



華東師範大學
EAST CHINA NORMAL
UNIVERSITY

《数据库系统及应用实践》课程项目报告

CovidLit Search(Milestone 1)

小组成员：

李鹏达	10225101460
武泽恺	10225101429
王 力	10225101434

2024 年 5 月

目 录

1	项目简介	1
2	数据库设计	1
2.1	概述	1
2.2	表结构	1
2.2.1	实体集	1
2.2.2	联系集	3
2.3	E-R 图	4
2.4	schema	5
3	数据集	5
3.1	数据来源	5
3.2	数据处理	5
3.2.1	数据集结构	5
3.2.2	处理方式	6
3.3	数据量	7
4	功能设计	7
4.1	用户功能	7
4.2	搜索功能	7
4.3	文献功能	8
5	用户界面设计	8
5.1	首页	8
5.2	搜索页	9
5.3	文献详情页	9
5.4	注册与登录页	10
5.5	用户页面	11
6	SQL	12
6.1	用户相关功能	13
6.2	搜索相关功能	13
6.3	文献相关功能	14

1 项目简介

本项目 (“CovidLit Search”) 致力于为研究人员提供一个方便的、用户友好的 COVID-19 相关文献检索工具，以帮助他们更快地找到相关文献进行参考研究。

本项目的目标是通过提供一个简单友好的界面，使用户能够快速搜索到与 COVID-19 相关的文献，并且能够根据作者、时间和期刊等信息进行检索。此外，本项目也允许用户通过研究方向、研究对象和研究问题等信息准确检索部分文献并对其他文献进行模糊搜索，以便用户查找到最相关的文献。

另外，本项目还提供了一个用户注册和登录系统。用户可以通过注册登录后，将自己的搜索历史保存在云端，以便在不同设备上查看自己的搜索历史。用户还可以将自己感兴趣的文献加入到自己的收藏夹中或订阅期刊，以便在以后查看。

本项目名称为 “CovidLit Search”，意为 “COVID-19 Literature Search”，即 COVID-19 相关文献检索。

2 数据库设计

2.1 概述

为了实现文献检索系统，我们需要设计一个数据库来存储文献、期刊、作者和引用等信息。我们将使用 MySQL 数据库来存储这些信息。

为了实现相关功能，我们考虑使用 Entity-Relation 模型来设计数据库。我们考虑设计文章 (article)、作者 (author)、期刊 (journal) 和用户 (user) 等实体集，以及撰写 (write)、引用 (cite)、订阅 (subscribe)、收藏 (collect) 和浏览历史 (history) 等联系集。

其中，文章 (article) 与作者 (author) 通过撰写 (write) 联系集相连，表示作者撰写了文章；文章 (article) 与期刊 (journal) 通过发表 (publish) 联系集相连，表示文章在期刊上发表；文章 (article) 与文章 (article) 通过引用 (cite) 联系集相连，标识不同文章之间的引用与被引用关系；用户 (user) 与文章 (article) 通过收藏 (collect) 浏览历史 (history) 联系集相连，表示用户收藏了文章和用户的历史浏览文章；用户 (user) 与期刊 (journal) 通过订阅 (subscribe) 联系集相连，表示用户订阅了该期刊，可能希望获取该期刊的最新文章。

2.2 表结构

2.2.1 实体集

用户 (user) 表存储用户的基本信息，其结构如下：

字段名	类型	主键	外键	说明
<i>id</i>	INT	是		用户 ID, 自动递增
<i>nickname</i>	VARCHAR(100)			用户名 (昵称)
<i>email</i>	VARCHAR(200)			邮箱
<i>password</i>	VARCHAR(100)			密码 (加密后)
<i>avatar</i>	VARCHAR(500)			头像
<i>motto</i>	VARCHAR(1000)			座右铭
<i>collage</i>	VARCHAR(100)			学院 (学校)
<i>subscribe_email</i>	BOOLEAN			是否订阅邮件
<i>save_history</i>	BOOLEAN			是否保存历史记录

表 2.1 用户 (*user*) 表

文章 (*article*) 表存储文章的基本信息, 其结构如下:

字段名	类型	主键	外键	说明
<i>id</i>	VARCHAR(50)	是		文章 ID
<i>title</i>	VARCHAR(1000)			文章标题
<i>abstract</i>	TEXT			摘要
<i>doi</i>	VARCHAR(50)			数字对象唯一标识符
<i>license</i>	VARCHAR(50)			许可
<i>publish_time</i>	DATETIME			发表时间
<i>url</i>	VARCHAR(800)			文章 URL
<i>study_type</i>	VARCHAR(500)			研究类型
<i>addressed_population</i>	VARCHAR(1000)			研究对象人群
<i>challenge</i>	VARCHAR(2000)			挑战/研究问题
<i>focus</i>	VARCHAR(100)			研究重点

表 2.2 文章 (*article*) 表

期刊 (*journal*) 表存储期刊的基本信息, 其结构如下:

字段名	类型	主键	外键	说明
<i>name</i>	VARCHAR(100)	是		期刊名称
<i>description</i>	VARCHAR(1000)			期刊描述

表 2.3 期刊 (*journal*) 表

作者 (*author*) 表存储作者的基本信息, 其结构如下:

