

对集成检索与全文阅读 的探索

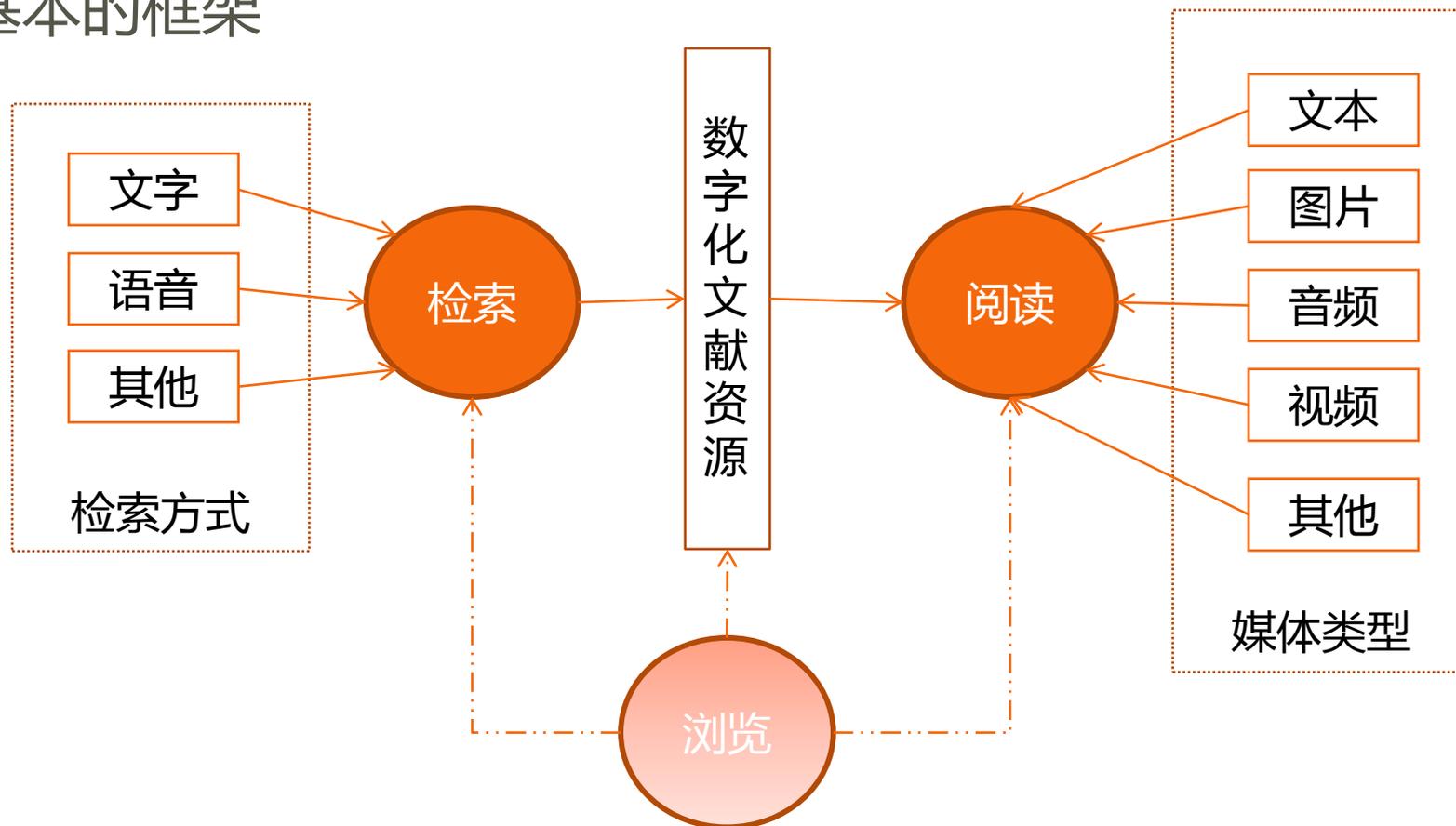


顾宏

2013年12月

读者对数字图书馆的使用

基本的框架



对检索（Find）的认识

基本的发​​展轨迹



一些浅薄的认识

- 检索：retrieval，讲求精准度，主要应用于元数据级，一维方式
- 搜索：search，讲求广度，主要应用于内容级，二维方式
- 发现：discovery，讲求关联，源于检索而高于搜索，准三维方式

图书馆的弱势与自救

∞ 现代环境下，数字图书馆在与商业搜索引擎的竞争中明显处于弱势，用户群体不断被蚕食。

- 精于检索/索引系统 (Retrieval System/Index System)
- 过于囿于精度/准度的要求，在广度和关联上下的功夫不足
- 智能化水平相对较低，对于用户浏览的需求简单化处理

