Web信息处理与应用

实验1：基于DBWorld的搜索引擎实现

实验内容

- 本实验是开发出面向DBWorld实时信息的灵活、多功能的搜索引擎

  本次实验的内容主要包括以下几部分:

  1. 针对研究人员对DBWorld的信息搜索需求，设计了一个DBWorld搜索引擎，给出总体设计

  2. 针对DBWorld原始数据的特点，对DBWolrd进行信息抽取，包括DBWorld网页中的地点、时间、主题等信息

  3. 实现了一个基于WEB的DBWolrd搜索引擎

DBWorld 介绍

❏ https://research.cs.wisc.edu/dbworld/browse.html

Sent	Message Type	From	Subject	Deadline	Web Page
28-Sep-2018	conf. ann.	Licong Cui	The 1st International Workshop on Quality Assurance of Biological and Biomedical Ontologies	7-Oct-2018	web page
28-Sep-2018	conf. ann.	David Silva	BigDat 2019: early registration October 15	15-Oct-2018	web page
28-Sep-2018	journal ann.	Marcelo Armentano	Special Issue on Knowledge Discovery and User Modelling for Smart Cities - Call for Papers		web page
28-Sep-2018	conf. ann.	Elvira Popescu	CFP (extended): International Conference on Smart Learning Environments (ICSLE 2019)	19-Oct-2018	
28-Sep-2018	conf. ann.	Nabil El Ioini	12th WISTP International Conference on Information Security Theory and Prac...	1-Oct-2018	web page
28-Sep-2018	journal ann.	Gabriele D'Angelo	CFP: (EXTENDED) "Cryptocurrencies and Blockchains for Distributed Systems"		web page
28-Sep-2018	journal ann.	Bernhard Peischl Dr.	CfP: SQJ, SI on Quality of Adaptive Software Systems		web page
28-Sep-2018	job ann.	Elisa Marengo	Research Associate Position at the Free Univ. of Bozen-Bolzano, Italy	5-Nov-2018	web page
28-Sep-2018	job ann.	Grégory Smits	18 months Post-Doc. position (IRISA laboratory - France)	15-Oct-2018	web page
28-Sep-2018	journal ann.	JCSE Editorial Office	Newly published papers		web page
27-Sep-2018	conf. ann.	Mohamed Anis Bach Tobji	Call for paper of ICDEc 2019	22-Dec-2018	web page
27-Sep-2018	conf. ann.	Feng Chen	Call for Papers: IEEE BPOD 2018 collocated with IEEE BigData 2018	10-Oct-2018	web page
27-Sep-2018	job ann.	Jennifer Schmidt	Tenure-Track Faculty Positions in Computer Science	15-Dec-2018	web page
27-Sep-2018	conf. ann.	Ansar Yasar	cfp: EDI40-2019 (NextGen IoT)	06-Dec-2018	web page
27-Sep-2018	conf. ann.	Ansar Yasar	CFP: ANT - 2019	06-Dec-2018	web page
27-Sep-2018	conf. ann.	Natália Reis	CFP - 12th IADIS International Conference on Information Systems 2019	9-Nov-2018	web page
27-Sep-2018	journal ann.	Bamshad Mobasher	CFP: UMUAI Special Issue on Session-Based and Sequential Recommender Systems		web page
27-Sep-2018	journal CFP	Olga C. Santos	CfP IEEE Special Issue: Data Capture and Analysis to Support Learning Engagement - IEEE-TLT	1-Oct-2018	web page
27-Sep-2018	conf. ann.	Katerina Doka	Deadline extension: 1st International Workshop on Next Generation Clouds for Extreme Data Analytics	8-Oct-2018	
27-Sep-2018	conf. ann.	Adrian Paschke	CfP: Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS 2018)	1-Oct-2018	web page
27-Sep-2018	conf. ann.	George Papadopoulos	The 27th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2019): Student Grants	1-Nov-2018	web page
27-Sep-2018	conf. ann.	Carl Pontillas	Last CFP: 7th International Conference on Informatics and Applications (ICIA2018) - JAPAN	5-Oct-2018	web page
27-Sep-2018	book ann.	Bart Baesens Prof. dr.	NEW BOOK: Principles of Database Management		web page
27-Sep-2018	conf. ann.	Jalou Batilong	Last Call for Papers: 4th International Conference on Electronics and Software Science - JAPAN	5-Oct-2018	web page
27-Sep-2018	journal CFP	Pradip Sharma Dr.	Special Issue on Emerging ICT Applications and Service - Big Data, IoT, and Cloud Computing	28-Feb-2019	web page
27-Sep-2018	book CFC	Alessandro Fiori PhD	Call for chapter proposals: Trends and Applications of Text Summarization Techniques	15-Oct-2018	
26-Sep-2018	journal ann.	Harindu Korala	CFP	Special Issue: Smart IoT Sensing (MDPI Sensors Open Access Journal)		web page