字段名	类型	主键	外键	说明
<i>name</i>	VARCHAR(100)	是		作者姓名
<i>email</i>	VARCHAR(1000)			邮箱
<i>lab</i>	VARCHAR(1000)			所在实验室
<i>institution</i>	VARCHAR(1000)			所在机构
<i>country</i>	VARCHAR(100)			国家
<i>post_code</i>	VARCHAR(100)			邮政编码
<i>settlement</i>	VARCHAR(100)			定居点 (城市)

表 2.4 作者 (*author*) 表

2.2.2 联系集

撰写 (*write*) 联系集存储文章与作者之间的联系，其结构如下：

字段名	类型	主键	外键	说明
<i>author_name</i>	VARCHAR(100)	是	<i>author(name)</i>	作者姓名
<i>article_id</i>	VARCHAR(50)	是	<i>article(id)</i>	文章 ID

表 2.5 撰写关系 (*write*) 表

发表 (*publish*) 联系集存储文章与期刊之间的联系，其结构如下：

字段名	类型	主键	外键	说明
<i>journal_name</i>	VARCHAR(100)	是	<i>journal(name)</i>	期刊名称
<i>article_id</i>	VARCHAR(50)	是	<i>article(id)</i>	文章 ID
<i>volume</i>	VARCHAR(50)			卷号
<i>pages</i>	VARCHAR(200)			页码

表 2.6 发表关系 (*publish*) 表

引用 (*cite*) 联系集存储文章与文章之间的引用关系，其结构如下：

字段名	类型	主键	外键	说明
<i>citing_id</i>	VARCHAR(100)	是	<i>article(id)</i>	引用文章 ID
<i>cited_id</i>	VARCHAR(100)	是	<i>article(id)</i>	被引用文章 ID

表 2.7 引用关系 (*cite*) 表

收藏 (*collect*) 联系集存储用户与文章之间的收藏关系，其结构如下：

字段名	类型	主键	外键	说明
<i>user_id</i>	INT	是	<i>user(id)</i>	用户 ID
<i>article_id</i>	VARCHAR(50)	是	<i>article(id)</i>	文章 ID

表 2.8 收藏关系 (*collect*) 表

订阅 (subscribe) 联系集存储用户与期刊之间的订阅关系，其结构如下：

字段名	类型	主键	外键	说明
<i>user_id</i>	INT	是	<i>user(id)</i>	用户 ID
<i>journal_name</i>	VARCHAR(200)	是	<i>journal(name)</i>	期刊名称

表 2.9 订阅关系 (*subscribe*) 表

浏览历史 (*history*) 联系集存储用户与文章之间的浏览历史关系，其结构如下：

字段名	类型	主键	外键	说明
<i>user_id</i>	INT	是	<i>user(id)</i>	用户 ID
<i>article_id</i>	VARCHAR(50)	是	<i>article(id)</i>	文章 ID
<i>time</i>	DATETIME	是		浏览时间