∞ 解决之道

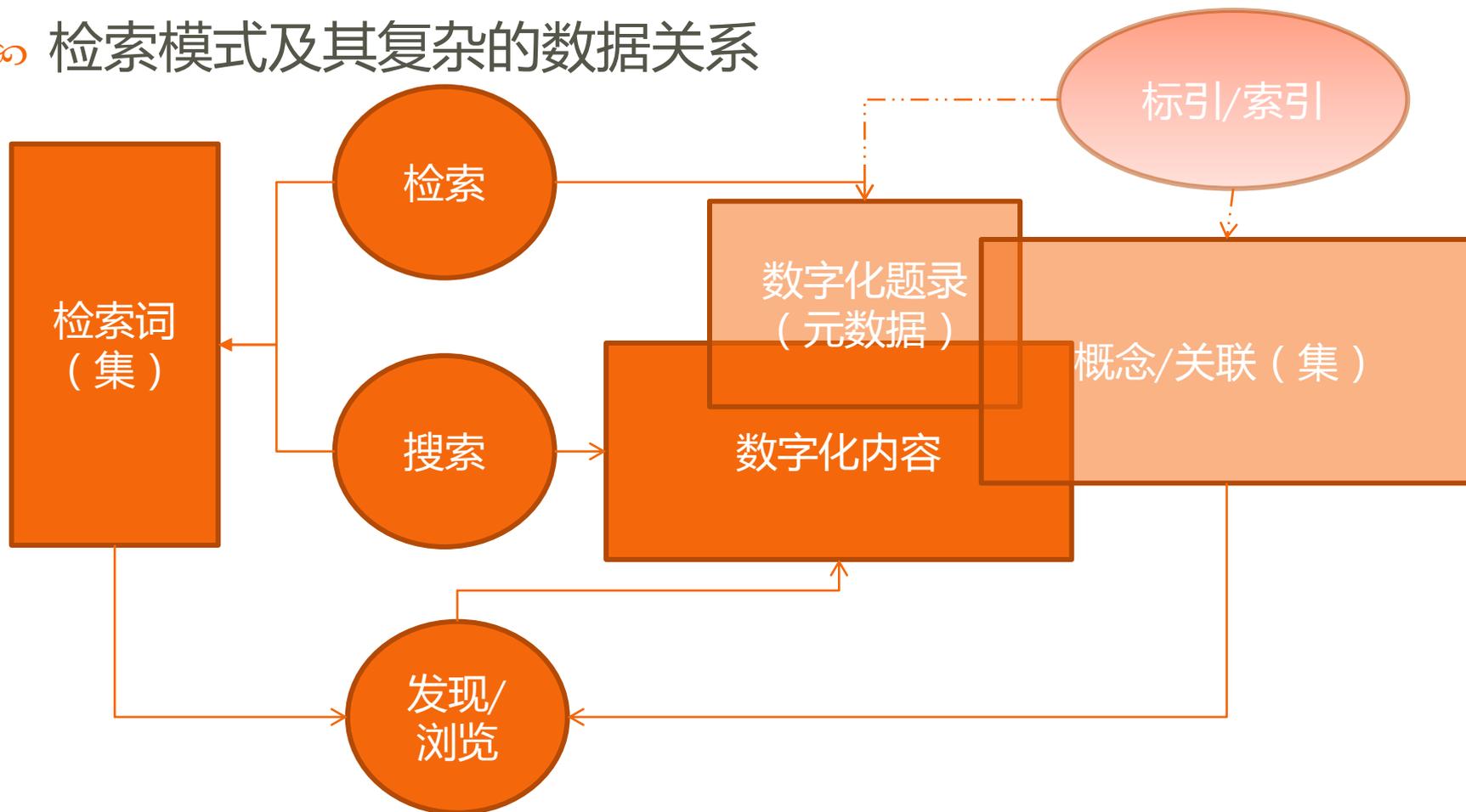
- 大量引进商业化数据库系统，发展联邦式检索系统
- 不断强化自身的OPAC系统，提供集成检索系统
- 引进与合作研发发现系统，目前的成熟度并不理想

∞ 平衡之道

- 精度与广度是一对矛盾，如何平衡是图书馆既能参与用户争夺战，又能保持自身特质的关键所在

检索模式和数据关系认识

检索模式及其复杂的数据关系



我们的尝试

基本思路

- 以OPAC为核心，尽力提供“一框式”检索服务
- 以可控资源为重心，逐步构建一体化检索服务
- 先易后难，逐步推进I³服务（ Integrated , Intelligent , Interesting ）

在OPAC中已经实现的主要检索功能

- 搜索词提示：借鉴了搜索引擎的搜索助手技术
- 相关词提示：纠错助手，对常用检索词效果明显
- 集成中外文电子书检索：电子资源集成检索
- 集成学位论文/博士后出站报告检索：自建资源集成检索
- 相似度排序的数字资源目录信息呈现：显示关联资源
- 虚拟书架：按馆藏排架顺序显示纸质关联资源
- 阅读与下载：对电子资源可以直接在线阅读，部分资源可以下载

搜索词提示

- 借鉴了搜索引擎的搜索助手技术
- 检索词集（检索命中集）是核心，采用前方匹配方式

The screenshot shows the library's search page at opac.njnu.edu.cn/opac/search.php. The main header is "iLIB 南京师范大学图书馆书目检索系统". Below the header is a navigation bar with links: "书目检索", "分类浏览", "期刊导航", "新书通报", "公共书架", "信息发布", "其它馆藏", "读者荐购", "我的图书馆". A secondary navigation bar includes: "简单检索", "多字段检索", "全文检索", "热门借阅", "热门评价", "热门收藏", "热门图书".

The "馆藏书目检索" section features a search input field containing "h". A dropdown menu displays suggestions:

Search Term	Number of Results
heart of darkness	15 结果
hadoop	7 结果
html5	22 结果
hsk	53 结果
human development	38 结果
html	70 结果
how to teach english	5 结果
head first	12 结果
henry fielding	6 结果
hn	5 结果

Additional interface elements include a "检索" button, a "存在与虚无 | 国际经济合作 | 普通话 | 菊与刀 | 人论 | 犯罪构成 | 高等教育学硕士研究生教学参考用书 |" banner, radio buttons for "书", "西文图书", "中文期刊", "西文期刊", and sorting options: "升序排列" and "降序排列". A "关闭" link is at the bottom of the dropdown.

