摘要：移动应用程序评论是直接来自应用程序用户的有价值的知识库。这样的想法涵盖了各种主题，在本文中我们展示了其中23.3％代表功能请求，即用户为应用程序建议新功能或者为应用程序的已有功能重新设计重新设计偏好的评论。应用程序开发人员在尝试使用此类反馈时面临的挑战之一是大量的可用评论。这使得很难确定各个评论的具体主题和反复出现的趋势。通过这项工作，我们的目标是通过设计MARA（移动应用程序审查分析器）来支持这样的过程，这是一个自动检索移动应用程序功能请求的原型  
在线评论。原型的设计是：a）通过调查用户通过评论表达特征请求的方式，b）围绕一组预定义的语言规则制定，以及c）在大量的在线评论样本上进行评估。利用潜在狄利克雷分配进一步分析评估结果，以确定特征请求中的共同主题，并在本文中报告分析结果。

关键词：在线评论，移动应用程序，功能请求。

引言：

在线评论是促进客户购买决策和产品评估的工具[1，2]。 特别是移动应用程序评论，是直接来自用户的有价值的想法和意见库，并使其他用户和应用程序开发人员受益。 一方面，用户可能会根据应用程序的其他用户的体验，在决定是否购买/安装应用程序之前，先查看应用程序的可用性和潜在错误。 另一方面，对于开发者来说，评论是可以包括改进想法或功能请求的评估。 我们认为功能请求是评论，通过这些评论用户可以为应用程序推荐新功能，或者表示对已有功能进行重新设计的偏好设置。 以前的工作主要集中在从应用程序商店描述文本中提取功能请求信息[4]，以及通过继承和类重用来理解Android移动应用程序市场中的软件重用[5]。

通过这项工作，我们调查了应用程序用户使用在线评论来表达功能请求的程度，以及在网上可用的评论中自动执行此类请求的挖掘过程的可能方式。 有几个挑战涉及。 首先，与一个应用程序相关的评论数量往往超过了人类阅读全部内容的能力，以便识别经常出现的问题和趋势。 仅Facebook应用就有近600万条评论。 此外，Google应用商店中最受欢迎的非免费应用每天平均得到45.5条评论。 其次，评论有自己的风格; 他们通常很短，非结构化，很少遵守语法和标点规则。 此外，用户以非常规的方式表达意见，经常使用非常规的语法或讽刺来购买应用程序后发泄自己的感受。 尽管人类通常可以解释这样的评论，但是自动化这种解释是具有挑战性的[6]。

我们首先考虑了分析的评论样本，我们的目标是确定多少用户的反馈包括功能请求，以及如何通过评论来表达这些功能请求。 这样的发现通知了能够识别和检索来自移动应用的在线评论的特征请求的原型系统的设计。 我们使用大量的评论评估了原型，我们使用潜在的Dirichlet分配[7]来确定作为评估结果获得的特征请求中的共同主题。 本文的结构如下：第2节报告在线评论分析，第3节描述了为自动检索来自在线评论的移动应用特征请求而设计的原型，第4节给出了对在 原型的评估。 最后，本文结束了对原型未来迭代的结论和想法。

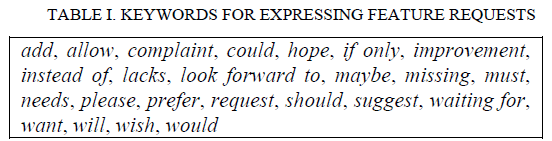
在我们试图了解功能请求包含多少反馈的情况下，我们从Google应用商店中选择了前6个最受欢迎的类别，对于每个类别，我们随机生成了一个数字nk的应用标识符，其中nk是所采样本的大小从该类别中的应用程序数量来看，与该类别中的应用程序总数成正比。所选类别和每个类别的应用程序数量为：个性化（54个应用程序），工具（14个应用程序），书籍和参考（45个应用程序），教育（27个应用程序），生产力（19个应用程序）和健康和健身10个应用程序）。对于每个随机选择的应用程序，我们自动提取并存储用户提供的评论。对于每个评论，我们都会自动收集发布日期，用户给出的评分，关联的设备以及应用的版本，以及实际评论的标题和文字。在随机选择的169个应用程序中，有8个应用程序没有分配给他们的评论，这给我们留下了161个已查看的应用程序，总共有3279个评论。所考虑的161个应用程序的平均评分是4.27（1到5分），而每个应用程序的平均价格是1.92英镑。