表 2.10 浏览历史 (*history*) 表

2.3 E-R 图

根据上述设计，我们绘制了数据库的 E-R 图，如图 2.1 所示。

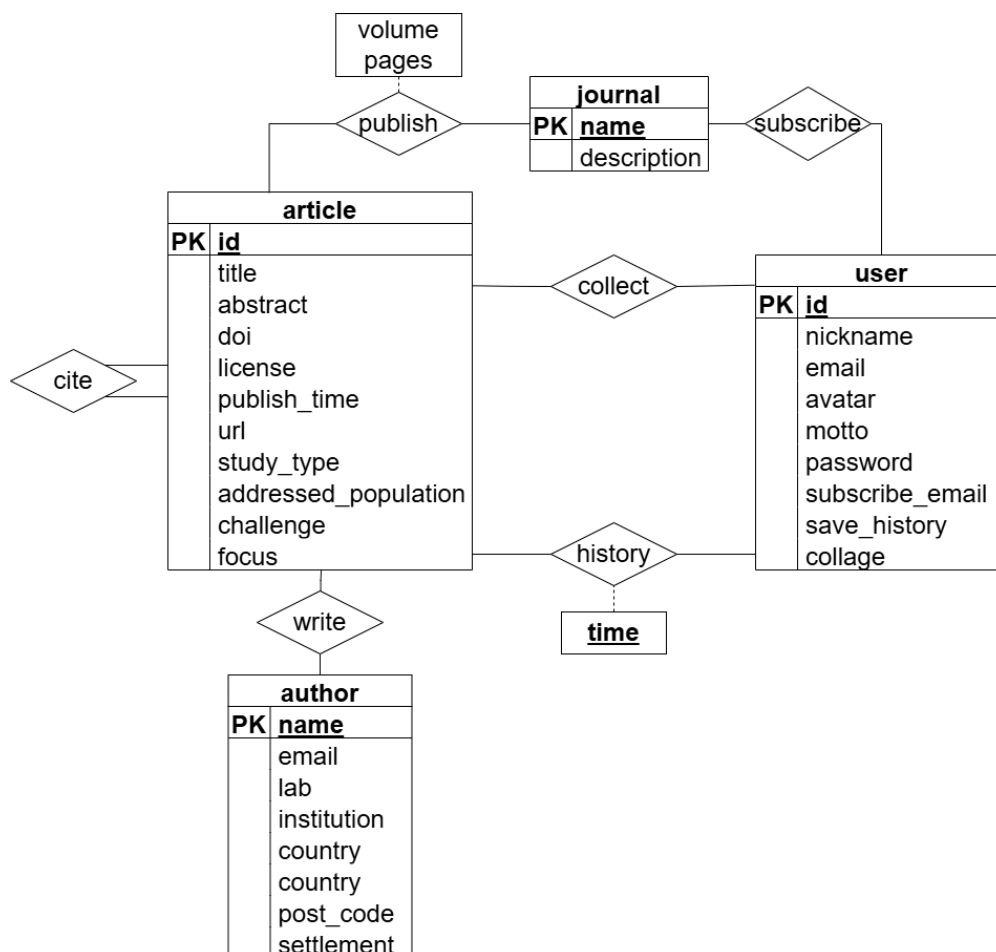


图 2.1 数据库 E-R 图

其中，PK 表示主键，FK 表示外键。

2.4 schema

根据上述设计，我们可以写出数据库的 `schema`，如下所示：

```
article(id, title, abstract, doi, license, publish_time, url, study_type,
        addressed_population, challenge, focus)
author(name, email, lab, institution, country, post_code, settlement)
write(author_name, article_id)
journal(name, description)
publish(journal_name, article_id, volume, pages)
cite(citing_id, cited_id)
user(id, nickname, email, password, avatar, motto, collage, subscribe_email,
     save_history)
collect(user_id, article_id)
subscribe(user_id, journal_name)
history(user_id, article_id, time)
```

3 数据集

3.1 数据来源

本项目使用的数据集是由美国白宫联合一系列顶尖研究机构提供的 COVID-19 Open Research Dataset (CORD-19)。该数据集可以在 Kaggle 上下载¹，它包含了超过 1,000,000 篇来自 PubMed、PMC、bioRxiv 和 medRxiv 等来源的 COVID-19 相关文献的元数据，其中 400,000 篇提供全文。此外，该数据集还包括超过 6,000 篇按研究方向分类的文献元数据，包括研究方向、研究对象和研究问题等信息。

该数据集的元数据包括文献标题、作者、摘要、发布时间、期刊、全文链接等信息。我们将使用这些信息来构建我们的文献检索系统。

3.2 数据处理

3.2.1 数据集结构

我们首先对数据集进行了初步的探索，其结构如图 3.1 所示。数据集主要包括以下几个部分：

1. `metadata.csv`：包含了文献的元数据，包括文献标题、作者、摘要、发布时间、期刊、全文链接等信息。
2. `metadata.readme`：包含数据集内容的更新日志。

¹ <https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge>

3. json_schema.txt: 包含了数据集中的 JSON 文件的结构。
4. COVID.DATA.LIC.AGMT.pdf: 包含了数据集的使用许可协议。
5. document_parsers/: 文件夹, 包含了数据集中的文献全文, 以 JSON 格式存储。
6. Kaggle/target_tables/: 文件夹, 包含了数据集中的研究方向分类的文献元数据, 包括研究方向、研究对象和研究问题等信息。
7. cord_19_embeddings/: 文件夹, 包含了数据集中的文献的嵌入向量。

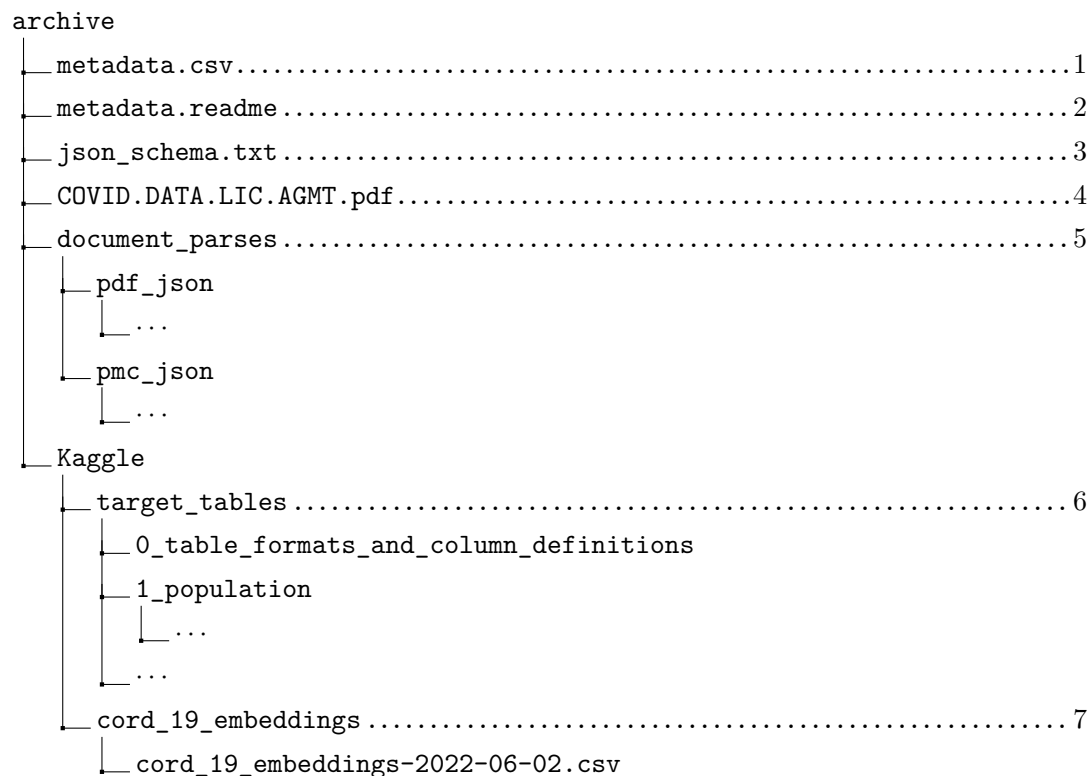


图 3.1 数据集结构

3.2.2 处理方式

由于数据集较大, 包含了大量无用信息, 且格式不是我们期望的格式, 我们需要对数据集进行预处理, 提取出我们需要的信息, 并插入数据库中。

我们使用 Python 脚本²进行处理, 处理方式如下:

