图书馆自画像技术报告

王德涵　　冯泉弼　　郑荣菲　　杨博乔　　王子铭

2024.1.5

1　项目背景

随着高校办学规模的扩大，各个高校图书馆的数量也在不断增多，图书资源及相应资源也变得更加丰富。然而，面对海量的资源，高校学生如何选择和利用这些资源成为一个问题。同时，高校图书馆如何分析和预测学生的需求，根据学生的阅读倾向进行阅读推广，实现个性化服务也是一个难题。

在这个大数据时代，国家“十三五”规划提出要实施国家大数据战略，这也给高校图书馆提供了机会和挑战。通过使用大数据和用户画像技术，高校图书馆可以精确分析、预测读者的资源需求，并进一步发掘读者的潜在需求，从而帮助高校图书馆为读者提供个性化服务。

为了更好地服务学生，提高图书馆的工作效率，我们计划开发一个大学图书馆自画像配套的查询网页。该网页将包含登录页面和查询页面，提供一系列的查询功能，以满足学生和图书馆工作人员的需求。

对于图书馆难以高效分析和预测读者的资源需求的问题，我们还设计了基于用户画像的图书馆毕业书单系统。该系统通过接入校园 CAS 认证系统，保障系统安全性，杜绝 CORS 请求。同时，采用 React.JS 框架开发美观流畅的前端界面。系统利用 TF-IDF 算法和 RAKE 算法对用户信息进行处理，从而构建词云。该平台实现了前后端分离的跨平台用户信息分析及展示，并同时部署在移动端和 Web 端，具备良好的实用性和可迁移性。

通过这个基于用户画像的系统，高校图书馆可以更加准确地了解读者的需求，为他们提供个性化的阅读推荐和服务。同时，图书馆工作人员也可以更加高效地管理图书资源，提高工作效率。这将使得高校图书馆能够更好地满足学生和教职员工的需求，为他们提供更好的学习和研究环境。

2　项目介绍

该项目在前人的分类方法的基础上，制作了登录页面，对进入图书馆的学生的分类（博览者、钻研者、思辨者、勤学者、学习者、无闻者）提供查询人数等方面的功能，具体功能如下：

- 查询在某一分类下具有的读者数量，方便表现南方科技大学的学生阅读以及自习水平。
- 查询在每一种分类中，每个年级的学生数量，以便于进行年级之间的学生比较。

- 可以单独获取某一用户的分类，并把多个用户的数据同时呈现在网页上，方便图书馆管理员和学生进行查阅。
- 可以上传含多个用户学号/工号的 Excel 文件，生成多个读者画像的两张图片，便于下载打印读者的图书馆自画像。
- 查询在指定日期之间访问图书馆自画像网站的人数

3 代码组成

3.1 网页部分代码组成

代码是一个基于 React 框架构建的网页应用。React 是由 Facebook 开发的用于构建用户界面的 JavaScript 库。它采用组件化开发模式，使得用户界面的构建更加模块化和可复用。React 使用虚拟 DOM 技术，通过高效的更新策略减少对实际 DOM 的操作，提高性能。

(1) 样式定义部分：在代码中使用了 CSS 语法来定义.App、.App-logo、.App-header 等类的样式，包括居中对齐、动画效果等。

(2) 页面结构部分：通过 、<header>、<img>、<p>、<a> 等 HTML 元素构成了页面的结构，并使用了 React 的组件化思想来构建页面。

(3) 路由部分：使用了 React Router 来定义了多个页面的路由，包括 <Login>、<Home>、<Fun1> 等页面组件。

主体代码位于 src/pages 页面下，Fun1-4 对应 4 个功能，Login 为登录界面，Home 为功能选择界面

有关依赖项目，首先你的电脑上需要安装有 node。你可以通过 npm install -g cnpm –registry=https://registry.npmmirror.com，npm config set registry https://registry.npmmirror.com 命令来安装环境，并运行 npm install all。我们需要通过 npm start 来启动网页，更多的功能可以查看 README.md 文件

