浙江大学计算机科学与技术学院

Java 应用技术课程报告

2023—2024 学年秋冬学期

题目	微型学术论文搜索引擎
学号	3210102377
学生姓名	徐文皓
所在专业	软件工程
所在班级	软工 2101

目　录

1 引言

本次开发的是一个微型学术论文搜索引擎，这是一个综合性的题目，可以对Java 语言中的各项功能有更好的理解和使用，通过具体的程序来加深对 Java 语言的掌握，提高自己的编程水平，为以后的工作打下一定的基础。

1．1 设计目的

学术论文搜索引擎是文献检索网站所使用的搜索引擎，是一种垂直搜索引擎，通常可以通过文献标题、作者、年份和摘要等信息检索学术论文。本文使用Java 语言编写一个与其类似的微型学术论文搜索引擎。具体功能如下：

(1) 简易学术论文搜索引擎的主体业务逻辑由网络爬虫、PDF 解析、建立索引和文献检索共 4 个部分组成。为了提升用户友好性，我们增加了一些辅助模块，将普通用户和管理员区分开来，他们具有不同的权限。

(2) 在网络爬虫模块中，使用 JSOUP 实现爬取数据库论文的功能，支持在运行时设置爬虫的参数，将爬取结果暂存到论文类中，等待稍后获得论文类的所有信息后将结果通过文件持久化。

(3) 在 PDF 解析模块中，使用 PDF Box 将从 JSOUP 中爬取得到的 PDF 解析成文本，同时获取 PDF 中关于图表的描述。将这些信息暂存到论文类中。

(4) 在建立索引模块中，使用 Lucene 对文件的主要属性建立索引，对于文章内容、图表信息、年份、会议等项仅建立索引，对于标题、PDF 的 URL 等需要在搜索结果中给出的项同时存储到搜索引擎中。

(5) 文献检索模块中，允许用户选择刚刚建立好的任意一项索引，输入关键词进行检索。搜索引擎显示检索到的文章题目、PDF 的 URL 等信息，允许用户直接在这里打开 PDF 文件。用户也可以根据搜索返回的结果序号查看某一搜索结果的详细内容。

(6) 普通用户和管理员拥有的权限不同。用户可以使用文献检索功能，而管理员可以使用所有功能。搜索引擎保证在使用者只通过搜索引擎对文件更改时数据与文件保持同步。

1．2　设计说明

本程序采用 Java 程序设计语言，在 VSCode 平台下编辑、编译与调试，使用 Maven 进行项目管理。具体程序由本人独立完成。

表 1 各成员分工表

成员名称	完成的主要工作	
	程序设计	**课程报告**
徐文皓	负责整个程序前期的需求分析和整体功能的架构 负责程序的网络爬虫模块 负责程序的 PDF 解析模块 负责程序的建立索引模块 负责程序的文献检索模块 负责其他用以提升用户友好性的模块	本报告的全部内容及格式工作