关键词提示

- 智能化纠错助手，在命中结果有偏差时会有帮助
- 检索词集（命中集）是其核心，采用相似度匹配方式
- 寻找更好的切分词组件是提高可用性的重要途径

检索条件：题名=大海的女儿 结果数：0 只显示可供借阅的图书 RSS

按照：

纸型书刊查询结果(结果数: 0)

抱歉，没有找到相关的纸型书刊，建议您用如下关键词进行查询：

相关搜索 海的女儿(5) 大梭的女儿(1) 乐观者的女儿(3) 上尉的女儿(6) 饥饿的女儿(4) 女儿(43) 不存在的女儿(1) 女儿魂(1)

相关资源   

[↑ TOP返回顶部](#)

电子资源查询结果(结果数: 3)

1. 大海的女儿 编委会编 7-5027-4939-X	超星电子书 阅读 / 下载
2. 大海的女儿：梨花姑娘传奇 李虹 华文出版社 7-5075-0916-8	方正电子书 阅读 / 下载
3. 大海的女儿——颜一烟的生平和创作 刘庆俄 7-80037-266-9	超星电子书 阅读 / 下载

集成中外文电子书检索

电子资源集成检索的尝试，已经集成了约115.5万册中外电子图书，包括超星、方正、书生中文电子书，TAO台湾学术数据库电子书，Ebrary、金图国际、EBSCO、Springer、Mylibrary等外文电子书

检索条件: 题名=大尉 结果数: 0 只显示可供借阅的图书 [RSS](#)

按照:

纸型书刊查询结果(结果数: 0)

抱歉，没有找到相关的纸型书刊

相关资源 [e读](#) [读秀](#) [外文期刊网](#)

[↑ TOP](#) 返回顶部

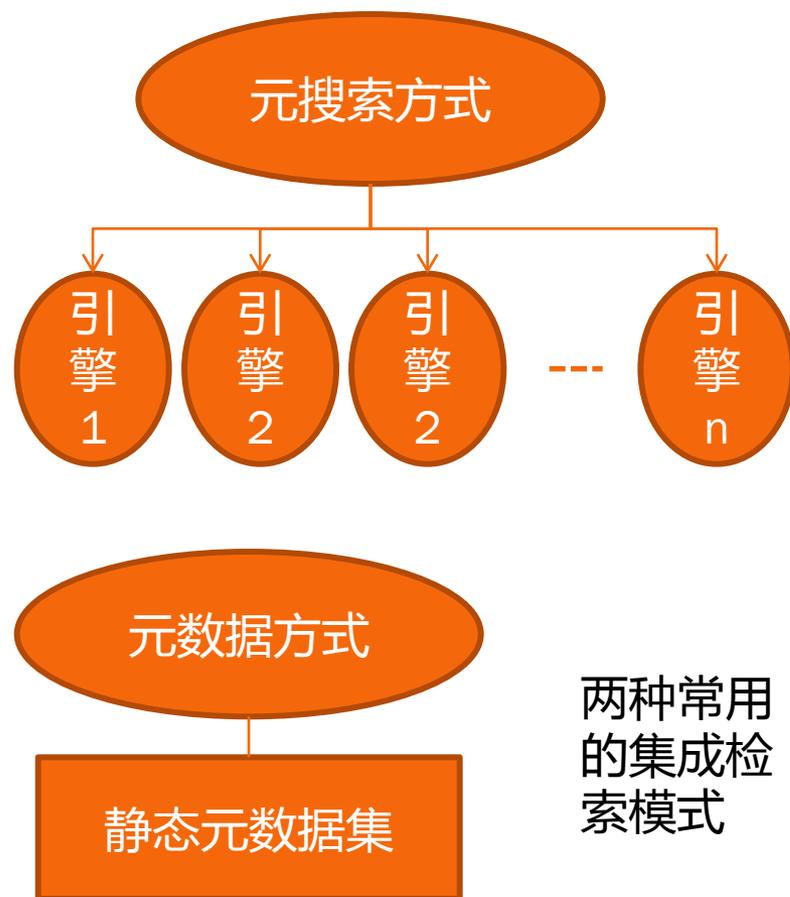
电子资源查询结果(结果数: 4)

1. 船长与大尉 赵迪生注释 7-5600-0352-4	超星电子书 阅读 / 下载
2. 船长与大尉 (上、下册) [苏联]卡维林著 于光译 10208-104	超星电子书 阅读 / 下载
3. 船长与大尉 (上、下册) [苏联]卡维林著 于光译 7-5016-0175-5	超星电子书 阅读 / 下载
4. 十八世纪俄国炮兵大尉新疆见闻录 [俄]伊·温科夫斯基著 [俄]尼·维谢洛夫斯基编 7-5316-3565-8	超星电子书 阅读 / 下载

集成中外文电子书检索

主要的考虑和实现方法

- 电子书更新周期较长，具有稳定性好的特点，易于实现
- 以本地镜像数据为主，具有一定可控性和实现上的便利
- 基于响应速度和服务稳定性考虑，采用静态元数据方式（数据库连接方案）实现
- 正考虑对电子期刊的集成检索问题，但元数据采集是个难题
- 建设元数据仓储时的数据清洗和合并处理需要很多细致的工作



两种集成方案的优缺点

元搜索方案：动态反映数据

- 以代理技术和WEB数据挖掘技术为主，适应性较好，可动态更新数据，但引擎处理算法需随着数据源的变化而时常更新
- 数据的实时有效性好，但结果集的延时响应问题比较突出
- 对网络依赖度极高，通讯、网站、服务器等环节均有重大影响

元数据方案：静态反映数据

- 以数据库管理技术为核心，静态存储元数据，易于长期保存
- 检索表达式的处理相对简单，检索响应速度快，服务稳定性好
- 数据存在同步时滞，元数据的及时采集和数据规范有一定困难