1. 创建一些存储过程或函数, 用于简化后续大量数据的插入。
2. 根据 json_schema.txt 文件, 对 document_parsers/ 文件夹中的文献全文的 JSON 结构使用 Python 类进行建模, 以便后续提取信息。
3. 读取 metadata.csv 文件, 对每一条文献的元数据进行处理。

(a) 提取文献的基本信息, 包括标题、摘要、发布时间、期刊等。

²见附录

- (b) 根据原数据中的文献全文地址，读取对应的全文文件。
 - i. 提取文献的详细作者信息，包括作者单位、邮箱、国籍等。
 - ii. 提取文献引用信息，包括引用文献的标题、作者、期刊、发布时间等。
 - (c) 将提取的信息存储到数据库中。
4. 读取 Kaggle/target_tables/ 文件夹中的研究方向分类的文献元数据。
- (a) 提取文献的研究方向、研究对象和研究问题等信息。
 - (b) 将提取的信息存储到数据库中。

由于数据量较大而 Python 处理速度较慢，我们使用了数据库连接池与多线程来加速处理。

3.3 数据量

对于前期开发测试，我们使用数据集中的 1000 + 5000 余篇文献进行测试。对于后期开发，我们将使用全部数据集（1000000 + 5000 余篇文献）进行测试。最终，我们将使用全部数据集进行部署。

4 功能设计

4.1 用户功能

1. 用户注册：用户可以通过邮箱注册账号。
2. 用户登录：用户可以通过邮箱和密码登录账号。
3. 用户修改密码：用户可以通过邮箱验证或旧密码修改密码。
4. 用户信息修改：用户可以修改自己的昵称、头像、座右铭等信息。
5. 用户订阅：用户可以订阅感兴趣的期刊。
6. 用户收藏：用户可以收藏感兴趣的文献。
7. 用户浏览历史：用户可以查看自己的浏览历史。
8. 邮件订阅：用户可以订阅邮件，以从邮件中获取订阅的文献的更新。

4.2 搜索功能

1. 文献搜索：用户可以通过关键词搜索文献。
2. 高级搜索：用户可以通过作者、时间、期刊等信息进行高级搜索。
3. 研究方向搜索：用户可以通过研究方向、研究对象和研究问题等信息进行搜索。
4. 文献推荐：系统可以根据用户的搜索历史推荐相关文献。

4.3 文献功能

1. 文献详情：用户可以查看文献的详细信息。
2. 文献引用：用户可以查看文献的引用信息，包括引用与被引用，直接引用和间接引用。
3. 文献跳转：用户可以跳转到文献全文所在的网址。
4. 文献推荐：系统可以根据文献的内容推荐相关文献。
5. 文献分享：用户可以将文献分享到社交媒体或邮件。