2 总体设计

2．1 功能模块设计

本程序需实现的主要功能有：

(1) 网络爬虫模块支持对指定 URL 的论文信息与 PDF 文件进行爬取；

(2) PDF 解析模块支持将指定 PDF 的文字内容解析提取成文本；

(3) 建立索引模块支持将论文信息的指定属性和 PDF 文字内容建立索引；

(4) 文献检索模块支持用户对已经建立索引的论文进行检索；

(5) 权限控制使得不同角色可以使用的功能隔离开来；

(6) 比较友好的命令行界面与用户进行交互。

程序的总体功能如图 1 所示：

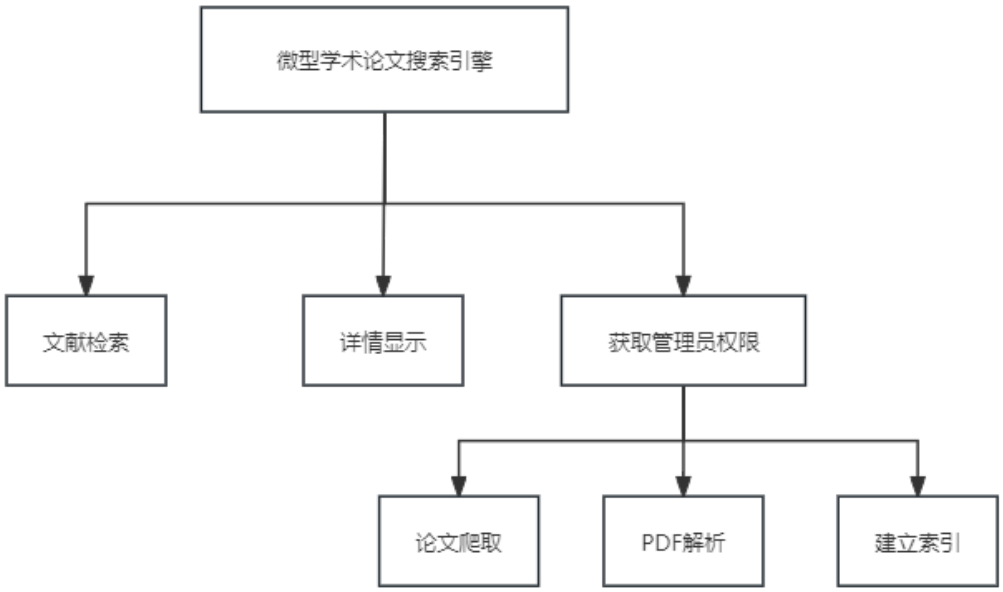

图 1 总体功能图

2. 2　流程图设计

程序总体流程如图 2 所示：

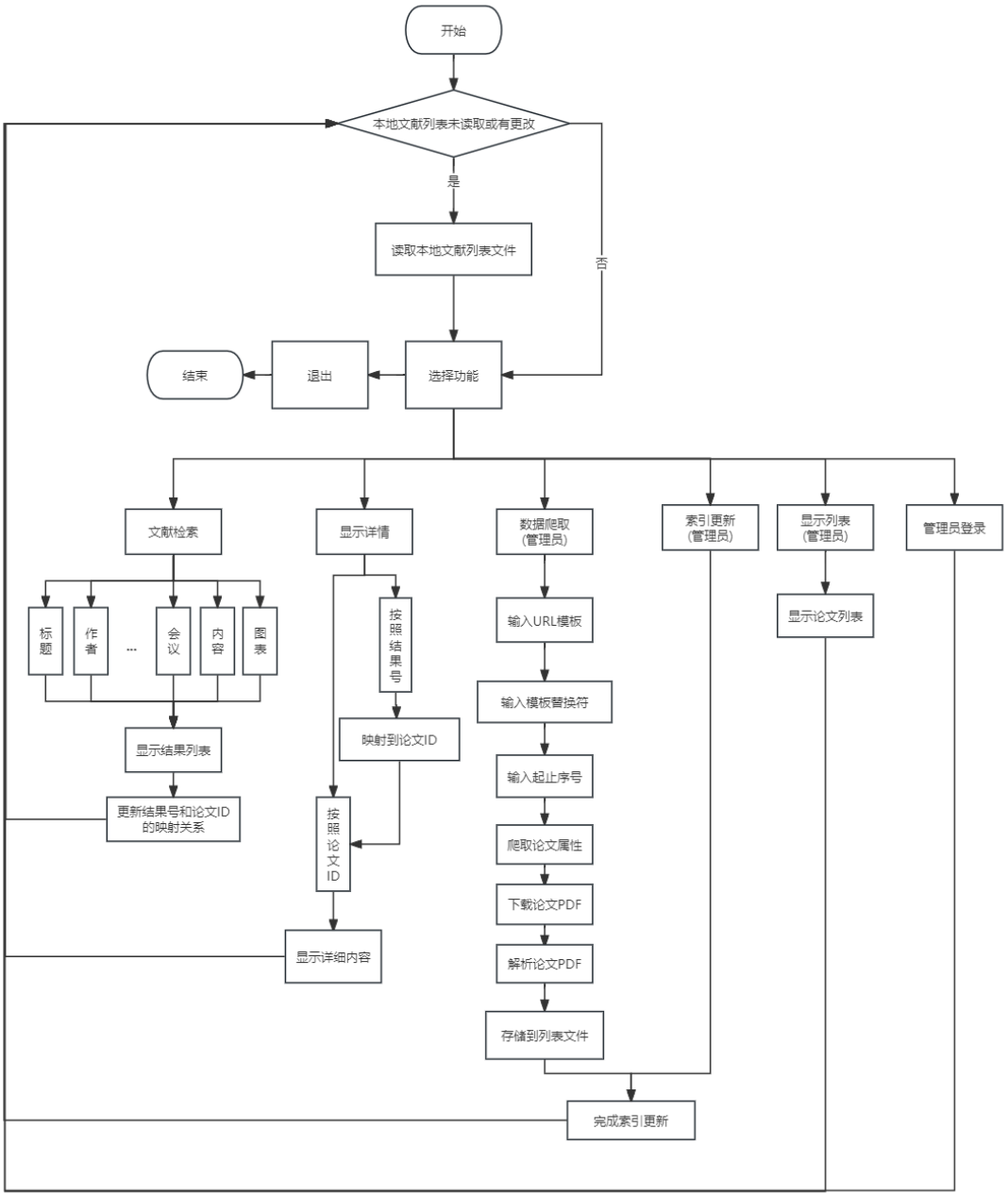

图 2 总体流程图

3　详细设计

3．1　论文类设计

Paper 类是对论文属性的抽象，定义了 title、author、abstractContent 等私有域，外部由其 get、set 方法进行操作，还定义了 toSstring()方法用于展示详情。特别地。PDFURL 域保存了该文献 pdf 的相对路径，在输出之后用户可以直接通过它超链接打开。