混合模式：动静结合的一种集成方案

- 通常的方案：前端为元数据模式，后端加载元搜索引擎
- 工程实现难度较大，牵涉的技术面较多，多用于改进改良系统

集成自建资源检索

- 力图改变自建资源分散孤立、自成一体现状
- 目前OPAC中集成了本校博硕士学位论文近2.5万篇，与中外电子书采用相同的技术实现方案，两者融为一体
- 目前正在准备加入更多的自建资源，包括多媒体资料

检索条件: 题名=去吧, 摩西 结果数: 2 只显示可供借阅的图书 [RSS](#)

按照:

纸型书刊查询结果(结果数: 2)

1. New essays on Go down, Moses = 《去吧, 摩西》新论 / I712.074/W133 Linda Wagner-Martin编, Peking University Press, 2007.	西文图书 可借3册, 共3册
2. 去吧, 摩西 47.6352/13.315/(4) (美)威廉·福克纳著, 上海译文出版社 1996	中文图书 可借2册, 共3册

相关搜索 韩学研究(2) 社会学研究(16) 庄子研究(4) 老子研究(7) 多学科研究(1) 社会科学研究(65) 欧美文学研究导引(1)

[TOP](#) 返回顶部

电子资源查询结果(结果数: 3)

1. 去吧, 摩西 [美]威廉·福克纳著 李文俊译 7-5327-1752-6	超星电子书 阅读 / 下载
2. 《去吧, 摩西》中的男女合作与男性统治 冯燕 外国语学院	南师大博硕士学位论文 阅读
3. 《去吧, 摩西》之生态主义解读 王蔚 外国语学院	南师大博硕士学位论文 阅读

显示关联资源—基于检索词

基于检索词，将纸质资源与数字资源同时显示，便于发现数字资源颗粒度仍需要调整，以提高发现范围

检索条件：题名=ontological 结果数：2 只显示可供借阅的图书 [RSS](#)

按照：[入藏日期](#) [降序](#) [排列](#)

纸型书刊查询结果(结果数: 2)

1. 基于本体特色的汉语研究:庆祝薛凤生教授八十华诞文集 H1-53/8.452 中文图书
侍建国, 耿振生, 杨亦鸣主编, 中国社会科学出版社 2011 可借2册, 共2册
2. 符号透视:传播内容的本体诠释:an ontological analysis of media content H0/7.110 中文图书
李彬著, 复旦大学出版社 2003 可借1册, 共4册

[↑ TOP](#)返回顶部

电子资源查询结果(结果数: 4)

1. **Ontological Engineering** EBSCO电子书
Gomez-Perez, Asuncion.-Fernandez-Lopez, Mariano-Corcho, Oscar Springer 9781852338404 阅读
2. **Ontological Fundamentals for Ethical Management** Springer电子书
Dominik Heil Springer Science+Business Media B.V. 978-94-007-1875-3 阅读
3. **Depth Psychology, Interpretation and the Bible : An Ontological Essay on Freud** Ebrary电子书
Polka, Brayton McGill-Queens University Press 9780773568853 阅读
4. **Ontolinguistics : How Ontological Status Shapes the Linguistic Coding of Concepts** Mylibrary电子书
Schalley, Andrea C. Mouton de Gruyter 9783110197792 阅读

显示关联资源—基于题录

- 实现基于题名、作者等要素的按相似度排序的数字资源发现
- 相似度算法组件及其参数调整是可用性的关键

书目信息
读者预约
通借通还
机读格式
引文格式
实时虚拟书架

题名/责任者: 太阳花 向阳开:1949--1979年部队创作歌曲选/昆明部队政治部歌舞团编
出版发行项: 昆明:云南人民出版社,1979
ISBN及定价: /0.57
载体形态项: 199页;19cm
团体责任者: 昆明部队政治部歌舞团 编
科图法分类号: 48.952/8.223
一般附注: 献给国庆30周年
馆际资源: e读资源
相关资源: [豆瓣douban](#) [Google](#) [读秀](#) [SwetsWise Linker](#) [e读](#) [外文期刊网](#)

总体评价: ☆☆☆☆☆ (共0票) 我的评价: ☆☆☆☆☆

纸型书刊列表

索书号	条码号	年卷期	馆藏地	馆藏地地理位置
48.952/8.223	60062133	19	华夏中文样本图书借阅室 <small>实时虚拟书架</small>	华夏中文样本图书借阅室
48.952/8.223	2000900644	1979	敬文中文图书借阅室 <small>实时虚拟书架</small>	敬文馆六楼东中文图书第五借阅室
48.952/8.223	2000900647	1979	随园中文图书借阅室 <small>实时虚拟书架</small>	随园馆二楼北区中文图书第四借阅室

相关电子资源

	资源名	名称相似度	作者	ISBN
超星	太阳花	<div style="width: 100%; height: 10px; background-color: #ccc;"></div>	文爱艺著	7-5060-1996-5
方正	太阳花水晶心(上)	<div style="width: 100%; height: 10px; background-color: #ccc;"></div>	北京师联教育研究...	7-89998-977-9
方正	太阳花水晶心(下)	<div style="width: 100%; height: 10px; background-color: #ccc;"></div>	北京师联教育研究...	7-89998-977-9
书生	种一片太阳花	<div style="width: 100%; height: 10px; background-color: #ccc;"></div>	李天芳	10151◆◆836

显示关联资源—虚拟书架

- 按书库、排架规则显示前后相关的纸质资源
- 与借阅信息关联产生动态的模拟效果
- 浏览发现，本质是分类/主题的聚合体

The screenshot shows a library interface with a navigation bar at the top containing tabs: 信息, 读者预约, 通借通还(COOP), 机读格式, 引文格式, and 实时虚拟书架(COOP). Below the navigation bar, there are dropdown menus for '当前馆藏地: 敬文中文样本图书借阅室' and '显示选项: 显示图书复本', along with a link '关于本系统'. The main area displays a list of book spines. The spine for '投资项目评估与工程项目' is highlighted in red. Green arrows on the left and right sides of the book list indicate navigation capabilities.