我们通过评论的样本，手动提取表达功能请求的评论片段（例如，“你需要添加一个退出按钮”，“希望它有一个数学语言包”，“希望我们可以有图标旁边的计数器”）。我们发现，评论被广泛用于向开发者提供有关可以改进或添加到应用程序的功能的详细信息。在所分析的反馈的23.3％中观察到这种行为：在3279个被认为表达的特征请求中，有763个评论。评分较高的应用获得的功能要求要比评分较低的应用多得多评级低于3的应用程序中只有37.5％与功能请求关联，而超过3的应用程序中有70％以上的应用程序与至少一项功能请求相关联。在价格方面，用户对更昂贵的应用程序和更便宜的应用程序表达的功能要求数量之间没有显着差异。几乎有四分之一的评论提供功能请求，我们认为支持自动检索来自在线评论的功能请求是必要的。

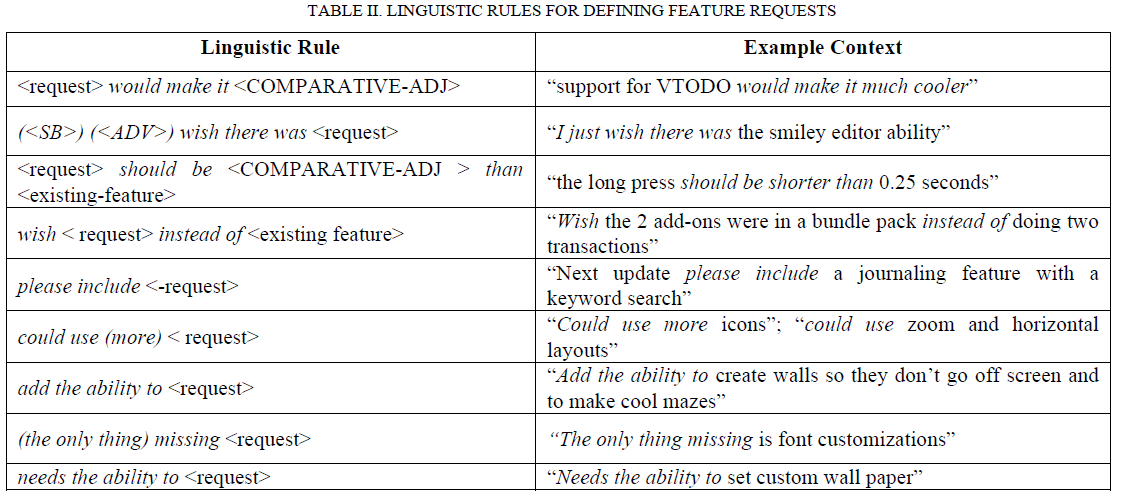
MARA是开发的原型，用于从移动应用程序的在线评论中检索功能请求。该系统被设计为（图1）：1）检索可用于应用的所有评论（评论检索），2）挖掘用于识别表达特征请求的句子或片段的评论的内容（特征请求挖掘），3 ）总结这样的内容（特征请求摘要），并且4）以用户友好的方式（特征请求可视化）呈现它。在检索检索阶段，网络爬虫将提取构成给定应用（原始评论）评论的页面源，并分析其内容。对这项工作感兴趣的是评论的实际内容，但是与每个评论相关的元数据也被收集用于进一步的分析。这样的元数据包括发布评论的日期，用户给出的评价，与评论相关联的设备，用户使用的应用的版本以及用户与评论相关联的标题。进一步的分析可能包括按时间顺序评估应用程序功能请求的演变。存储评论的内容和评论的元数据，将内容标准化以减少最终结果中的噪音。另外，评论内容被分成句子。为此，我们使用支持句子分割的自然语言处理工具LangPipe [8]。

该功能请求挖掘阶段将拆分评论的内容作为输入，并针对用户表示的功能请求挖掘地雷。特征请求挖掘算法使用一组定义的语言规则来支持引用这些请求的句子的识别。例如，特征请求更倾向于用“添加一个退出按钮将会很棒”这样的句子来表达，这意味着“添加<请求>将会是<POSITIVE-ADJECTIVE>”的语言规则。在特征请求汇总阶段，系统根据一组预定义的规则汇总提取的特征请求。这些规则的目标是根据提取的请求的频率和长度来对提取的请求进行排序;在总结中首先显示更频繁和更长的特征请求。由于用户提供了大量的反馈，因此这是必要的。最后，在功能请求可视化阶段，汇总结果显示给用户。通过本文，我们将解决特征请求挖掘和评估的过程。

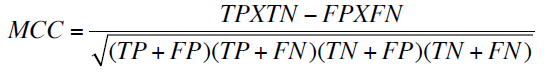
1. 功能请求挖掘  
   基于对第二部分描述的随机抽样中人工确定的特征请求的分析，我们探讨了定义表达请求的语言的可能性。 首先，我们遍历所有标记为特征请求的句子，并将每个句子与表示句子的关键字相关联。 我们选择了所有与3个以上句子相关的关键字，以避免单词与特征请求的意外关联。 我们确定了24个这样的关键字（表1），其中包含80％的句子。