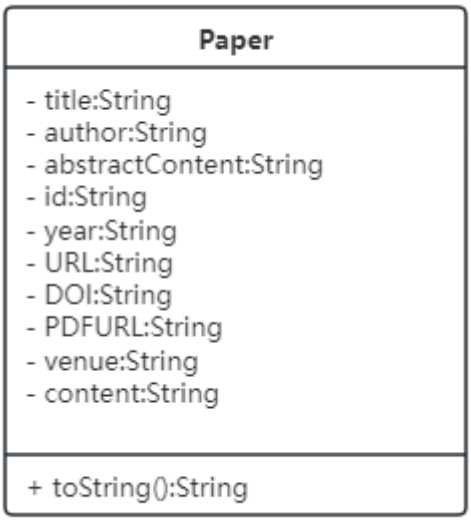

图 3 Paper 类

3．2　网络爬虫模块设计

AclanthologyCrawler 类包含了一个静态的对外接口 fetch()，要求传递待爬取的 URL 模板、模板替换符、起止序号、爬取概率和输出文件流，使用 JSOUP 爬取相关论文。其 UML 图如图 4 所示：

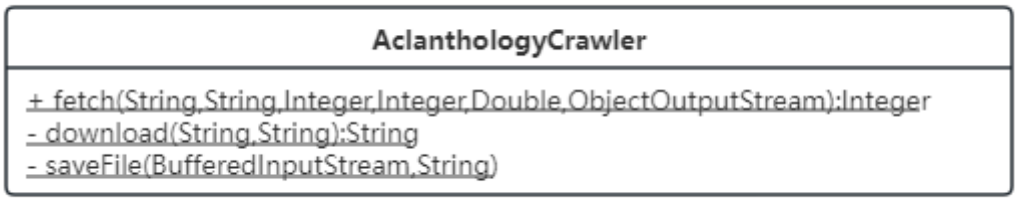

图 4 AclanthologyCrawler 类

以下是 UML 图中有关方法的详细说明：

① fetch(String,String,Integer,Integer,Double,ObjectOutputStream)方法是网络爬虫的主要方法，也是其对外接口。它首先根据传递的 URL 模板、模板替换符、序号起止范围确定爬取的 URL 范围，然后使用 JSOUP 库中的一些方法对这些 URL 进行爬取，并将 PDF 下载到相应文件夹中，对 PDF 进行解析后将 Paper 实例写入到输出流中。文献的爬取分为三个阶段，一是对文献的标题、作者、摘要进行爬取，二是对文献的年份、DOI 等其他信息进行爬取，三是下载并解析 PDF 文件。爬取成功后，实例将被追加到相应文件列表中；

② download(String,String)是用于下载 PDF 的辅助方法，供 fetch()调用。它根据传递的 URL 和名称，前往该 URL 下载文件并且命名为该名称，返回下载后的文件相对路径；

③ saveFile(BufferedInputStream,String)方法是用于保存下载文件的辅助方法，供 download()调用。它得到 download()下载得到的文件流，并将流中的信息保存到给定的路径。

3．3　PDF 解析模块设计

PDFParser 类包含了一个静态的对外接口 parsePDF(String),它用于根据传入的路径解析 PDF 文件，使用 PDFBox 库进行解析，并返回内容解析结果。其 UML 图如图 5 所示：

图 5 PDFParser 类

3．4　建立索引模块设计

Lucene 类是建立索引和实现索引查询的类。它主要实现了两个静态的对外接口，分别用于建立索引和通过索引查询。其 UML 图如图 6 所示：

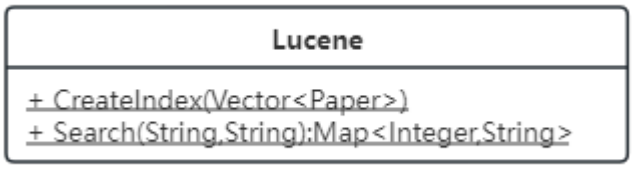

图 6 Lucene 类

以下是 UML 图中有关方法的详细说明：

① CreateIndex(Vector<Paper>)方法用于建立索引。它接收传递的 Paper 类向量，为其中的每个 Paper 类的 ID、标题、作者等一系列信息建立索引，并将 ID、标题和 PDFURL 存储下来用于搜索结果的显示；

② Search(String,String)是用于根据索引搜索的方法。它接收搜索所基于的索引和搜索关键词，进行搜索。在搜索到结果后，它将为每一条搜索结果赋结果号，并打印其标题和 PDFURL。用户可以直接点击 PDFURL 查看 PDF，或者根据结果号查询该论文的详细信息。这个方法返回了一个 Map，其键为结果号，值为论文 ID，也就建立了这样的映射，方便之后进行详情展示。

3．5　文献检索模块和界面设计

Display 类是用于和用户交互的前端界面。它维护了 Scanner 类读取用户输入，并且提供了启动交互界面的 Welcome()这一静态对外接口，由其调用菜单显示方法，由菜单调用搜索方法。其 UML 图如图 7 所示：