书名	作者/编者	出版社
全球市场中的企业与政府	(美)默里·	上海三联书店
建设项目规范高效操作规...	丁晓欣、聂...	中国时代经济...
世纪工程与未来中国	李五等...	中国社会科学出...
项目可行性研究与评估	王勇、方志...	中国建筑工业出...
投资项目评估	李晓蓉编...	南京大学出版社
投资项目评估与工程项目	张明等著	中国物价出版社
投资项目评估	张启振、...	厦门大学出版社
建设项目经济社会评价	林晓言等...	中华工商联合出...
项目决策分析与评价第...	全国注册...	机械工业出版社
项目决策分析与评价第...	全国注册...	机械工业出版社
投资项目后评价	姜伟新、...	中国石化出版社
投资项目评估第2版	徐强主编	东南大学出版社
基于计算机实验的工程供...	盛昭瀚、...	上海三联书店

对全文搜索的试验

全文搜索的前提条件

- 文本型全文：OCR文本，TXT/WORD/PDF文本。
- 标引：题录，原文文件。
- 元数据定义、标引深度、对象文件组织将决定未来的应用

可资参考的目标：读秀/百链、google/百度等搜索引擎

- 特征：出处（原文），命中的文字片段



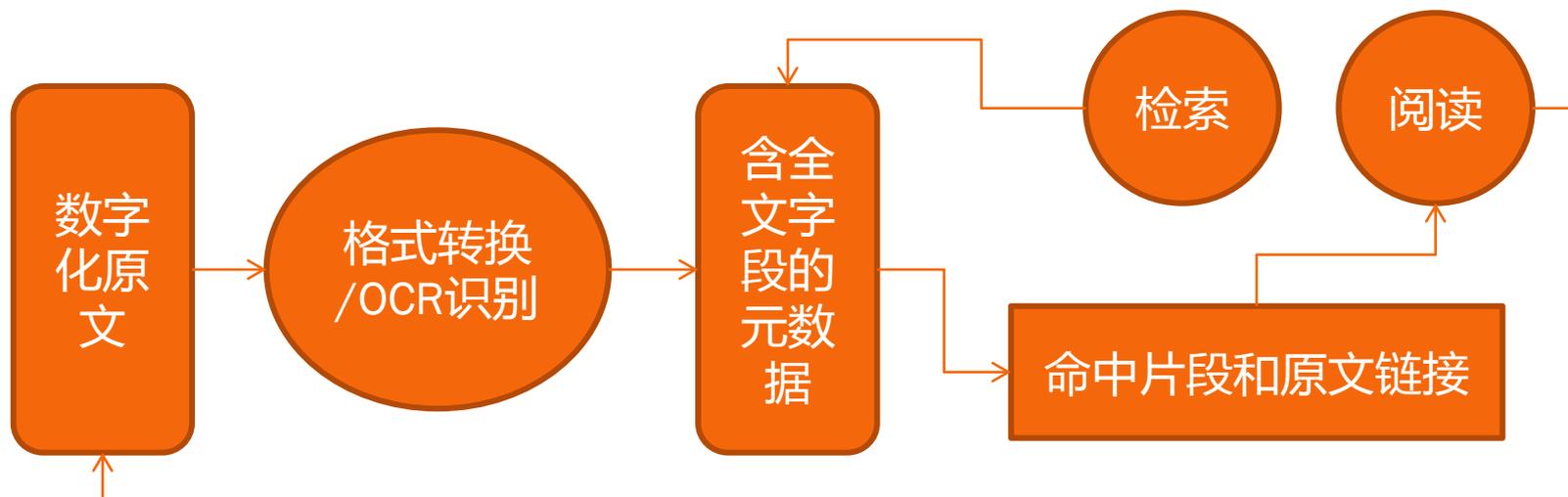
The screenshot shows a Google search page with the following content:

- Address bar: <https://www.google.com.hk/search?q=白流苏和范柳原&oq=bailiu&>
- Search bar: 白流苏和范柳原
- Navigation: 网页 (selected), 图片, 地图, 更多, 搜索工具
- Results: 找到约 69,500 条结果 (用时 0.35 秒)
- Result 1: [倾城之恋- 维基百科，自由的百科全书](https://zh.wikipedia.org/zh-cn/倾城之恋)
zh.wikipedia.org/zh-cn/倾城之恋 转为简体网页
来自上海大户人家的小姐白流苏，在经历一次失败的婚姻后，身无分文，在亲戚间备受冷嘲热讽，看尽世态炎凉。在偶然的机缘中认识了多金潇洒的单身汉范柳原，便拿 ...
- Result 2: [范柳原_百度百科](http://baike.baidu.com/view/3609698.htm?noadapt=1)
baike.baidu.com/view/3609698.htm?noadapt=1
让他意想不到的是，白流苏竟小心谨慎，闪躲腾挪，不愿轻易上钩。其实，白流苏已经28岁，离婚数年，很愿意嫁给范柳原作为终身依靠。但必须是“嫁”，明媒正娶，而 ...
- Result 3: [《倾城之恋》中之范柳原与白流苏-搜狐娱乐频道](http://yule.sohu.com)
yule.sohu.com 娱乐频道
2005年8月29日 - 范柳原: 英国长大的广东生意人，事业成功，对女人存有戒心。他游戏于情感与女人之间，对女人就像是对生意一样知根知底。他风流自持，精神空虚， ...
- Result 4: [《倾城之恋》中的白流苏与范柳原\(评论: 倾城之恋\)-豆瓣](http://www.douban.com/review/3482190/)
www.douban.com/review/3482190/
★★★★★ 评分: 5 - 评价者: 杭之
2010年7月28日 - 《倾城之恋》是一个动听的而又近人情的故事。.....我喜欢参差的对照的写法，因为它是较近事实的。《倾城之恋》里，从腐朽的家庭里走出来的流苏， ...