5 用户界面设计

5.1 首页

首页包含搜索，用户注册、登录，文献推荐等功能，如图 5.1 所示。

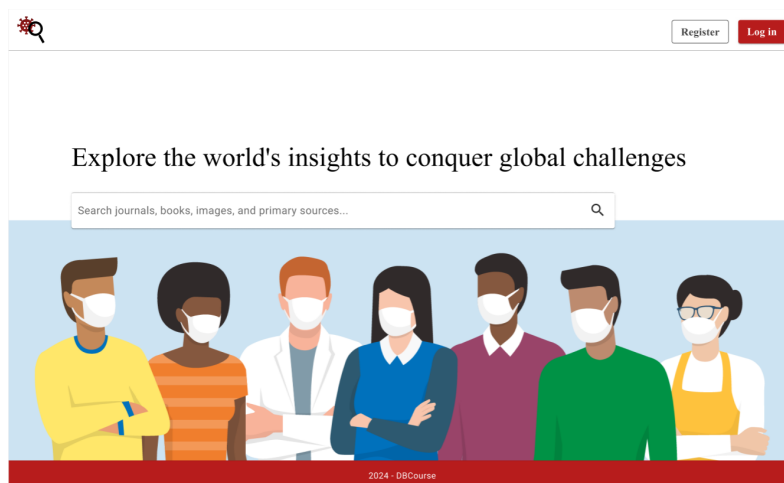


图 5.1 首页

用户在登入后，首页会显示用户昵称、头像等信息，并且可以跳转至用户个人主页、收藏、安全设置等页面，如图 5.2 所示。

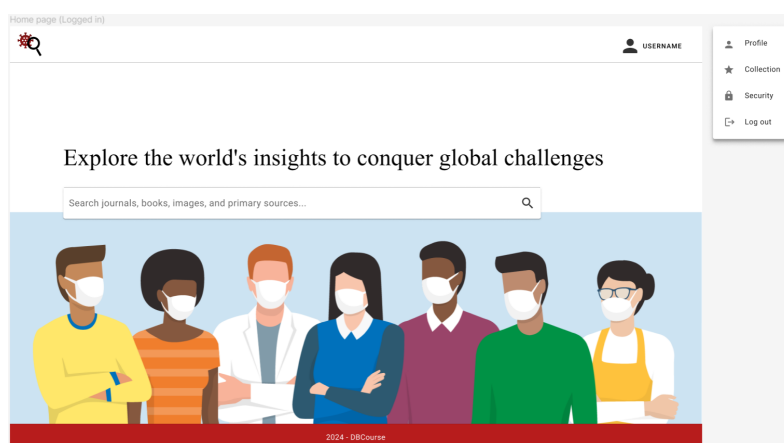


图 5.2 首页（登录后）

5.2 搜索页

用户在任意界面都可以进行搜索，搜索结果会显示在搜索页。搜索页包含搜索框、高级搜索、研究方向搜索等功能，用户可以对搜索结果进行排序，或对搜索结果进行筛选，如图 5.3 所示。

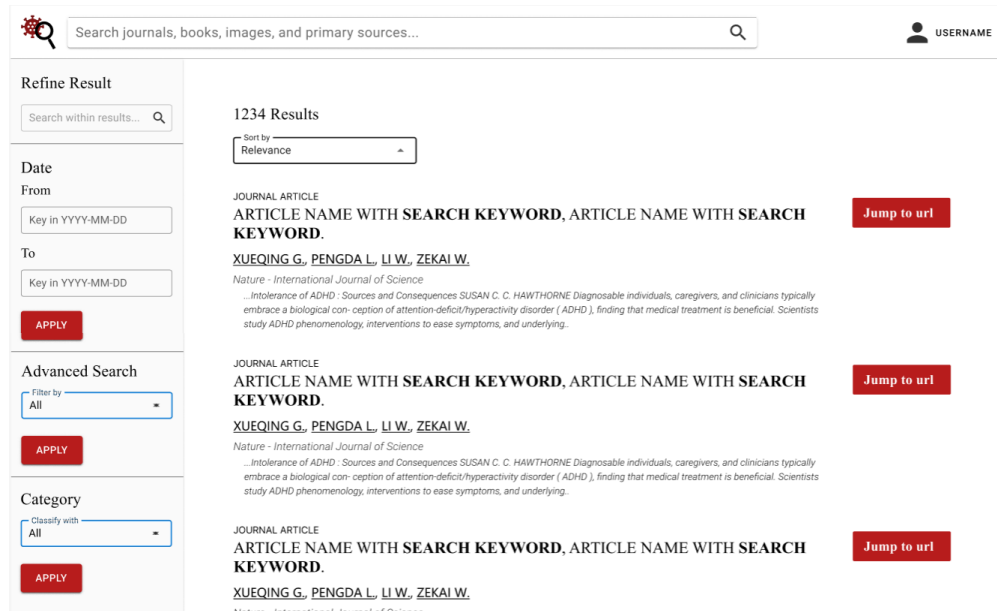


图 5.3 搜索页

5.3 文献详情页

用户点击文献，将跳转至文献页。用户可以查看文献的详细信息，包括标题、作者、摘要、发布时间、期刊等信息。用户可以查看文献的引用信息，包括引用与被引用，直接引用和间接引用。用户也可以跳转到文献全文所在的网址。如图 5.4 所示。