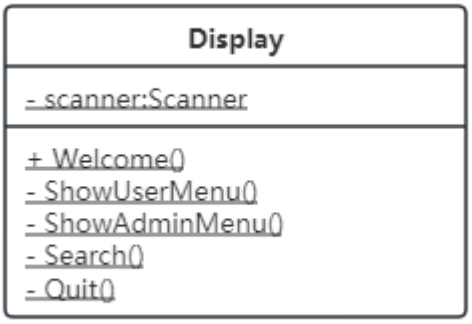

图 7 Lucene 类

以下是 UML 图中有关方法的详细说明：

① Welcome()方法用于启动搜索引擎交互界面。它向用户打印欢迎信息并调用用户菜单显示方法 ShowUserMenu()进行后续交互，将欢迎单独实现的目的是，在其中处理了 NoSuchElementException 异常，使得用户在通过键盘中断退出程序时能够得到合理的处理；

② ShowUserMenu()是用于展示用户菜单的方法。它提示用户选择功能，提供了查询论文、根据结果号查看论文详情和根据论文 ID 查看论文详情功能，以及管理员登录的入口，引导用户和搜索引擎进行交互；

③ ShowAdminMenu()是用于展示管理员菜单的方法。当用户在用户菜单进行管理员登录成果后调用此方法。相对与用户，管理员可以使用论文列表展示、论文爬取和主动更新索引功能，引导管理员和搜索引擎进行交互；

④ Search()是用于引导用户发起搜索的方法。它提示用户选择搜索所依据的索引并输入关键词进行搜索。

以上提到的功能会在稍后的 Engine 类中详述。

3．6　搜索引擎服务端设计

Engine 类用于为前端提供服务。它维护了搜索引擎的状态，并为前端界面中提到的可选择的所有功能提供服务。其一切属性和方法均为静态。其 UML 图如图 8 所示：

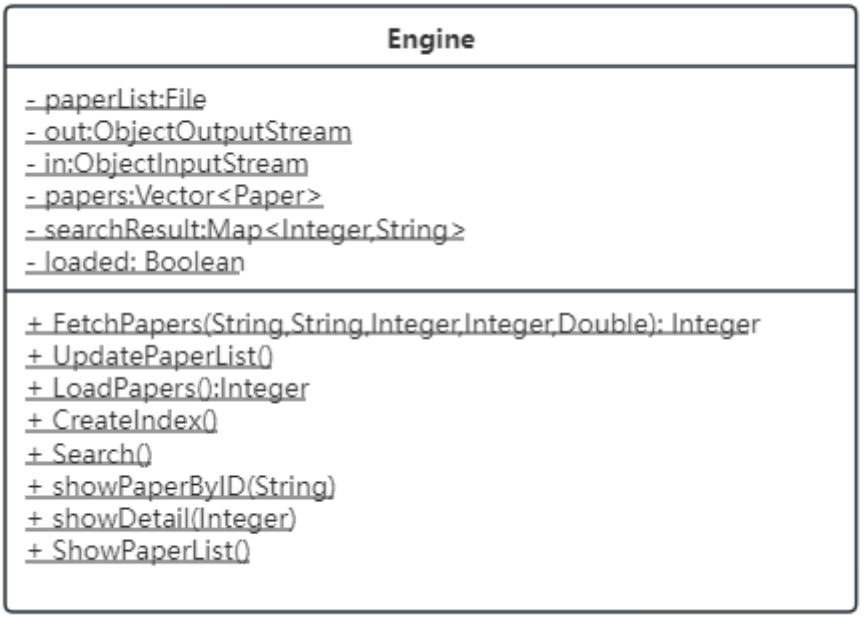

图 8 Engine 类

以下是 UML 图中有关属性和方法的详细说明：

（1）成员变量

① paperList 是用于存储论文属性的文件，我们使用这一文件实现论文数据的持久化；

② out 是用于更新 paperList 的对象输出流；

③ in 是用于从 paperList 中读取论文的对象输入流；

④ papers 是从 paperList 中读取得到的 Paper 向量，存储了本地所有的论文属性信息；

⑤ searchResult 是上一次搜索中搜索结果同论文 ID 的映射；

⑥ loaded 记录了 paperList 的加载情况。当从 paperList 向 papers 中完成加载时，它被设置为 true；当 paperList 发生了更新时，它被设置为 false。对于那些需要读取 papers 的方法，它们要先检查 loaded，如果 loaded 为 false，

说明 papers 与存在本地的 paperList 内容未同步，需要重新读取后再进行后续功能。

（2）方法

① FetchPapers(String,String,Integer,Integer,Double)方法为管理员的论文爬取功能提供服务，它接收 URL 模板、模板替换符、起止序号和爬取概率，调用网络爬虫模块爬取论文信息,爬取结束后自动调用CreateIndex()建立索引；

② LoadPapers()方法从 paperList 中同步数据，并赋值给 papers 变量；

③ UpdatePaperList()方法用于检索 paperList 中的 PDF 文件是否还存在，它将删去 PDF 不存在的论文信息——也就是说，我们可以通过删除本地的 PDF 文件来间接地删除搜索引擎中存储的论文信息；