对全文搜索的试验

主要的策略

- 对扫描文件进行OCR识别与转换，对其他文件进行格式转换
- 将转换后的文本作为一个字段放入标引的元数据中
- 在元数据中进行检索，提取前后文字片段进行显示
- 对命中记录，按照其标引的内容生成原文阅读链接



对全文搜索与阅读的试验

试验中主要解决的问题

- OCR识别：文字内容基本正确，图表不甚理想，中外文混排、繁简字混排的识别率可以接受，不进行人工纠错，总体正确率>95%。
- 文档拆分、转换与合并：按页拆分，基本可以做到自动化转换，一个文件处理后组织为一个目录，合并生成标准PDF格式单文件。
- 应用级别：利用文件命名规则可以批量标引页码，利用标引信息可以做到按页显示；生成的PDF文档可以满足下载需求。
- 版权保护：使用SWF浏览窗显示，拷贝、打印等操作需授权。

未解决的问题

- 文字识别不能批量处理，文档拆分不能批量组织，效率较低。
- 不能进行公式、分子式等特殊应用的检索（识别后的词序错乱）。
- 使用多个不同的软件工具，一些目录组织依靠人工处理，标引自动化程度低，难以进行标准化流程处理。

全文搜索试验的效果

实验环境下的效果

找到相关的**条目**约 89 条,用时 0.001 秒

查找相关的外文关键词 **知识产权保护**

完善和加强我国知识产权保护制度

PDF下载

收藏

第四节完善和加强我国**知识产权保护**制度近年来,我国一直根据国民经济发展的需要以及世贸组织的规则,对**知识产权保护**制度进行不断的修订、补充和完善,以便使**知识产权保护**制度更好地适应国民经济发展的需要,并与国际... **阅读**

来自 《科技发明与创新及科研不端行为处罚法律依据实务全书 第1卷》 -张明海主编

- P182

来源与出处

在线阅读所在页内容

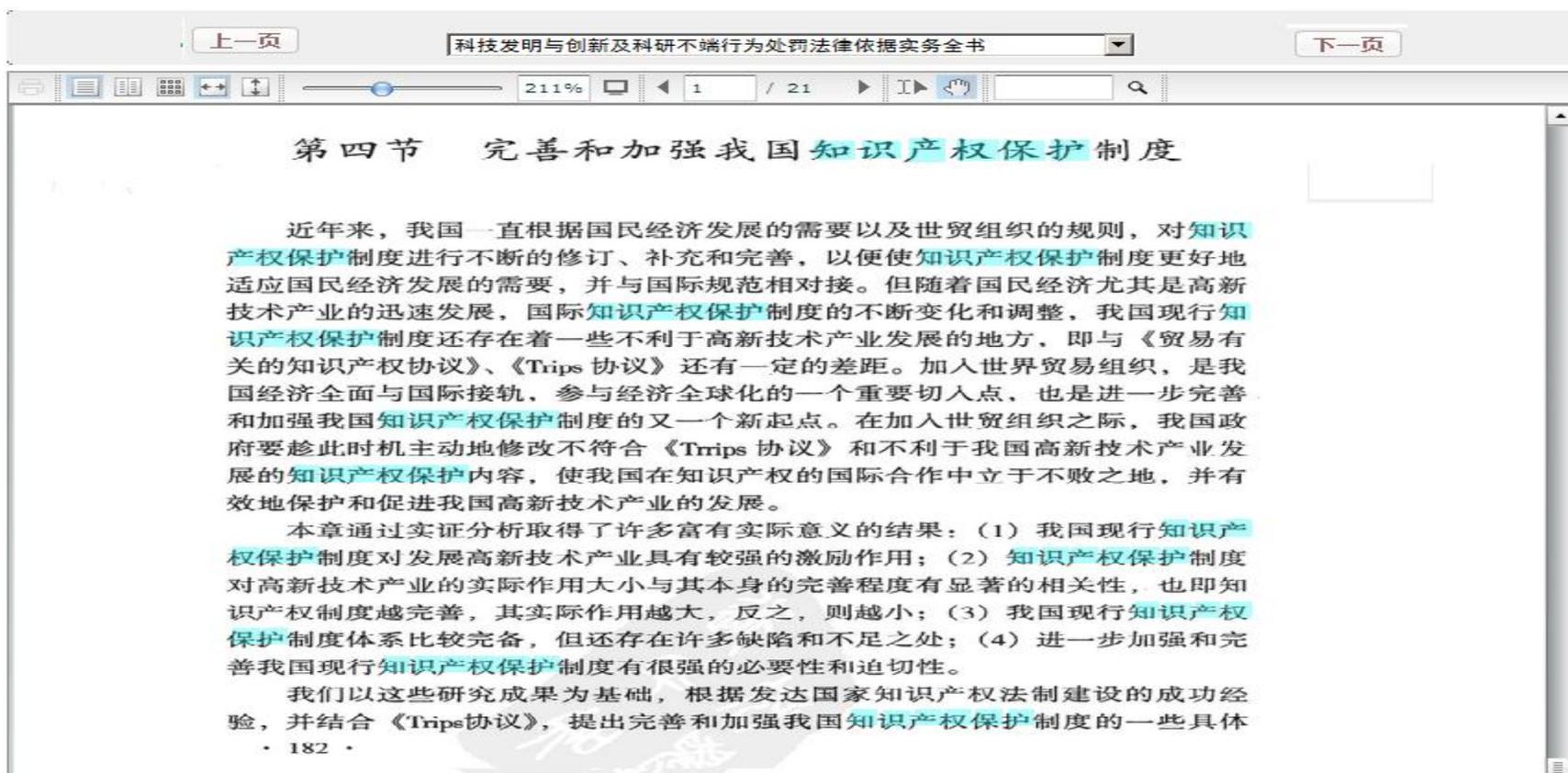
标注
命中
页码

授权用户可以
进行下载

命中词的前后
片段 (单页)

