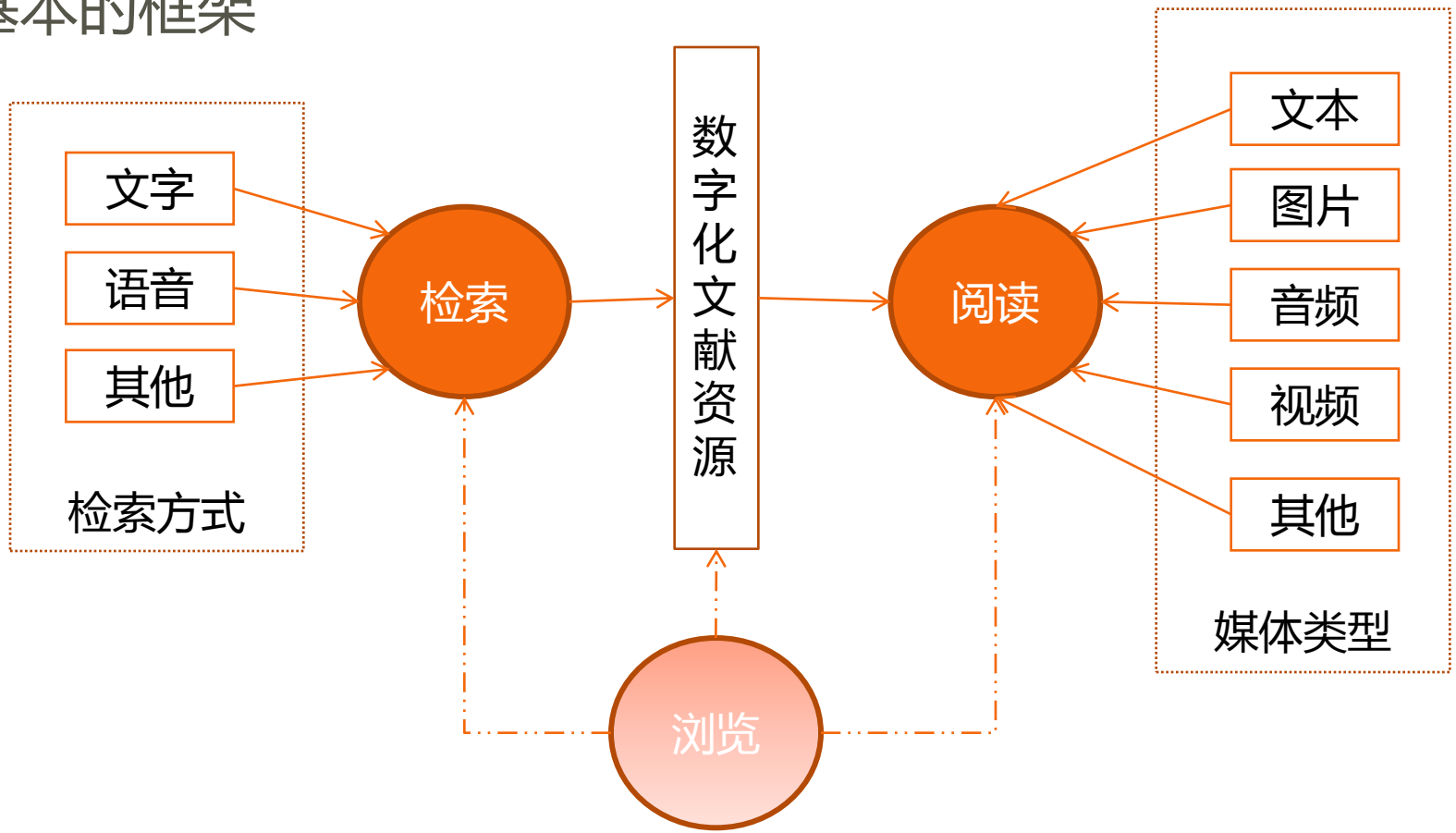
对集成检索与全文阅读的探索

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

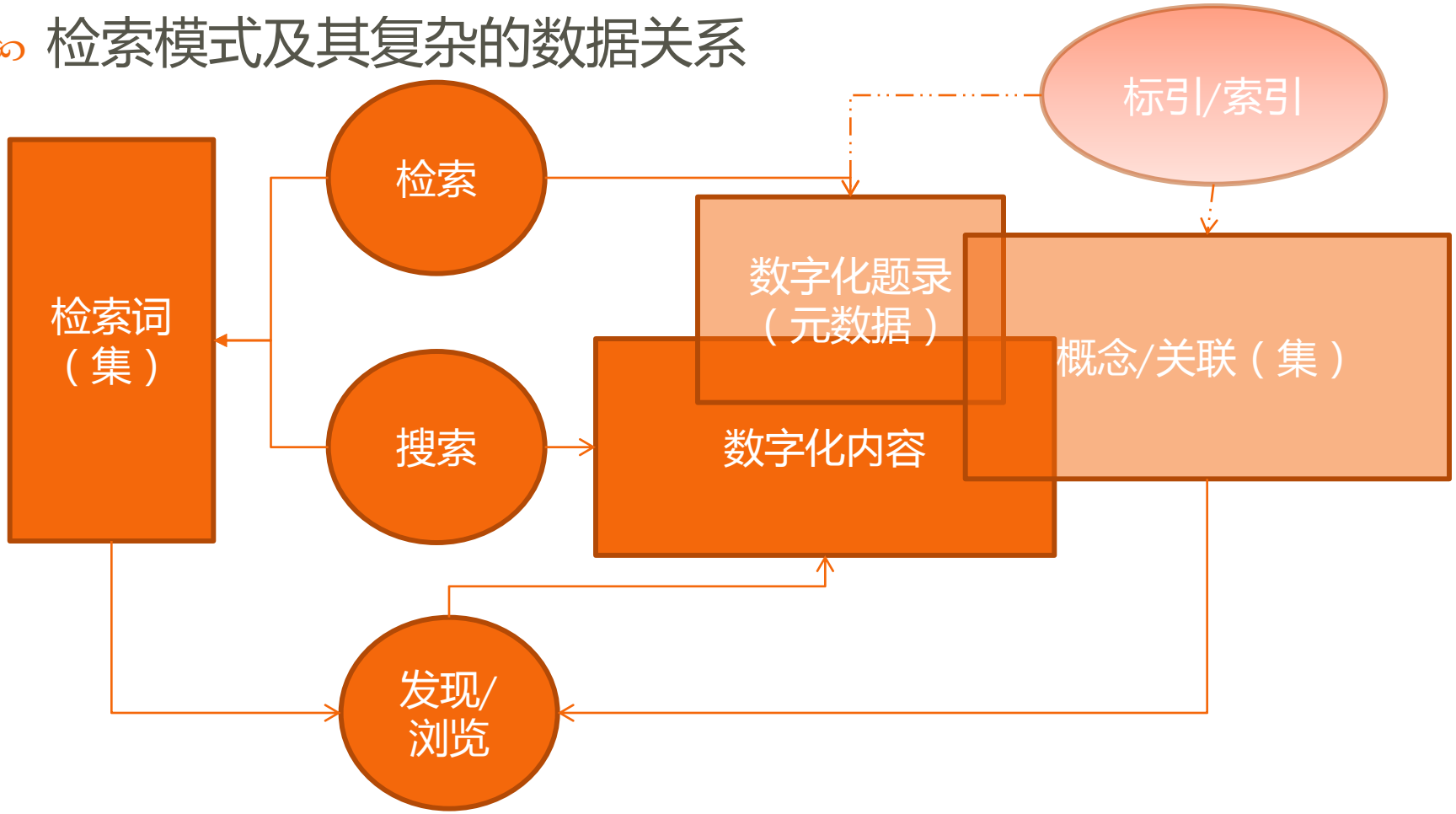
- 一些浅薄的认识
  - 检索：retrieval，讲求精准度，主要应用于元数据级，一维方式
  - 搜索：search，讲求广度，主要应用于内容级，二维方式
  - 发现：discovery，讲求关联，源于检索而高于搜索，准三维方式

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I^3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