④ CreateIndex()方法根据 papers 建立索引；

⑤ Search()方法调用搜索引擎的搜索功能，同时更新 searchResult；

⑥ showPaperByID(String)方法接收需要显示详情内容的论文 ID,在 papers 中根据论文 ID 检索论文并输出详细信息；

⑦ showDetail(Integer)方法先把接收到的结果号映射为论文 ID，再调用 showPaperByID(String)输出该论文的详细信息；

⑧ ShowPaperList()方法显示所有 papers 的论文 ID、标题和 PDFURL。

3．7　对象输入输出流设计

为了实现能够在 paperList 中追加 Paper，而不是每次都覆盖掉它，我们实现了 MyObjectInputStream 和 MyObjectOutputStream 类，重写了关于头部 I/O 的方法。其 UML 图如图 9 所示：

图 9 MyObjectInputStream 和 MyObjectOutputStream 类

4 测试与运行

4．1 程序测试

在程序代码基本完成后，经过不断的调试与修改，最后测试本次所设计的微型学术论文搜索引擎能够正常运行，没有出现明显的错误和漏洞，但是在一些细节方面仍然需要完善，比如网络爬虫的多线程化等。总的来说本次设计在功能上已经基本达到要求，其他细节方面有待以后完善。

4．2 程序运行

程序运行主界面及用户菜单如图 10 所示：

图 10 程序主界面及用户菜单

管理员菜单如图 11 所示：

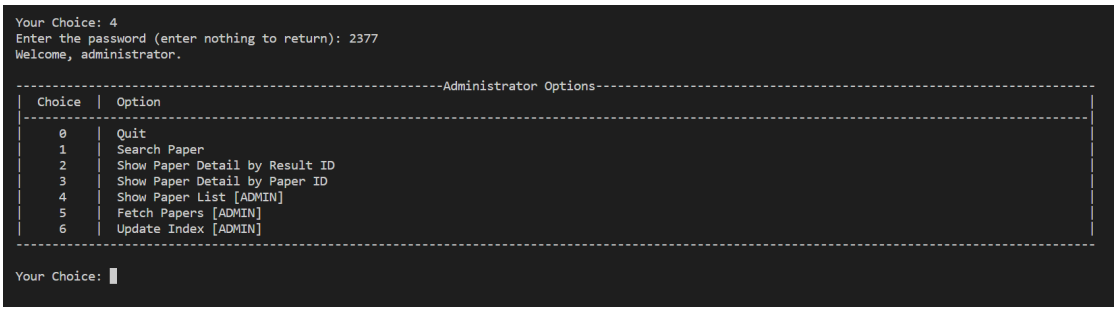

图 11 管理员菜单

论文爬取功能显示情况如图 12 所示：

=

```
Your Choice: 5
Enter the URL Template (enter nothing to return): https://aclanthology.org/P15-*/
Enter the replacer (enter nothing to return): *
Enter the start number (enter nothing to return): 1001
Enter the end number (enter nothing to return): 1010
Enter the fetch probability (enter nothing to return): 1

Start fetching...

Succeed to fetch paper[P15-1001].
Succeed to fetch paper[P15-1002].
Succeed to fetch paper[P15-1003].
Succeed to fetch paper[P15-1004].
Succeed to fetch paper[P15-1005].
Succeed to fetch paper[P15-1006].
Succeed to fetch paper[P15-1007].
Succeed to fetch paper[P15-1008].
Succeed to fetch paper[P15-1009].
Succeed to fetch 9 paper(s).
Succeed to create index.
```