原文在线阅读效果

全文阅读（从命中页起始，按页读）



The screenshot shows a PDF viewer interface. At the top, there are navigation buttons for '上一页' (Previous Page) and '下一页' (Next Page). The title bar indicates the document is '科技发明与创新及科研不端行为处罚法律依据实务全书'. The viewer shows page 1 of 21, with a zoom level of 211%. The main content is a document page with the following text:

第四节 完善和加强我国知识产权保护制度

近年来，我国一直根据国民经济发展的需要以及世贸组织的规则，对知识产权保护制度进行不断的修订、补充和完善，以便使知识产权保护制度更好地适应国民经济发展的需要，并与国际规范相对接。但随着国民经济尤其是高新技术产业的迅速发展，国际知识产权保护制度的不断变化和调整，我国现行知识产权保护制度还存在着一些不利于高新技术产业发展的地方，即与《贸易有关的知识产权协议》、《Trips 协议》还有一定的差距。加入世界贸易组织，是我国经济全面与国际接轨，参与经济全球化的一个重要切入点，也是进一步完善和加强我国知识产权保护制度的又一个新起点。在加入世贸组织之际，我国政府要趁此时机主动地修改不符合《Trips 协议》和不利于我国高新技术产业发展的知识产权保护内容，使我国在知识产权的国际合作中立于不败之地，并有效地保护和促进我国高新技术产业的发展。

本章通过实证分析取得了许多富有实际意义的结果：（1）我国现行知识产权保护制度对发展高新技术产业具有较强的激励作用；（2）知识产权保护制度对高新技术产业的实际作用大小与其本身的完善程度有显著的相关性，也即知识产权制度越完善，其实际作用越大，反之，则越小；（3）我国现行知识产权保护制度体系比较完备，但还存在许多缺陷和不足之处；（4）进一步加强和完善我国现行知识产权保护制度有很强的必要性和迫切性。

我们以这些研究成果为基础，根据发达国家知识产权法制建设的成功经验，并结合《Trips 协议》，提出完善和加强我国知识产权保护制度的一些具体

• 182 •

对阅读的理解

☞ 阅读的变化

- 片段化泛读逐渐成为习惯。人们普遍追求快速阅读，只对感兴趣的内容部分进行重点阅读，一般不对通篇文章进行精读。
- 越来越多的人阅读时间碎片化。能抽出大段时间专注于阅读的人越来越少，快餐式阅读倾向明显。

☞ 阅读的内容形态

- 简短的信息：题录，摘要，缩略图、文字/影视片段等
- 完整的内容：原文（全文），高分辨图像，影视全片等

☞ 我们的对策

- 改进信息展示方式，在有限屏幕空间可展示尽可能多的内容
- 改进全文阅读方式，努力提供章节以下级别的阅读服务
- 提供全文在线阅读和下载服务，满足通篇阅读需要

博士后出站报告中的尝试

80 可选择的详尽页面

 返回首页

博士后信息

博士后姓名	崔大成
流动站名称	04数学博士后流动站
专业名称	应用数学
	显示更多.....

研究报告信息

研究报告题名	二维定常亚音速环流的存在性和稳定性 一类非线性薛定谔方程约束态的存在性和中性
研究报告题名(英文)	无
报告提交日期	2011-05-01
	显示更多.....

博士后信息

博士后姓名	崔大成
流动站名称	04数学博士后流动站
专业名称	应用数学
学术联系教师姓名	张吉惠 教授
学术联系教师单位	数学科学学院
第二联系教师姓名	
第二联系教师单位	

[隐藏](#)

研究报告信息

研究报告题名	二维定常亚音速环流的存在性和稳定性 一类非线性薛定谔方程约束态的存在性和中性
研究报告题名(英文)	无
报告提交日期	2011-05-01
离站科研评审会日期	2011-05-01
研究报告保密级别	公开
发布授权	全部公开
中文关键词	亚音速环流;Euler方程组;定常等熵无旋流;广义的质量通量条件;加权的Holder估计;约束态;非线性薛定谔方程;若Harnack不等式;Poincare不等式;越山引理;集中紧致性引理
英文关键词	subsonic circulatory flow;Euler equations;steady isentropic irrotational flow;generalized mass-flux condition;weighted Holder estimates;Bound state;nonlinear Schrodinger equation;Harnack inequality;Poincare inequality;the mountain pass theorem;concentrat
中文摘要	无
英文摘要	无
研究报告总页数	76
参考文献总数	51

分章节的全文阅读

按章次拆开原文，对章节进行标引并提供阅读服务

研究报告信息

研究报告题名 二维定常亚音速环流的存在性和稳定性 一类中性

研究报告题名(英文) 无

报告提交日期 2011-05-01

[显示更多.....](#)

全文在线阅读

- 封面-谢辞-摘要-目录 [\(在线阅读\)](#)
- 第一章 Preface[P1-4] [\(在线阅读\)](#)
- 第二章 On the existence and stability of 2-D perturbed steady s [\(在线阅读\)](#)**
- 第三章 Existence and concentration of bound states of a class α [\(在线阅读\)](#)
- References[P65-68] [\(在线阅读\)](#)
- Papers list[P69] [\(在线阅读\)](#)
- 全本 [\(在线阅读\)](#)

第二章 On the existence and stability of 2-D perturbed steady s

tions or in one direction. Therefore, it is important to derive the infinity behaviors of the subsonic flow under physical conditions, since essentially one should deal with elliptic equations in unbounded domains.