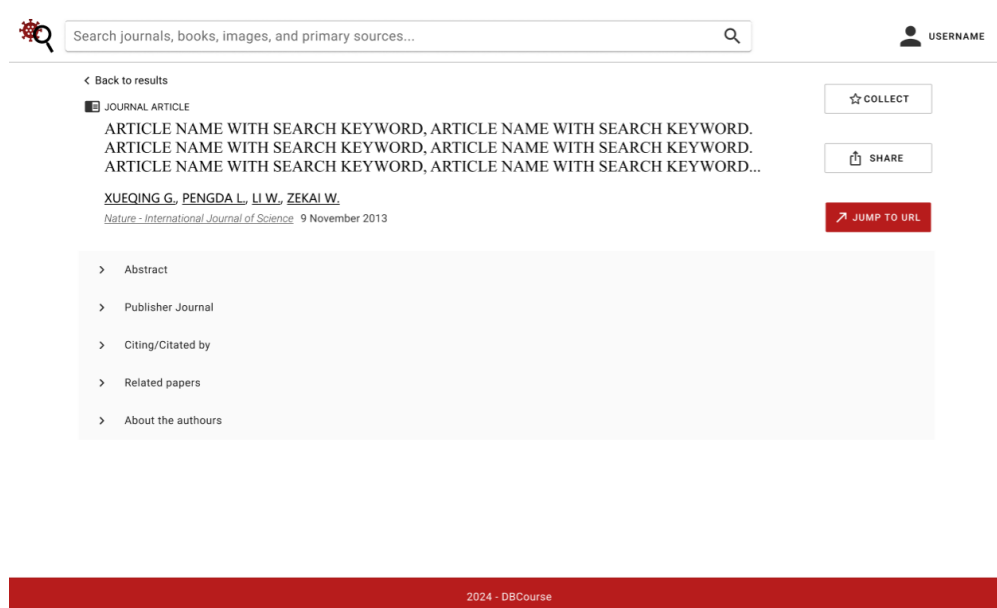


图 5.4 文献详情页

用户点击摘要、期刊、作者等栏目，可以展开查看更多相关信息，如图 5.5 所示。

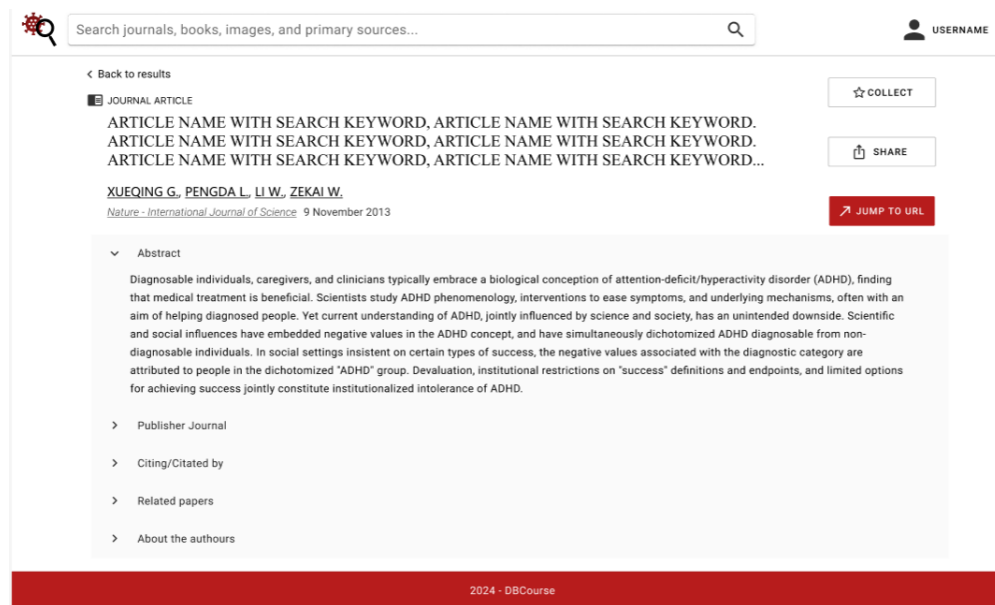


图 5.5 文献详情页（更多信息）

5.4 注册与登录页

用户在任意界面点击注册，将跳转至注册页。用户在注册页输入邮箱，通过验证码认证后，设置密码完成注册，如图 5.6 所示。

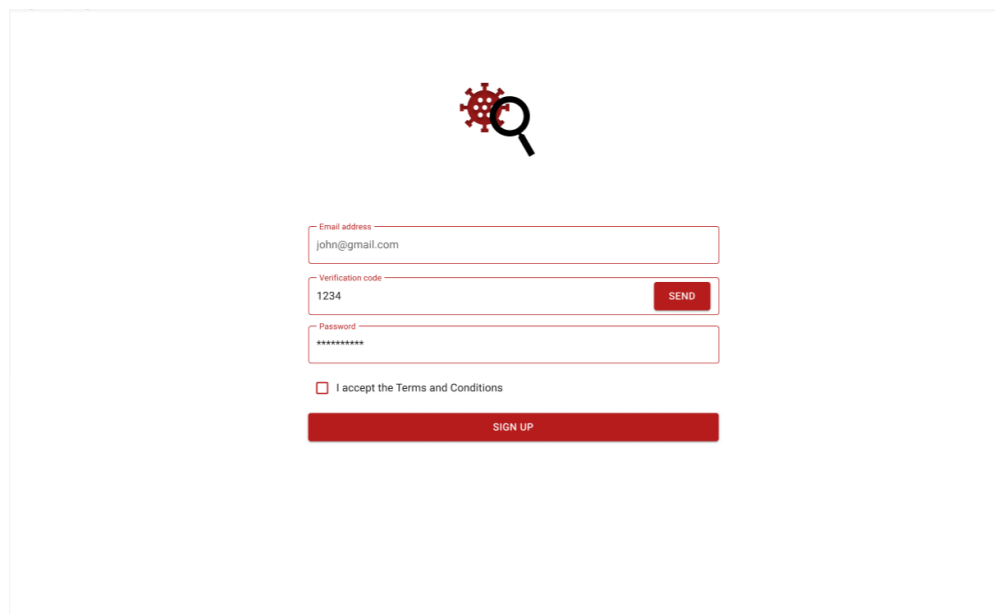
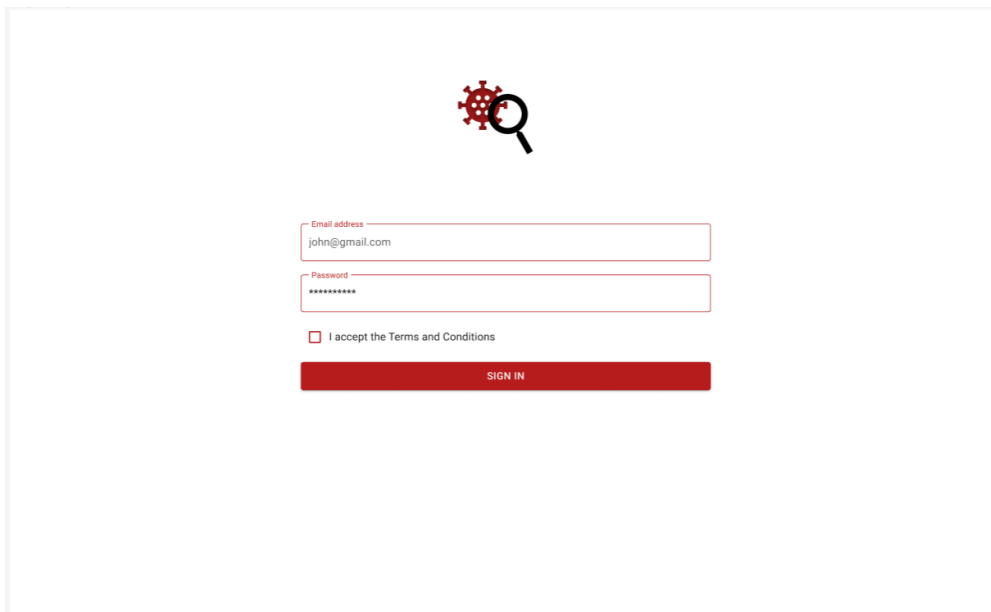


