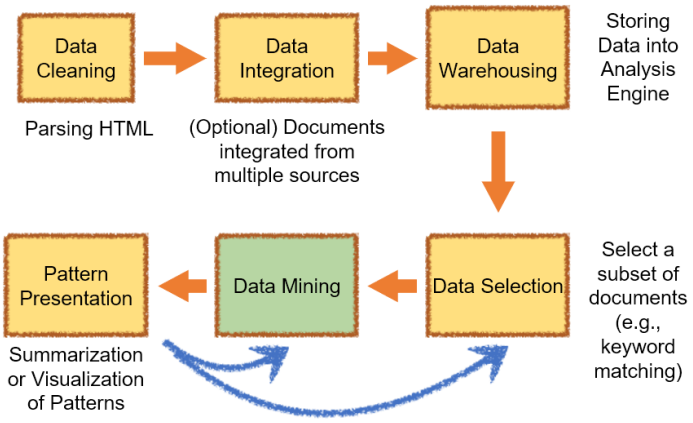


图七：数据挖掘技术的发展与数据量增长的速度对比[12]

[12] Data mining for scientific and engineering applications[M]. Springer Science & Business Media, 2013.

数据集可以来自数据库，也可以是图表或者信息网络，但是一定需要具备某些特征。在前文中数据集是用户评论，每条用户评论都具备很强的独立性并且对于不同的产品而言，用户评论具备不同的特征。

对于一个数据集，数据挖掘技术提供了一些方法用来分析和处理数据。包括特征化、模式搜索、分类器、聚类分析、孤立点分析、序列分析和网络分析。其中特征分析是指利用可视化等方法总结并比较数据之间的不同特征，这也是前文的例子中进行的工作的一部分（事实上，这就是附录的代码中“wordcloud”包的作用）。模式搜索是指搜索数据之间的关联性，如区域物流和经济发展间的关联[13]。分类器是对数据的分类和标注，在前文的例子中同样有所涉及（代码中的“jieba”包就利用了这一技术）。聚类分析与分类器有些类似，它采用某种策略将数据分为一些聚类，同一聚类中的数据关联性较大而不同聚类中数据关联性较小。孤立点分析顾名思义是指对远离拟合函数的数据的分析，这些数据往往是值得关注的，孤立点分析可以用于发现被审计数据中的疑点[14]。序列分析基于具备时序的数据集，用来预测未来的趋势，在金融领域有着广泛的应用，如金雪军等人对我国区域金融成长的分析[15]。最后的网络分析相对比较复杂，综合了分类器、聚类分析和孤立点分析，主要用于对社交网络的分析以及 Web 分析。这些方法构成了数据挖掘技术的主要组成部分。下面是一个基于 Web 的数据挖掘框架：



图八： 基于 Web 的数据挖掘框架

[13] 参见戢晓峰,张雪,陈方,李杰梅.基于多源数据的区域物流与经济发展关联特性分析——以云南省为例[J].经济地理,2016,36(01):39-45.

[14] 周喜,曾丽.孤立点数据挖掘技术在审计信息化中的应用研究[J].南华大学学报(社会科学版),2011,12(05):55-57.

[15] 金雪军,田霖.我国区域金融成长差异的态势:1978-2003 年[J].经济理论与经济管理,2004(08):24-30.

数据挖掘技术的应用有很多，其中最引人注目的就是在商业智能领域的应用，利用数据挖掘技术前文构造了基于用户评论优化产品结构的框架。实际上，在未来的时间里，如果数据挖掘技术能进一步提升与用户的交互体验以及对不同种类数据的兼容性，我相信用户评论能为企业的发展计划制定带来更多关键信息，从而进一步促进企业的智能化运营，同时也让用户能体验到真正符合他们的真实需求的产品。