图 12 论文爬取功能

论文列表展示功能显示情况如图 13 所示：

```
Your Choice: 4
Result
| Paper ID | Title | PDF |
| 2020.acl-main.1 | Learning to Understand Child-directed and Adult-directed Speech | ./documents/pdf/2020.acl-main.1.pdf |
| 2020.acl-main.2 | Predicting Depression in Screening Interviews from Latent Categorization... | ./documents/pdf/2020.acl-main.2.pdf |
| 2020.acl-main.3 | Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling | ./documents/pdf/2020.acl-main.3.pdf |
| 2020.acl-main.4 | Designing Precise and Robust Dialogue Response Evaluators | ./documents/pdf/2020.acl-main.4.pdf |
| 2020.acl-main.5 | Dialogue State Tracking with Explicit Slot Connection Modeling | ./documents/pdf/2020.acl-main.5.pdf |
| 2020.acl-main.6 | Generating Informative Conversational Response using Recurrent Knowledge... | ./documents/pdf/2020.acl-main.6.pdf |
| 2020.acl-main.7 | Guiding Variational Response Generator to Exploit Persona | ./documents/pdf/2020.acl-main.7.pdf |
| 2020.acl-main.8 | Large Scale Multi-Actor Generative Dialog Modeling | ./documents/pdf/2020.acl-main.8.pdf |
| 2020.acl-main.9 | PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable | ./documents/pdf/2020.acl-main.9.pdf |
| 2020.acl-main.10 | Slot-consistent NLGfor Task-oriented Dialogue Systems with Iterative Rec... | ./documents/pdf/2020.acl-main.10.pdf |
| 2020.acl-main.11 | Span-ConveRT: Few-shot Span Extraction for Dialog with Pretrained Conver... | ./documents/pdf/2020.acl-main.11.pdf |
| 2020.acl-main.12 | Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialo... | ./documents/pdf/2020.acl-main.12.pdf |
| 2020.acl-main.13 | A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle | ./documents/pdf/2020.acl-main.13.pdf |
| 2020.acl-main.14 | TransS-Driven Joint Learning Architecture for Implicit Discourse Relatio... | ./documents/pdf/2020.acl-main.14.pdf |
| 2020.acl-main.15 | A Study of Non-autoregressive Model for Sequence Generation | ./documents/pdf/2020.acl-main.15.pdf |
| 2020.acl-main.16 | Cross-modal Language Generation using Pivot Stabilization for Web-scale ... | ./documents/pdf/2020.acl-main.16.pdf |
| 2020.acl-main.17 | Fact-based Text Editing | ./documents/pdf/2020.acl-main.17.pdf |
| 2020.acl-main.18 | Few-Shot NLGwith Pre-Trained Language Model | ./documents/pdf/2020.acl-main.18.pdf |
| 2020.acl-main.19 | Fluent Response Generation for Conversational Question Answering | ./documents/pdf/2020.acl-main.19.pdf |
| P15-1001 | On Using Very Large Target Vocabulary for Neural Machine Translation | ./documents/pdf/P15-1001.pdf |
| P15-1002 | Addressing the Rare Word Problem in Neural Machine Translation | ./documents/pdf/P15-1002.pdf |
| P15-1003 | Encoding Source Language with Convolutional Neural Network for Machine T... | ./documents/pdf/P15-1003.pdf |
| P15-1004 | Statistical Machine Translation Features with Multitask Tensor Networks | ./documents/pdf/P15-1004.pdf |
| P15-1005 | Describing Images using Inferred Visual Dependency Representations | ./documents/pdf/P15-1005.pdf |
| P15-1006 | Text to 3DScene Generation with Rich Lexical Grounding | ./documents/pdf/P15-1006.pdf |
| P15-1007 | MultiGranCNN: An Architecture for General Matching of Text Chunks on Mul... | ./documents/pdf/P15-1007.pdf |
| P15-1008 | Weakly Supervised Models of Aspect-Sentiment for Online Course Discussio... | ./documents/pdf/P15-1008.pdf |
| P15-1009 | Semantically Smooth Knowledge Graph Embedding | ./documents/pdf/P15-1009.pdf |
```

图 13 论文列表展示功能

搜索功能如图 14 所示：

```
Your Choice: 1
Search By
| Choice | Others | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Option | Quit | I D | Title | Author | Abstract | Year | Venue | DOI | Content/Figure |
Your Choice: 2
Search Keywords: study
Here is the result.

Result
| ResId | Title | PDF |
| 0 | A Study of Non-autoregressive Model for Sequence Generation | ./documents/pdf/2020.acl-main.15.pdf |
| 1 | Few-Shot NLGwith Pre-Trained Language Model | ./documents/pdf/2020.acl-main.18.pdf |
| 2 | Learning to Understand Child-directed and Adult-directed Speech | ./documents/pdf/2020.acl-main.1.pdf |
| 3 | Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling | ./documents/pdf/2020.acl-main.3.pdf |
| 4 | Large Scale Multi-Actor Generative Dialog Modeling | ./documents/pdf/2020.acl-main.8.pdf |
| 5 | Designing Precise and Robust Dialogue Response Evaluators | ./documents/pdf/2020.acl-main.4.pdf |
| 6 | Slot-consistent NLGfor Task-oriented Dialogue Systems with Iterative Rectification Network | ./documents/pdf/2020.acl-main.10.pdf |
| 7 | Generating Informative Conversational Response using Recurrent Knowledge-Interaction a... | ./documents/pdf/2020.acl-main.6.pdf |
| 8 | Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage | ./documents/pdf/2020.acl-main.16.pdf |
| 9 | Encoding Source Language with Convolutional Neural Network for Machine Translation | ./documents/pdf/P15-1003.pdf |
| 10 | MultiGranCNN: An Architecture for General Matching of Text Chunks on Multiple Levels o... | ./documents/pdf/P15-1007.pdf |
| 11 | PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable | ./documents/pdf/2020.acl-main.9.pdf |
| 12 | Predicting Depression in Screening Interviews from Latent Categorization of Interview ... | ./documents/pdf/2020.acl-main.2.pdf |
| 13 | Guiding Variational Response Generator to Exploit Persona | ./documents/pdf/2020.acl-main.7.pdf |
| 14 | Dialogue State Tracking with Explicit Slot Connection Modeling | ./documents/pdf/2020.acl-main.5.pdf |
| 15 | A Complete Shift-Reduce Chinese Discourse Parser with Robust Dynamic Oracle | ./documents/pdf/2020.acl-main.13.pdf |
| 16 | Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums | ./documents/pdf/P15-1008.pdf |
| 17 | Fluent Response Generation for Conversational Question Answering | ./documents/pdf/2020.acl-main.19.pdf |
```

