**CSDN社区用户兴趣标签挖掘报告**

1. **数据与目标**

CSDN（www.csdn.net）是中国最大的开发者服务平台，也是全球最大的中文IT技术社区，拥有5000万注册用户，每天有数十万用户在论坛上进行IT技术的交流、咨询和分享。

现在已经有一些用户的兴趣数据，希望可以通过数据分析，聚焦CSDN用户画像问题，找到用户兴趣标签之间的关系，以期可以为新用户做精准地兴趣标签推荐。

数据包含1055名CSDN用户的的标签空间，共包含42个标签，每一行代表一个用户的3个标注的兴趣标签，共包含4个字段，依次为用户编号、兴趣标签1、兴趣标签2、兴趣标签3，用\001分开。

1. **数据分析过程**

整个数据分析过程分为下面几个步骤：

数据分析的代码在附的Python文件中。

1. 规整数据

为了能够很好地做兴趣标签的分析，我首先对拿到的数据进行数据收集和整理编号，以利于下一步的分析处理。统计标签与用户数，针对每一个标签，生成与其共同出现过的标签集合

1. Jaccard系数建立标签之间的相似矩阵

这里我用的是Jaccard系数。Jaccard index,又称为Jaccard相似系数，用于比较有限样本集之间的相似性与差异性。Jaccard系数值越大，样本相似度越高。

观察相似矩阵，对每个标签取与其最相似的3个标签，得到下列结果：

# 推荐系统: 人机交互, 嵌入式开发, 桌面开发

# 企业信息化: 移动开发, 云计算, 信息安全

# 电子商务: 项目管理, 软件工程, 数据库

# 人工智能: 项目管理, 电子商务, 并行及分布式计算

# 地理信息系统: 网络管理与维护, 数据挖掘, 计算机辅助工程

# 数据恢复: 游戏开发, 算法, 硬件

# 信息安全: 移动开发, 软件工程, 企业信息化

# 数据挖掘: 增强现实, 计算机辅助工程, web开发

# 嵌入式开发: 地理信息系统, 机器人, 算法

# 文字识别: 机器学习, 商业智能, 网络管理与维护

# 数据库: 网络与通信, 物联网, web开发

# 虚拟化: web开发, 企业信息化, 图像处理

# 虚拟现实: 虚拟化, 数据库, 数据挖掘

# 数据可视化: 多媒体处理, 虚拟化, web开发

# 网络与通信: 移动开发, 数据库, 软件工程

# 计算机辅助工程: 数据挖掘, 虚拟现实, 硬件

# 计算机视觉: 自然语言处理, 数据挖掘, 多媒体处理

# 桌面开发: 项目管理, 网络与通信, 虚拟现实

# 游戏开发: web开发, 系统运维, 算法

# 机器学习: 桌面开发, web开发, 虚拟化

# 机器人: web开发, 电子商务, 地理信息系统

# 商业智能: 算法, 计算机辅助工程, 地理信息系统

# 语音识别: 计算机视觉, 网络管理与维护, 机器人

# 软件工程: 移动开发, 信息安全, 云计算

# 云计算: 移动开发, 软件工程, 企业信息化

# 物联网: web开发, 云计算, 项目管理

# web开发: 物联网, 虚拟化, 数据库

# 人脸识别: 虚拟现实, 计算机视觉, 桌面开发

# 自然语言处理: 计算机视觉, 机器学习, 人脸识别

# 图像处理: 移动开发, 软件工程, 企业信息化

# 硬件: 大数据技术, 物联网, 项目管理

# 多媒体处理: 企业信息化, 大数据技术, 云计算

# 项目管理: 大数据技术, 软件工程, 网络与通信

# 深度学习: web开发, 桌面开发, 数据挖掘

# 算法: 系统运维, 商业智能, 地理信息系统

# 移动开发: 软件工程, 云计算, 信息安全

# 并行及分布式计算: 企业信息化, 云计算, 多媒体处理

# 系统运维: 图像处理, 大数据技术, 项目管理

# 大数据技术: 项目管理, 图像处理, 移动开发

# 增强现实: 数据挖掘, 项目管理, 企业信息化

# 人机交互: 推荐系统, 深度学习, 机器学习

# 网络管理与维护: 地理信息系统, 虚拟现实, 并行及分布式计算

这里我们可以看到利用Jaccard 系数来做相似度分析的结果，整体上看有一定的准确度，比如“虚拟现实: 虚拟化, 数据库, 数据挖掘”，“人机交互：推荐系统、深度学习、机器学习”，但是我们可以看到的是，并不是所有的相似度都是按照相同类型来的，实际上，用户的兴趣标签很可能是很多个方向的，像是图像处理对应的有企业信息化，语音识别对应的有网络管理与维护。这样保证了不会所有的推荐都是基于同类兴趣，也更容易被用户所接受。

1. 获取标签向量特征

为了聚类，我们希望把每个标签对应一个特征向量，且假设上述相似矩阵任意的元素，即任两个标签的相似度，为他们特征向量的内积。一般的特征向量维度可任意指定，这里为了方便，我们取向量的维度和样本数量相同，即k=42，此时只要求相似矩阵x的平方根即可，这在umpy里只要下面一行代码：

X = linalg.sqrtm(m)

1. 用k-means做标签聚类

根据4，得到特征向量后，进行聚类。我们采用常用的K-means算法，并经过若干次尝试后，发现聚为七类效果比较好， 输入的结果如下：

{'游戏开发', '文字识别', '商业智能', '数据恢复', '算法', '嵌入式开发'}

{'电子商务', '机器人', '人工智能', '人机交互'}

{'机器学习', '并行及分布式计算', 'web开发', '多媒体处理', '虚拟化', '深度学习', '增强现实'}

{'软件工程', '项目管理', '云计算', '大数据技术', '信息安全', '图像处理', '数据库', '系统运维', '移动开发', '网络与通信', '企业信息化'}

{'地理信息系统', '物联网', '硬件'}

{'桌面开发', '网络管理与维护', '虚拟现实', '数据可视化', '数据挖掘', '人脸识别', '计算机辅助工程', '推荐系统', '计算机视觉'}

{'语音识别', '自然语言处理'}

可以看到“机器视觉、人脸识别、虚拟现实”等在一类，“软件工程', '项目管理、系统运维”等被分在一起，还是具有一定的可靠性，但是还是存在一些奇怪的地方，比如“机器学习和多媒体处理”在一起，这些比较异常的情况，暂时还不知道发生的原因，猜测可能是数据样本的有限性。

**三、思考**

在这个项目中，获取标签向量时，其实k值得取值是指定的，可以无限逼近，选取特征k取42 其实只是为了计算更方便。本次处理结果虽然具有一定可靠性，但是一千个样本数量还不够多，如果可以对更多数据进行分析，会得到更精准地模型。

这个方法在实际中有很多地方可以应用，微博用户的话题推荐，微信公众号内容的推荐等，其实本质上和本次的数据分析都有一定的相似之处。