相关词提示

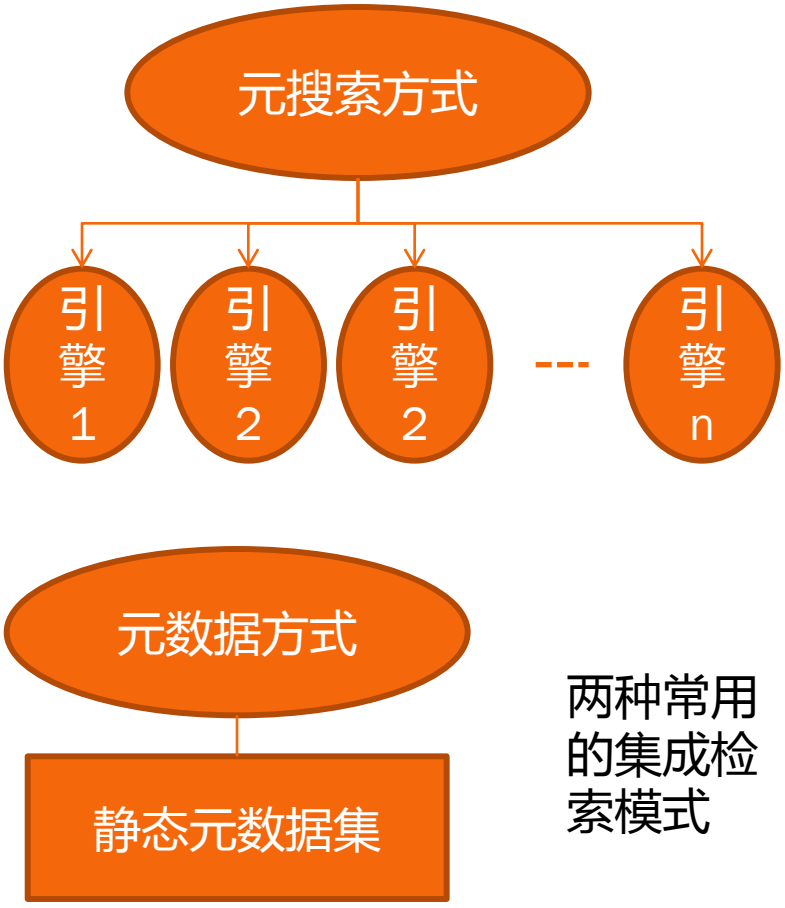
- 智能化纠错助手，在命中结果有偏差时会有帮助
- 检索词集（命中集）是其核心，采用相似度匹配方式
- 寻找更好的切分词组件是提高可用性的重要途径

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

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



两种常用的集成检索模式

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、自成一体的现状
- 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

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

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
G贸mez-P茅rez, Asunci贸n.-Fern谩ndez-L贸pez, 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　Myilibrary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

显示关联资源—虚拟书架

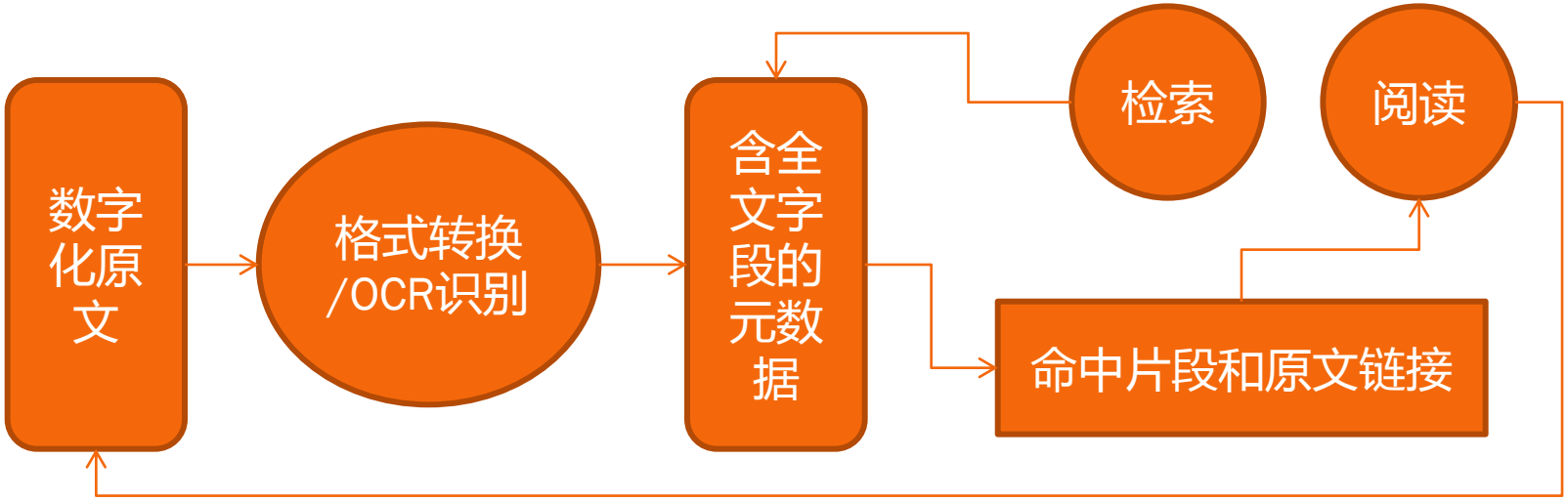
- 按书库、排架规则显示前后相关的纸质资源
- 与借阅信息关联产生动态的模拟效果
- 浏览发现，本质是分类/主题的聚合体