点击查看网页内容
(需要自己爬取)

DBWorld browse

网页内容

时间地点

```
--------------------
ICDEc 2019 - CALL FOR PAPERS
--------------------

The fourth annual International Conference on Digital Economy (ICDEc) will take place from April 15th to 18th, 2019 in the International University of Beirut, Lebanon.

The ICDEc conference is an opportunity to bring together researchers and experts in emerging technologies from both public and private sectors to discuss issues related to the development, implementation, use and management of emerging technologies and the move to digital economy.

The conference is an occasion to link researchers from computer science, business computing and information system areas to business fields such as marketing, finance, human resources, management etc.

It is an opportunity to explore ideas, exchange thoughts and implement research projects between this diversity of fields with a focus on application contexts like e-commerce, e-business, e-heath, e-learning, etc.

Accepted and presented papers will be included in the conference proceedings and submitted for inculsion in the Springer LNBIP series.

----------
GENERAL CHAIR
----------
Farid Abdallah, International University Of Beirut, Lebanon

----------
PC CO-CHAIRS
----------
Sehl Mellouli, Université Laval, Canada
Deny Belisle, Université de Sherbrooke, Canada

-----
TRACKS
-----
Submissions can consist of theoretical or applied research in a large range of tracks and topics including, but not limited to:

EMERGING TECHNOLOGIES

- Web Intelligence Technologies
- Big Data and Data Science
- Knowledge Discovery and Management
- Intelligent Agent Systems
- Web Security
- Uncertainty Management

BUSINESS INNOVATION

- Digital transformation processes in the private and public sectors
- E-Governance
- E-Commerce
- Digital Marketing
- E-Finance
- E-Learning and Cloud Education
- E-Health
- Balance of Trade and Big Data
- Utilizing Big Data in Sustainable Development

--------
SUBMISSION
--------
- Full papers (Up to 14 pages)
- Short papers (Up to 6 pages)
- Demo/case study papers (Up to 6 pages)
- Posters (with extended abstract)

IMPORTANT DATES
----------
Deadline of paper submission: December 22nd, 2018
Notification of acceptance and reviewing report: January 21st, 2019
Deadline of final version submission: February 10th, 2019
Deadline for early birds registration: March 15th, 2019
Deadline for late registration: April 3rd, 2019
Conference date: April 15th-18th, 2019
```