图 14 搜索功能

论文详情展示功能如图 15 所示：

```
Your Choice: 2
Enter the result ID (enter nothing to return): 0
Here is its detail information.

-------------------------------------------------------------------------Detail-------------------------------------------------------------------------
| ID       |  2020.acl-main.15
| Title    |  A Study of Non-autoregressive Model for Sequence Generation
| Author   |  Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, Tie-Yan Liu
| Venue    |  ACL
| Year     |  2020
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------
| Abstract |  Non-autoregressive (NAR) models generate all the tokens of a sequence in parallel, resulting in faster generation speed compared to
|          |  their autoregressive (AR) counterparts but at the cost of lower accuracy. Different techniques including knowledge distillation and
|          |  source-target alignment have been proposed to bridge the gap between AR and NAR models in various tasks such as neural machine tran
|          |  lation (NMT), automatic speech recognition (ASR), and text to speech (TTS). With the help of those techniques, NAR models can catch
|          |  up with the accuracy of AR models in some tasks but not in some others. In this work, we conduct a study to understand the difficul
|          |  y of NAR sequence generation and try to answer: (1) Why NAR models can catch up with AR models in some tasks but not all? (2) Why t
|          |  chniques like knowledge distillation and source-target alignment can help NAR models. Since the main difference between AR and NAR
|          |  odels is that NAR models do not use dependency among target tokens while AR models do, intuitively the difficulty of NAR sequence g
|          |  neration heavily depends on the strongness of dependency among target tokens. To quantify such dependency, we propose an analysis m
|          |  del called CoMMA to characterize the difficulty of different NAR sequence generation tasks. We have several interesting findings: 1
|          |   Among the NMT, ASR and TTS tasks, ASR has the most target-token dependency while TTS has the least. 2) Knowledge distillation redu
|          |  es the target-token dependency in target sequence and thus improves the accuracy of NAR models. 3) Source-target alignment constrai
|          |  t encourages dependency of a target token on source tokens and thus eases the training of NAR models.
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------
| DOI      |  10.18653/v1/2020.acl-main.15
| URL      |  https://aclanthology.org/2020.acl-main.15
| PDFURL   |  ./documents/pdf/2020.acl-main.15.pdf
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------
```