附录

京东评论 python 网络爬虫：

```
1. import os
2. import time
3. import json
4. import random
5. import jieba
6. import requests
7. import numpy as np
8. from PIL import Image
9. import matplotlib.pyplot as plt
10. from wordcloud import WordCloud
11.
12. # 词云形状图片
13. WC_MASK_IMG = 'wordcloudtem.png'
14. # 评论数据保存文件
15. COMMENT_FILE_PATH = 'jd_comment.txt'
16. # 词云字体
17. WC_FONT_PATH = 'msyh.ttc'
18.
19. def spider_comment(page=0):
20.     """
21.     爬取京东指定页的评价数据
22.     :param page: 爬取第几，默认值为 0
23.     """
24.     # 100009177388 可以替换为任意商品的商品号
25.     url = 'https://sclub.jd.com/comment/productPageComments.action?callback=
    fetchJSON_comment98vv4646&productId=100009177388' \
26.           '&score=0&sortType=5&page=%s&pageSize=10&isShadowSku=0&fold=1' % p
    age
27.     kv = {'user-
    agent': 'Mozilla/5.0', 'Referer': 'https://item.jd.com/100009177388.html'}
28.     try:
```

```
        r = requests.get(url, headers=kv)
        r.raise_for_status()
    except:
        print('爬取失败')
    # 截取 json 数据字符串
    r_json_str = r.text[26:-2]
    # 字符串转 json 对象
    r_json_obj = json.loads(r_json_str)
    # 获取评价列表数据
    r_json_comments = r_json_obj['comments']
    # 遍历评论对象列表
    for r_json_comment in r_json_comments:
        # 以追加模式换行写入每条评价
        with open(COMMENT_FILE_PATH, 'a+') as file:
            file.write(r_json_comment['content'] + '\n')
        # 打印评论对象中的评论内容
        print(r_json_comment['content'])

def batch_spider_comment():
    """
    批量爬取某东评价
    """
    # 写入数据前先清空之前的数据
    if os.path.exists(COMMENT_FILE_PATH):
        os.remove(COMMENT_FILE_PATH)
    for i in range(100):
        spider_comment(i)
        # 模拟用户浏览，设置一个爬虫间隔，防止 ip 被封
        time.sleep(random.random() * 5)

def cut_word():
    """
    对数据分词
    :return: 分词后的数据
    """
    with open(COMMENT_FILE_PATH) as file:
        comment_txt = file.read()
        wordlist = jieba.cut(comment_txt, cut_all=True)
        wl = " ".join(wordlist)
        print(wl)
        return wl

def create_word_cloud():
    """
```

```
    生成词云
    :return:
    """
    # 设置词云形状图片
    wc_mask = np.array(Image.open(WC_MASK_IMG))
    # 设置词云的一些配置，如：字体，背景色，词云形状，大小
    wc = WordCloud(background_color="white", max_words=2000, mask=wc_mask, scale=4, max_font_size=50, random_state=42, font_path=WC_FONT_PATH)
    # 生成词云
    wc.generate(cut_word())
    # 在只设置 mask 的情况下,你将会得到一个拥有图片形状的词云
    plt.imshow(wc, interpolation="bilinear")
    plt.axis("off")
    plt.figure()
    plt.show()

if __name__ == '__main__':
    # 爬取数据
    batch_spider_comment()
    # 生成词云
    create_word_cloud()
```

参考文献

[1] 汐元.麒麟990 vs 苹果A13全面解析：谁是2019最强芯？[EB/OL].https://www.ithome.com/0/444/644.htm, 2019-09-11.

[2]徐鑫.产品结构优化建模及应用研究[D].同济大学,2008.

[3]周少鹏.丽珠集团：逐渐走出阴影的医药龙头[J].股市动态分析,2019(34):27-28.

[4]叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究[J].信息系统学报,2007(01):79-91.

[5]孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.

[6]李实,叶强,李一军,Rob Law.中文网络客户评论的产品特征挖掘方法研究[J].管理科学学报,2009,12(02):142-152.

[7]高威,傅湘玲.基于用户评论的手机特征挖掘应用研究[J]. 计算机科学与应用, 2017, 07(08): 738-746.

[8]余长慧,潘和平.商业智能及其核心技术[J].计算机应用研究,2002(09):14-16+26.

[9]吴忠,丁绪武.大数据时代下的管理模式创新[J].企业管理,2013(10):35-37.

[10]Data mining for scientific and engineering applications[M]. Springer Science & Business Media, 2013.

[11]Goebel M, Gruenwald L. A survey of data mining and knowledge discovery software tools[J]. ACM SIGKDD explorations newsletter, 1999, 1(1): 20-33.

[12]Brookshear J G. Computer science: an overview[M]. Addison-Wesley Publishing Company, 2008.

[13]戢晓峰,张雪,陈方,李杰梅.基于多源数据的区域物流与经济发展关联特性分析——以云南省为例[J].经济地理,2016,36(01):39-45.

[14]周喜,曾丽.孤立点数据挖掘技术在审计信息化中的应用研究[J].南华大学学报(社会科学版),2011,12(05):55-57.

[15]金雪军,田霖.我国区域金融成长差异的态势:1978-2003年[J].经济理论与经济管理,2004(08):24-30.