主题

时间

参考网站

- http://dbis-group.uni-muenster.de/dbms/templates/conferences/conferences.php

只支持对会议名称的搜索

需求分析

- 功能需求
  - 提供定时抓取，定时更新DBworld信息的功能，以确保数据的及时、准确、有效性。
  - 提供多种搜索选项，包括提交截止日期，会议起止日期，地点，主题等。
  - 提供一个搜索和显示界面。
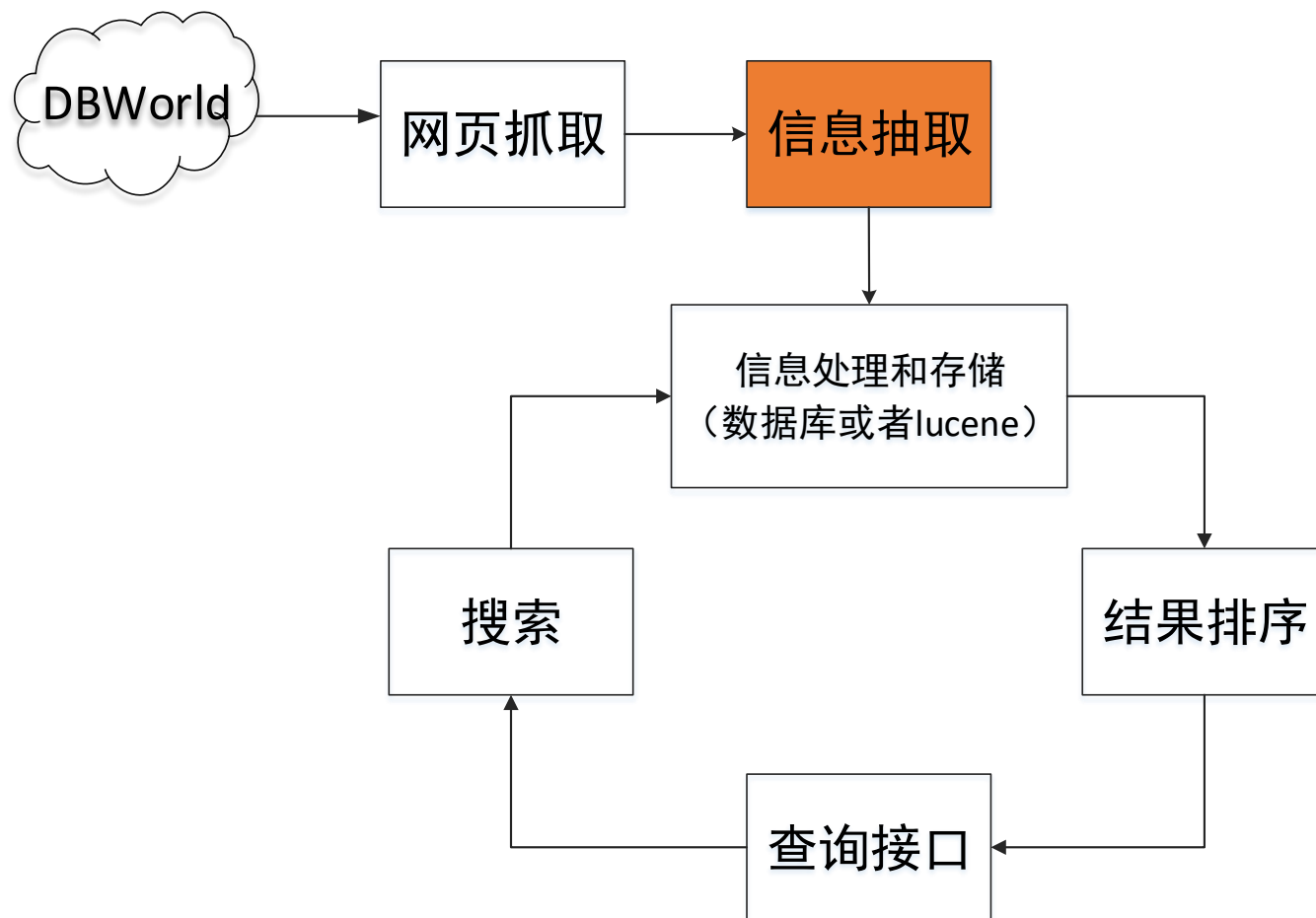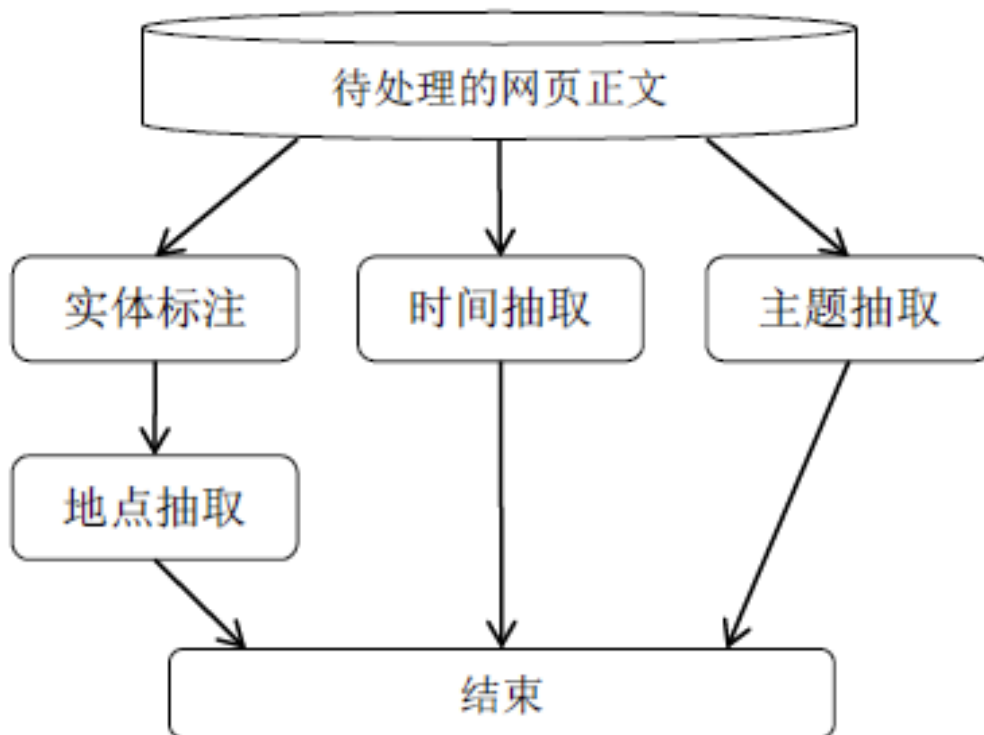- 性能需求
  - 需要保证DBWorld信息的查准率。
  - 需要保证搜索引擎的快速响应。

