如何在众包需求工程后自动确定项目或者产品类别

1. 众包需求工程之后获得的Artifacts有：产品简单描述和用户需求。那么如何根据这些内容自动确定一个产品或者项目的类别呢？开源社区Sourceforge.net中有一个良好的项目分类体协，一共有十个大类，每一个大类下面又至少有一个层级的小类别，有的还有两个、三个层级的小类别。本研究的主要目的就是根据众包需求工程的输出自动确定项目或者产品的类别，包括大类别和小类别。

2. 分类的标准就是Sourceforge.net中项目分类体系，那么首先需要搞清楚在Sourceforge.net中是如何根据项目的内容对项目进行分类的。探究基于哪一种算法计算的分类结果更接近该开源社区的分类方案。确定项目类别的主要目的是让需要的用户能够根据一般的分类体系精确的找到该项目，然而，在开源社区中或者软件repository中，一般会有另一种帮助用户更加直接找到需要的软件的方法，这就是推荐。精准的推荐机制能够让用户在最短的时间内找到想要的项目，即在浏览最少的项目具体内容之后便可以找到目标项目。那么，基于项目分类的标准是否能够辅助确定精准推荐的相关项目呢？是否可以通过调整项目简介的内容提高项目被检索到的可能性？

3. 基于以上描述，本研究中根据Sourceforge.net中项目知识需要解答的问题有：

RQ1. 不同粒度上属于相同类别的项目在哪些属性上有明显的共性，即不同类别的项目在哪些属性上有明显的差异？（主成份、词频、关键词、TFIDF、不同主题上的统计量）

RQ2. 基于项目属性的聚类方法适用于确定项目的类别吗？

RQ3. 基于项目属性的分类方法适用于确定项目的类别吗？

RQ4. Sourceforge.net的推荐项目逻辑大致是怎样的？如何验证？（现有推荐结果构图；基于图的划分算法结果怎样；标记类别、各种相似度等边信息后分析有关联项目之间的关系特征）

RQ5. 基于用户需求的聚类结果对推荐相关项目有什么样的作用？

4. 数据集准备

（1） 抓取工作要求：抓取sourceforge.net上项目数据，抓取那些包含15个以上用户需求的项目；统计各个类别项目数量；首先抓取项目地址页；然后组织分配抓取任务，设计一个云平台上任务管理机制，实现协作完成抓取任务。

（2） 具体爬取信息：

|  |  |
| --- | --- |
| 项目名称 | 项目简介：Description和Features |
| 项目主页地址 | 项目独立主页地址 |
| 项目所属分类（多层，多个） | 推荐项目的名称和主页地址 |
| 项目评分；评分人数；五分人数 | 用户需求数量 |
| 项目总下载量 | 用户需求起始地址 |
| 项目一周内下载量 | 所有的用户需求：Summary；priority; created time; content; labels; status |
| 上次更新时间；注册时间 |
| 目标用户群 |  |
| 平台或系统；开发语言 |  |