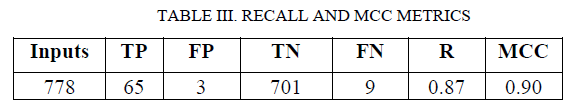
在确定了这些关键字之后，我们过滤了所有包含这些关键字中至少一个的句子，并且全部通过它们来定义这些关键字指向实际功能请求的上下文（即句子的片段）。这样的上下文的例子是：“退出按钮将是太棒了”，“添加更多的图标将是伟大的”，“提示和数学支持也将是很好的”。然后，我们将上下文抽象为语言规则，例如：“（添加）<请求>将（<ADV>）为<正面 - ADJECTIVE>”。 “增加”这个词是可选的，而ADV和POSITIVEADJECTIVE可以分别替换为任何副词或积极的形容词。使用所有已确定的上下文，我们定义了一组237条语言规则，其中一些例子如表2所示。识别特征请求转换成识别匹配至少一个语言规则的上下文。这项工作的局限性仅在于考虑用英文撰写的评论。然而，通过定义语言特定的语言规则，该方法可以应用于其他语言。另外，这个作品并没有专门讨论讽刺和讽刺。



在特征请求挖掘的情况下，我们考虑以下元素来定义性能指标：a）真正的正面（TP）作为正确返回的特征请求，b）误报（FP）作为返回的结果，而不是实际的特征 请求，c）作为非特性请求不作为结果返回的真否定（TN），d）作为实际特性请求不作为结果返回的错误否定（FN）。 作为指标，我们考虑精度，召回率和马修斯相关系数（MCC）。 精度是实际功能要求（TP）返回结果与返回结果总数（TP + FP）之间的比率。 回想一下，实际功能请求（TP）返回的结果与输入中的功能请求总数（TP + FN）之间的比率。 最后，MCC被定义为（MCC = 1作为完美的预测因子）：



出于评估目的，我们在Google应用商店（2013年2月7日）下载了半数可用非免费应用的评论，并对其进行分析以检索其内容和元数据。我们之所以选择这些应用程序，是因为它们涉及大量的评论;我们储存的用于分析的游泳池包括136,998条评论。我们使用这些评论作为特征请求挖掘算法的输入（在将其内容分割为句子之后），并且基于上述度量评估结果。我们随机地选择了3000个算法返回的特征请求，人类编码器分析了这个样本来检查哪些是FP。编码员的任务是为每个分析结果确定用户表达的实际请求。没有这种要求的结果被认为是FP。总的来说，样本的计算精度为P = 0.85。在召回和MCC方面，我们随机选择了一个应用程序，并选择其评论作为衡量这两个指标的样本。我们查看了480条评论的样本，分成了778个句子，结果如表3所示。



我们使用LDA来确定作为评估结果而获得的功能请求中的主题。我们将每个功能请求视为一个文档，我们的目标是确定一组由与整个文档语料库相关的关键字所描述的主题。为了减少最终结果中的噪音，我们从整个语料库中提取了英语停用词和识别功能请求的关键字（见表I）。为了更好的精确度，我们使用主题和关键词数量的不同组合来运行LDA算法（表IV）。此外，为整个语料库确定一个主题会导致与下列关键字（游戏（0.015），应用程序（0.011），使用（0.010），更新（0.008），级别（0.007），功能（0.006），工作（ （游戏（0.130）），（使用（0.097）），（更新（0.076）），（应用程序（0.061）），（功能（0.006）），更好（0.0057）游戏是一个与“更新”，“级别”，“更好”等关键词相关的反复出现的主题;用户经常要求游戏应用程序频繁更新作为一个话题，应用往往与“支持”，“方式”，“功能”，“能力”等关键词相关联，分析评论时，用户使用“方式” “选项”和“功能”是指在使用应用程序时执行特定操作的可能性，用户要求为其所使用的应用程序提供更好的支持和自定义选项，更新作为与关键字“”，“better“，”支持“，”星星“。

表5

通过这项工作，我们的目标是支持应用程序开发人员检索并理解用户通过在线评论表达的功能请求。 我们开发了一个能够从移动应用在线评论中提取功能请求的系统，我们用它从136,998个在线评论中抽取了这样的请求。 然后，我们使用LDA模型来确定可能与这些请求相关的主题。 我们了解到，大多数用户的请求涉及对应用程序的更好支持，更新频率更高，游戏应用程序的新级别以及更多自定义选项。