效果在答辩视频中进行展示

3.2 服务器部分代码组成

我们使用 Gin 构建的基础服务器，采用 GORM 与 Postgresql 数据库交互，JWT-GO 用于鉴权，go-swagger 生成 api 文档。docs 部分是 swagger 生成的文档，global 部分包含运行过程中需要用到的全局变量，initialize 部分负责服务器初始化，middleware 是代码在客户和具体的 handler 之间的处理函数，router 负责路由，services 处理具体的请求，table 包含与数据库交互的函数，test 是测试代码，utils 包含需要用到的小块代码。

你需要先在 config.json 中设置好与数据库连接的相关参数，随后通过 go run main.go 指令启动程序。服务器默认运行在本地，监听 9090 端口

4 算法设计

4.1 前人研究成果

对于用户画像中词云的构建，系统采用了 TF-IDF 算法和 RAKE 算法分别对中文和英文关键词进行提取。

4.1.1 TF-IDF 算法

TF-IDF（Term Frequency-Inverse Document Frequency）是一种常用于信息检索和文本挖掘的算法，用于评估一个词对于一个文档集合中的某个文档的重要程度。

TF-IDF 算法基于以下两个概念：

1. 词频（Term Frequency，TF）：表示某个词在一个文档中出现的频率。词频可以通过计算某个词在文档中的出现次数，并进行归一化处理，以避免文档长度的差异对结果的影响。

2. 逆文档频率（Inverse Document Frequency，IDF）：表示一个词在整个文档集合中的普遍重要程度。逆文档频率通过计算包含某个词的文档数目的倒数，并进行对数变换来降低高频词对结果的影响。

TF-IDF 的计算公式如下：

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

其中，t 表示词语（term），d 表示文档，D 表示文档集合。

TF-IDF 的计算过程如下：

1. 计算词频（TF）：对于给定的词语 t 和文档 d，计算词频，表示为 $\text{TF}(t, d)$。

2. 计算逆文档频率（IDF）：对于给定的词语 t 和文档集合 D，计算逆文档频率，表示为 $\text{IDF}(t, D)$。计算公式为：

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

其中，$|D|$ 表示文档集合的总文档数目，$|\{d \in D : t \in d\}|$ 表示包含词语 t 的文档数目。

3. 计算 TF-IDF：将词频和逆文档频率相乘，得到最终的 TF-IDF 值。

TF-IDF 算法的直观理解是，一个词在某个文档中的重要性随着它在该文档中的词频增加而增加，但同时又要考虑到它在整个文档集合中的普遍程度。因此，对于在某个文档中频繁出现但在整个文档集合中普遍出现的词，TF-IDF 值较低；而对于在某个文档中频繁出现且在整个文档集合中较少出现的词，TF-IDF 值较高。

4.1.2 RAKE 算法

RAKE 算法（Rapid Automatic keyword extraction）简单而高效，能够提取一些比较长的专业术语，适合用于在英文文本中提取关键词。RAKE 算法主要包含以下几个步骤：（1）使用标点符号（如半角的句号、问号、感叹号、逗号等）将一篇文本划分为若干分句，对于每一个分句，使用停用词作为分隔符将其分为若干短语，这些短语作为关键词的候选词。（2）每个短语通过空格分为

若干个单词，计算每一个单词在短语中的共现词数，构建词共现矩阵。(3) 计算每个单词的词频与度。(4) 将度与词频的商作为单词的得分。(5) 按照单词得分降序排列，输出单词所在的短语。

4.2 基于信息熵和序列频域性质的 k-means 算法优化

首先，计算每个维度的信息熵，表示为 $H(d)$。信息熵可以使用以下公式计算:

$$H(d) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

其中，n 是该维度上的属性值数量，x_i 是该维度上的第 i 个属性值，$P(x_i)$ 是该属性值出现的概率。

如，学生可以离散的借阅 0 到 18 本书，则在计算“借阅图书数”一维度的信息熵时，n=18，x_i=i-1

然后，选择信息熵最大的 k 个维度作为初始解的维度。表示为 $D_{\text{init}} = \{d_1, d_2, \ldots, d_k\}$。

最后，使用选择的维度作为初始聚类中心，应用 k-means 算法进行聚类。

这个过程可以用以下 LaTeX 公式总结:

$$D_{\text{init}} = \arg\max_{D} \{H(d) \mid d \in D, |D| = k\}$$

其中，D 是所有可能的维度集合。

接下来是优化后的 K-means 算法应用在数据库分类上的流程

令 $X = \{x_1, x_2, \ldots, x_n\}$ 表示包含 n 个数据点的集合，这里的 x_i 表示为一个向量对象，也就是数据库中的一行。K（在该 project 中表示为 6）表示聚类的数量。