图 5.6 注册页

用户在任意界面点击登录，将跳转至登录页。用户在登录页输入邮箱和密码，完成登录，如图 5.7 所示。

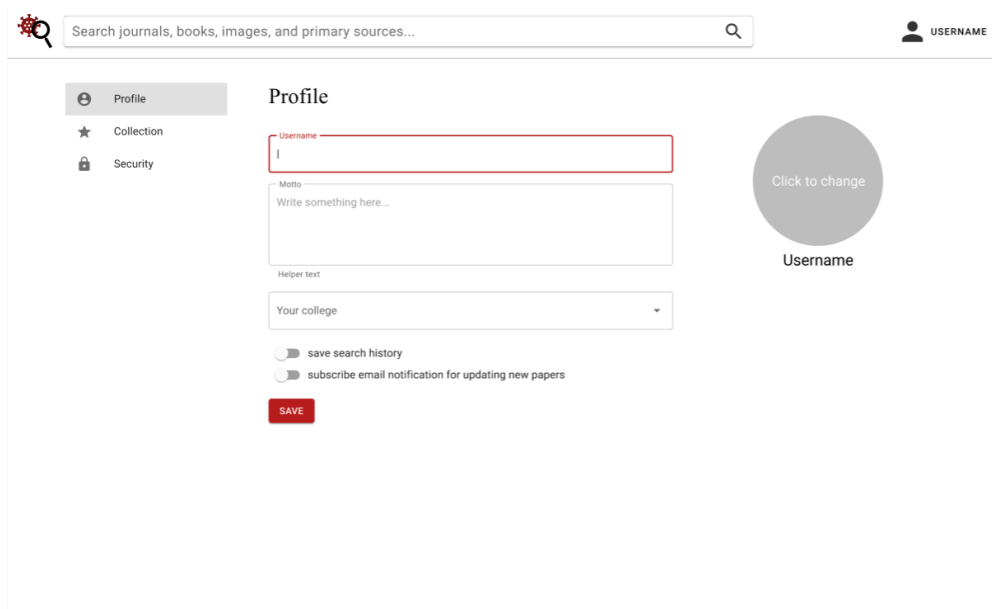


The login page features a red virus icon with a magnifying glass. Below it are two input fields: 'Email address' with the text 'john@gmail.com' and 'Password' with masked characters. A checkbox labeled 'I accept the Terms and Conditions' is positioned above a red 'SIGN IN' button.

图 5.7 登录页

5.5 用户页面

用户在任意界面点击用户信息按钮，将跳转至用户信息页面。在用户信息页面，用户可以查看和修改自己的个人信息，包括昵称、头像、座右铭、学院等信息。用户也可以修改个人信息偏好，如是否存储浏览历史等。如图 5.8 所示。



The user profile page has a top navigation bar with a search bar and a user icon labeled 'USERNAME'. A left sidebar contains links for 'Profile', 'Collection', and 'Security'. The main 'Profile' section includes a 'Username' input field, a 'Motto' text area, a 'Your college' dropdown menu, and two toggle switches for 'save search history' and 'subscribe email notification for updating new papers'. A red 'SAVE' button is at the bottom. On the right, there is a circular placeholder for a profile picture with the text 'Click to change' and 'Username' below it.