功能流程

信息抽取

简单查询界面实现

- 使用jsp实现简单的搜索界面

```
<html>
<head>
<title>First Page</title>
</head>
<body>
<H3>Today is:<%= new java.util.Date() %>
</H3>
</body>
</html>
```

JSP可以很简单的嵌入Java代码 <% Java代码 %>，在Java代码中可以使用out.write()直接写出到HTML页面中

Tomcat服务器

- Tomcat 服务器是一个免费的开放源代码的Web 应用服务器，具有处理HTML页面的功能，另外它还是一个Servlet和JSP容器。

  - 使用Tomcat需要系统安装了jdk
  - ftp中xampp里包含了tomcat，或者直接解压apache-tomcat-7.0.42-windows-x86.zip，使用startup.bat脚本进行启动
  - 程序目录放在webapps下

简单查询界面实现

本实验要点

- 学会使用网络爬虫爬取网页。

- 对网页进行分析，抽取网页中的时间，地点，主题。

- 学习如何对搜索结果进行ranking

- 了解基本JSP语法，实现一个简单的搜索界面

实验扩展

- 会议重要性
  - 根据会议重要性(参考ccf推荐会议)排序
  - https://www.ccf.org.cn/xspj/gyml/
- 用户相关
  - 实现用户登录注册功能
  - 根据用户信息，搜索结果优先返回用户领域相关文章
- 相关词搜索
- 结果多页显示
- 等等

- 大家可自由发挥，给予适当的加分。

实验相关资料

- ftp://114.214.167.60
  - 用户名：webinfo 密码：ustc2018

- 推荐使用java

- IDE建议使用Eclipse的J2EE版本，可以直接部署在 Tomcat里运行，调试方便。

- Stanford CoreNLP
  - https://stanfordnlp.github.io/CoreNLP/

实验考核

- 程序需要现场运行检查

- 实验需要提交的文档
  - 源程序打包。

  - 程序设计文档（包括程序运行截图），具体见ftp上程序设计文档示例。

有疑问？

- 联系我
  - mulin@mail.ustc.edu.cn 穆林 (科技楼1610)
  - jr1111@mail.ustc.edu.cn 周靖人(科技楼1610)

- 实验课
  - 每周六晚上上机，从10.13开始，大家可关注课程主页。