图 15 论文详情展示功能

PDF 的 URL 可以直接点击打开，如图 16 所示：

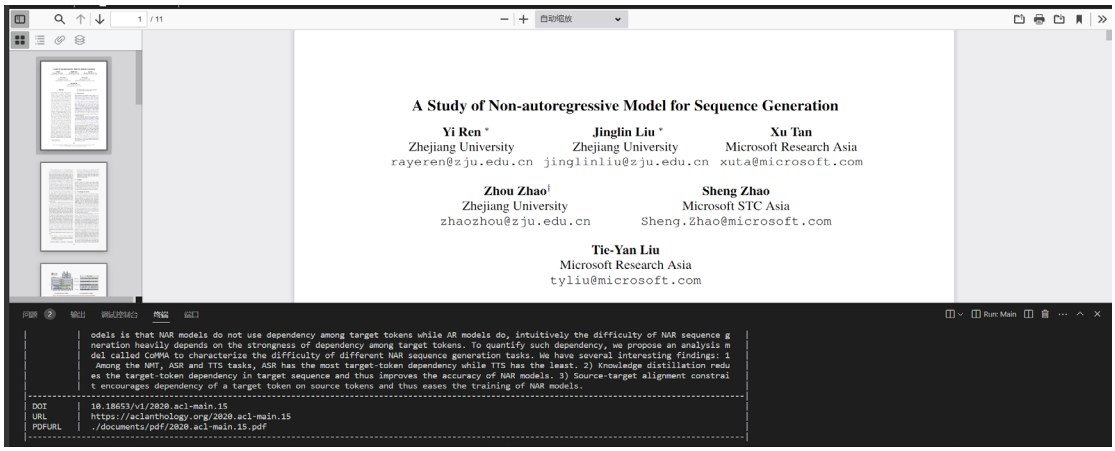

图 16 PDF 跳转功能

论文详情 URL 可以直接点击打开，跳转到原网址，如图 17 所示：

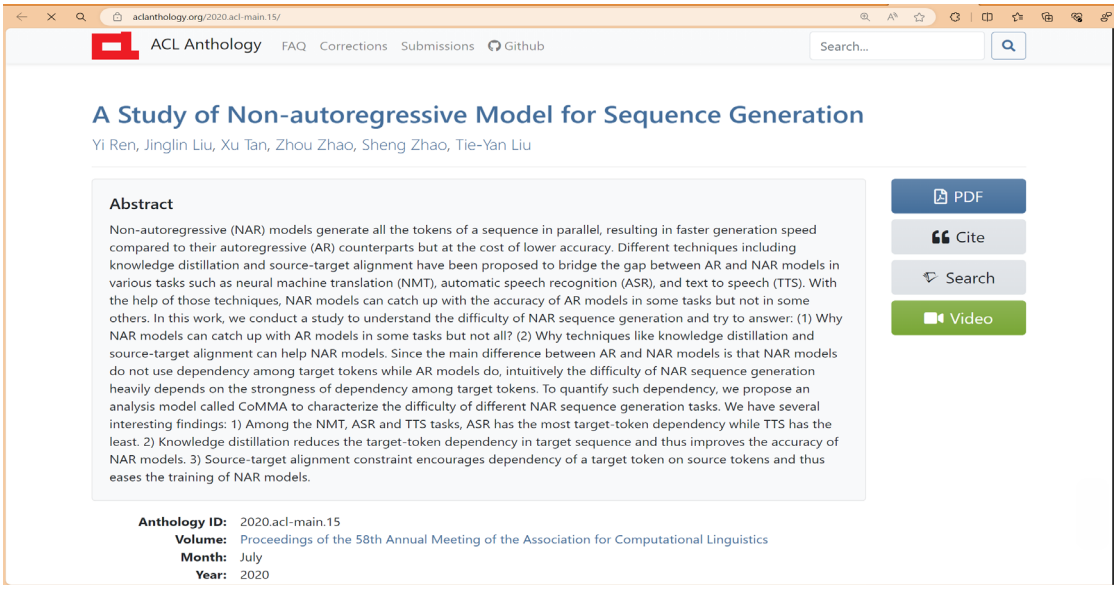

图 17 URL 跳转功能

可以发现，本引擎对图表的搜索也能够达到预期效果，如图 18 所示：

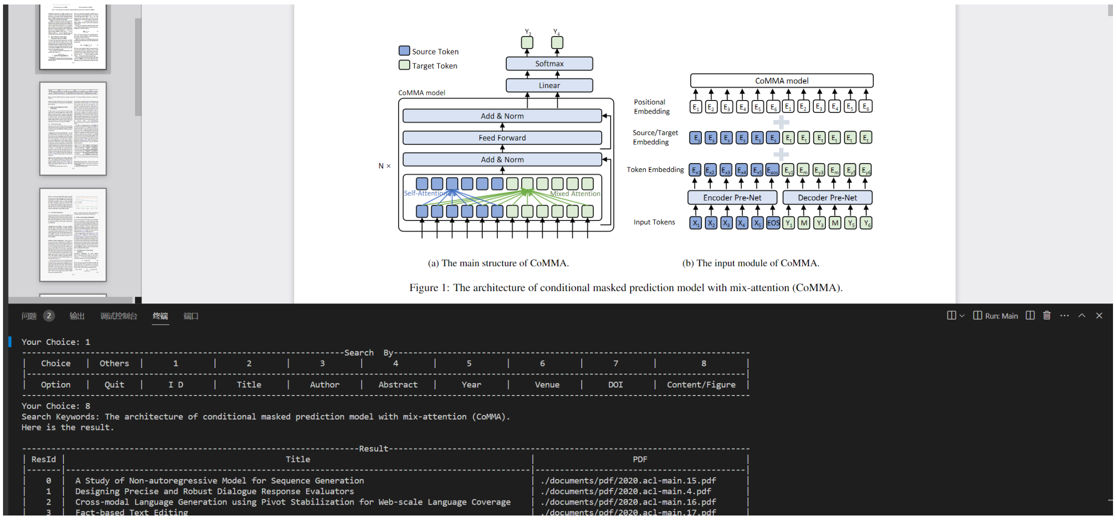

图 18 图表搜索情况

5 总结

整个微型学术论文搜索引擎项目聚焦在 Java 的 IO、异常处理和使用第三方库等知识点，项目要求结构清晰，非常适合使用面向对象思想进行设计。本项目主要还是客户端 Display 类和服务端 Engine 类之间的交互，采用了自底向上的思路，先实现服务，再调用它们。

在本次项目的编写过程中，我见识到了 Java 语言本身的严谨性，提高了自己的编程能力。这是我第一次使用爬虫和正则表达式，经过学习后成功爬取到内容令我十分兴奋。同时，在本次项目中我用到了像 JSOUP、Lucene 等非常强大的第三方库，这也让我感慨于计算机领域的开源、互助和这种“生态”精神，这是因为有了这些服务的提供，我们利用它们实现自己想要实现的东西才成为了可能——否则光是“造轮子”就会使人望而却步了。

6 数据格式说明

本项目中，论文属性会先存储到 Paper 中，之后同步到./paperList 文件中进行持久化。

文献索引文件储存在./index 中，文献 PDF 文件存储在./doucuments/pdf 中。

打开项目时，应将工作目录切换至 homework4 目录，即与 paperList 所在的目录。

参考文献

[1]《Java 语言程序设计（进阶篇）（英文版·第 10 版）》[美]梁勇（Y.DanielLiang）著，2017 年，机械工业出版社.

[2]《Learn Java 12 Programming: Astep by step guide to learning essential concepts in Java SE 10,11,and 12 (English Edition)》 [美]NickSamoylov 著，2019 年，Packt Publishing 出版社.