The problem is described by two-dimensional Euler system for steady isentropic irrotational flow:

$$\begin{cases} \operatorname{div}(\rho u) = 0, \\ \operatorname{rot} u = 0, \\ \operatorname{div}(\rho u \otimes u) + \nabla P = 0, \end{cases} \quad (2.1.1)$$

where $u = (u_1, u_2)$, ρ and P denote the velocity, the density and the pressure respectively. Moreover, for the polytropic gas, the equation of state is given by $P = A\rho^\gamma (1 < \gamma < 3)$ with $A > 0$ a fixed constant, $c(\rho) = \sqrt{P'(\rho)}$ is the local sound speed.

The last three equations in (2.1.1) yield the Bernoulli's law([7]):

$$\frac{1}{2}|u|^2 + \frac{\gamma}{\gamma-1} \frac{P}{\rho} = B, \quad (2.1.2)$$

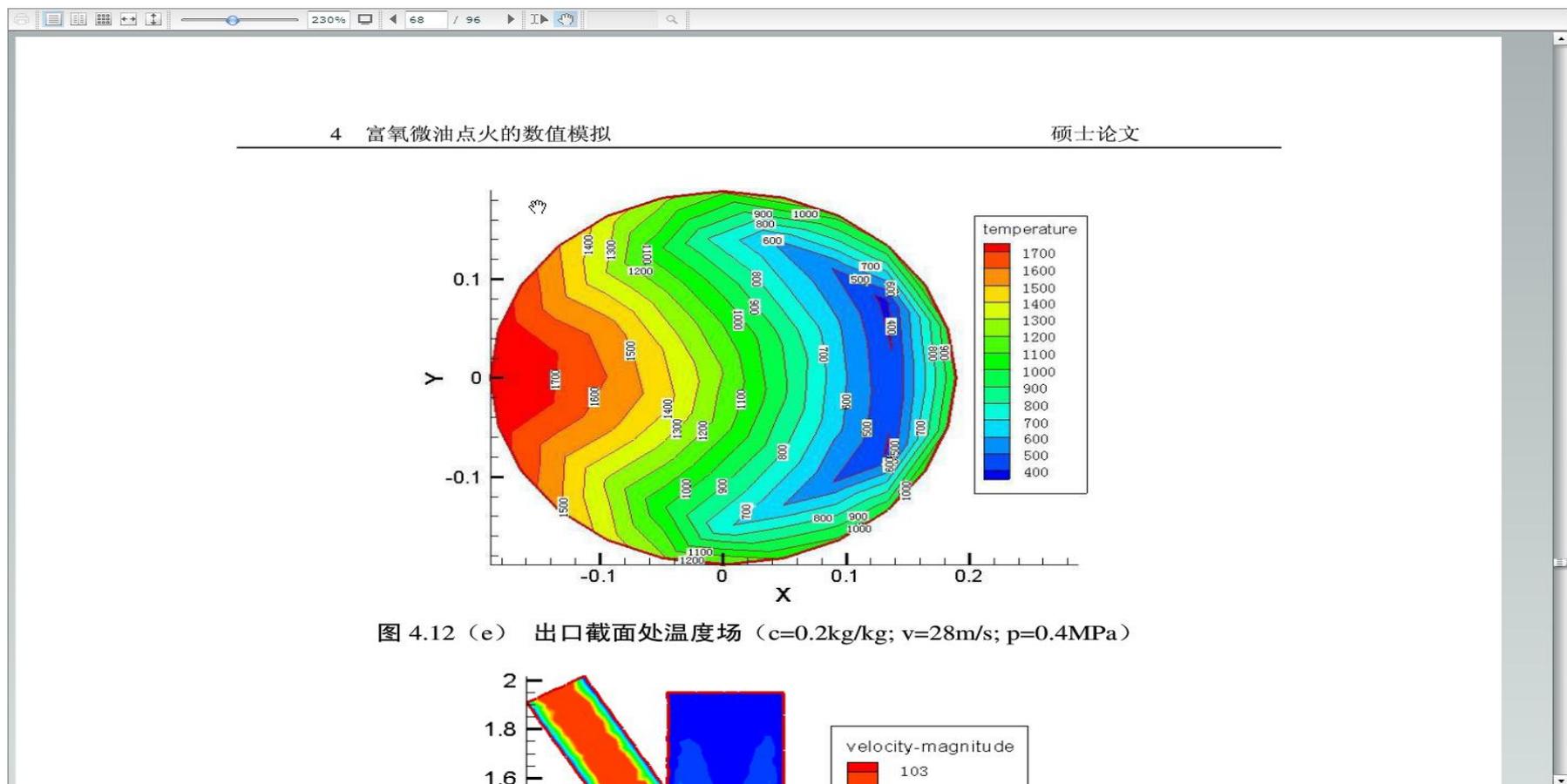
where B is called the Bernoulli's constant.

We intend to look for a subsonic circulatory flow in the unbounded domain D described by $\{x : |x| > 1 + b(x)\}$. Without loss of generality, it is assumed that $b(x)$

5

连续的全文阅读

在博硕士论文系统中的全文阅读



谢谢大家



欢迎会后交流！