1. 初始化：对归一化后信息熵最大的维度，在其频域上进行 k 等分以选择中心点，接下来对于信息熵最大的 K 个维度，按序进行中心的设置。

2. 分配数据点到最近的聚类中心：对于每个数据点 x_i，计算其与各个聚类中心的距离，表示为 $d(x_i, c_j)$，并将其分配到距离最近的聚类中心所属的类别，表示为 S_i。

3. 更新聚类中心：对于每个类别 S_i，计算该类别中所有数据点的均值，并将均值作为新的聚类中心 c_i。

4. 重复步骤 2 和 3，直到满足停止条件。停止条件可以是达到最大迭代次数，或者聚类中心不再发生明显变化。

5. 输出最终的聚类结果。

在算法中，最小化距离平方和的目标函数可以表示为（由于归一化，我们采用了曼哈顿距离计算）:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - c_j\|$$

4.3 基于 gpt 的图像生成

在这个 project 种，我们采用了 GPT-4 技术，以重新生成了六种不同类型的人物头像画像。GPT-4 是一种先进的自然语言处理模型，具有强大的生成能力。通过该技术，我们能够根据特定的人物描述和特征，生成画风统一的的头像画像，包括但不限于人物配色，周围的饰品等。

具体生成 6 张图片效果如 Fig1,2,3,4,5,6所示

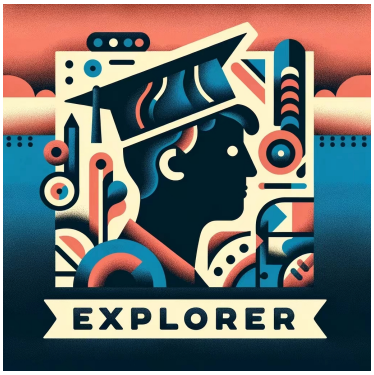

Figure 1: 博览者

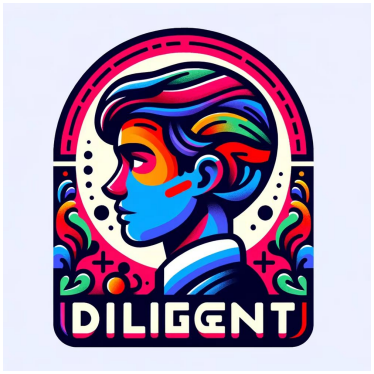

Figure 2: 勤学者

Figure 3: 思辩者

Figure 4: 无闻者

Figure 5: 学习者

Figure 6: 钻研者

5 结语

该系统主要面向南方科技大学的学生，旨在帮助学生了解自己对图书馆资源的利用情况，并为图书馆管理人员提供个性化、智能化的服务。

系统采用了 React.JS 框架搭建前端平台，同时引入了 CAS 系统和 K8s 容器集群管理系统，以确保用户信息的安全性和平台的高效性。通过用户画像作为平台主体，系统能够精准地分析、预测学生的需求，为学生提供个性化的图书馆服务。

该系统的主要功能包括学生个人书单的生成和管理、图书馆资源的推荐和个性化服务等。学生可以根据自己的兴趣和需求，在系统中选择自己感兴趣的图书，并生成个人的书单。系统会根据学生的选择和阅读记录，推荐相关的图书资源，帮助学生更好地发现和获取所需的知识。

对于图书馆管理人员来说，该系统提供了一个有效的工具来分析、预测学生的需求。通过对学生的个人书单和阅读记录的分析，管理人员可以了解学生的兴趣和需求，从而提供更加个性化、智能化的图书馆服务。这有助于提高图书馆资源的利用率，提升学生的学习体验。

总之，基于用户画像的校园图书馆毕业书单系统为南方科技大学的学生提供了个性化、智能化的图书馆服务。通过该系统，学生可以更好地了解自己对图书馆资源的利用情况，并获得相关的推荐和个性化服务。对于图书馆管理人员来说，该系统提供了一个有效的工具来分析、预测学生的需求，为学生提供更好的图书馆服务。这一系统的设计和应用为智慧校园建设提供了有益的借鉴和帮助。