图 5.8 用户信息

用户点击收藏按钮，将跳转至收藏页面。在收藏页面，用户可以查看自己收藏的文献，如图 5.9 所示。

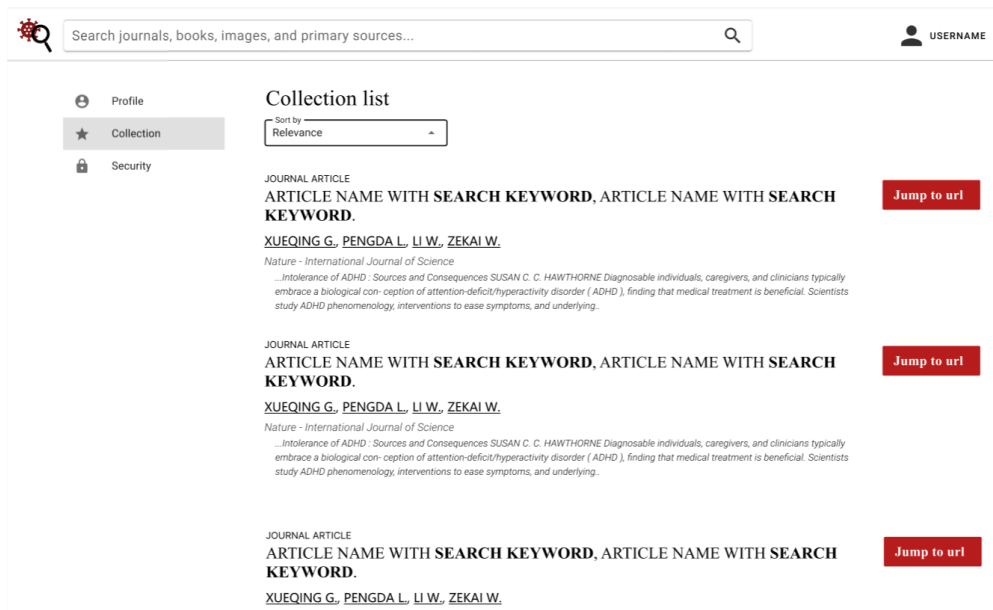


图 5.9 收藏

用户点击安全设置按钮，将跳转至安全设置页面。在安全设置页面，用户可以修改密码，如图 5.10 所示。

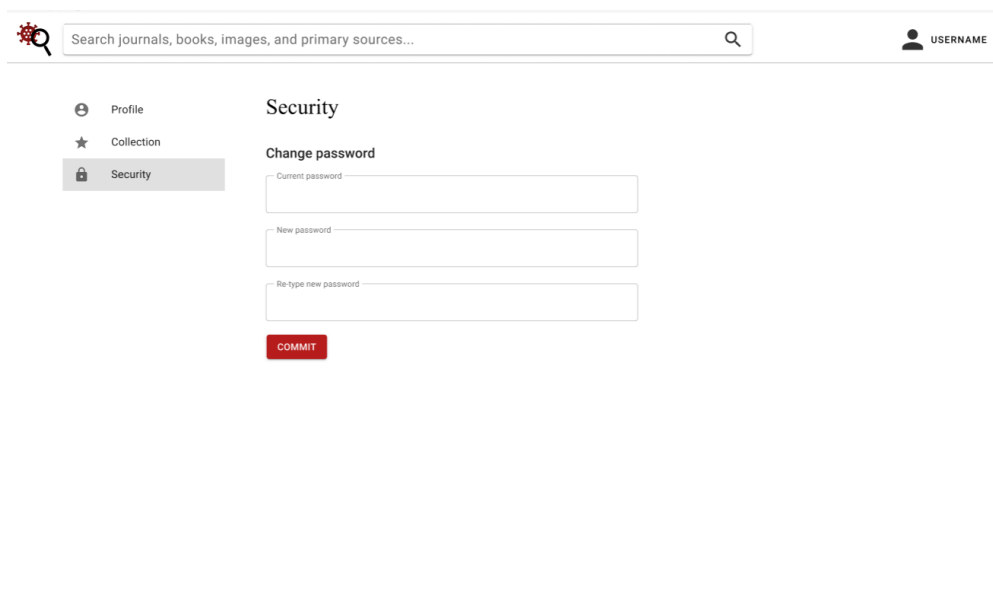


图 5.10 安全设置

6 SQL

在此部分，我们将展示部分 SQL 语句，用于实现系统相关功能。

由于数据量较大，我们考虑在查询时应该使用分页查询，以减少查询时间。我们使用 LIMIT 和 OFFSET 关键字来实现分页查询，例如，LIMIT n OFFSET m 或 LIMIT m, n 表示

从第 $m+1$ 行开始取 n 行数据。为了使报告中的 SQL 语句更加简洁，我们在此省略了分页查询的部分。

在下面的代码中，SQL 语句中的 ? 表示占位符，用于接收用户输入的参数。

6.1 用户相关功能

6.2 搜索相关功能

根据作者姓名（模糊）搜索作者信息：

```
1 SELECT * FROM `author` WHERE `name` LIKE "%?%";
```

根据作者姓名（模糊）搜索其撰写的文献：

```
1 SELECT `author_name`, `id`, `title`, `abstract`, `doi`,  
2     `license`, `publish_time`, `url`  
3 FROM `write` JOIN `article`  
4 ON `article_id` = `id`  
5 WHERE `author_name` LIKE "%?%";
```

根据文献标题（模糊）搜索文献：

```
1 SELECT `id`, `title`, `abstract`, `doi`, `license`  
2 FROM (  
3     SELECT *  
4     FROM `article`  
5     WHERE `title` LIKE "%?%"  
6 ) AS `art`  
7 JOIN `publish` ON `id`=`publish`.`article_id`  
8 JOIN `write` ON `id`=`write`.`article_id`;
```

根据期刊名模糊搜索期刊：

```
1 SELECT *  
2 FROM `journal`  
3 WHERE `journal`.`name` LIKE "%?%";
```

根据期刊名模糊搜索该期刊刊登的文章：

```
1 SELECT *  
2 FROM `publish` JOIN `article` ON `article_id` = `id`  
3 WHERE `journal_name` LIKE "%?%";
```

根据研究方向、研究对象和研究问题搜索文献：

```
1 SELECT `id`, `title`, `abstract`, `doi`, `license`,  
2     `publish_time`, `url`, `journal_name`  
3 FROM (  
4     SELECT *  
5     FROM `article`  
6     WHERE `study_type` LIKE "%?%"  
7     OR `addressed_population` LIKE "%?%"  
8     OR `challenge` LIKE "%?%"  
9     OR `focus` LIKE "%?%"  
10 ) AS `art`  
11 JOIN `publish` ON `id`=`publish`.`article_id`  
12 JOIN `write` ON `id`=`write`.`article_id`;
```

6.3 文献相关功能