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

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

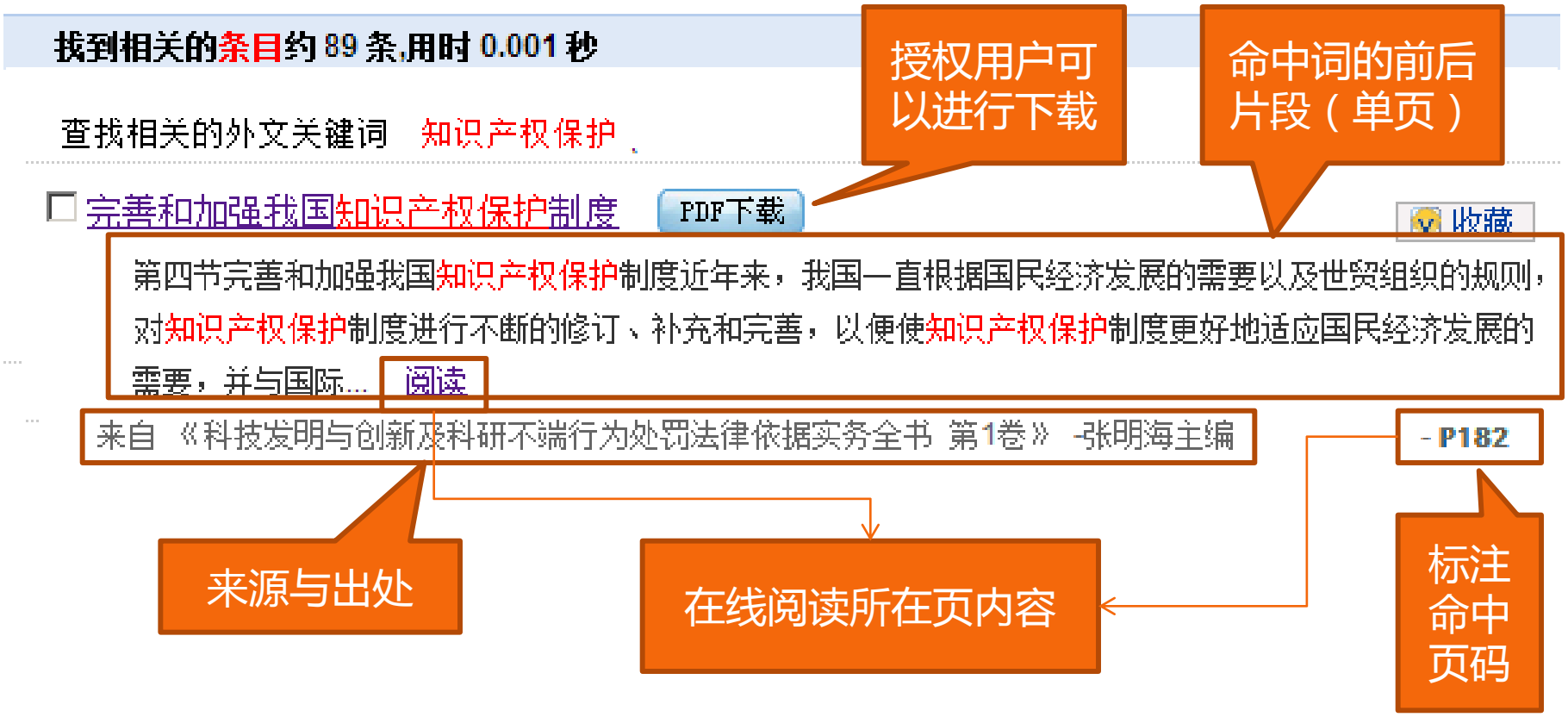


对全文搜索与阅读的试验

- 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。

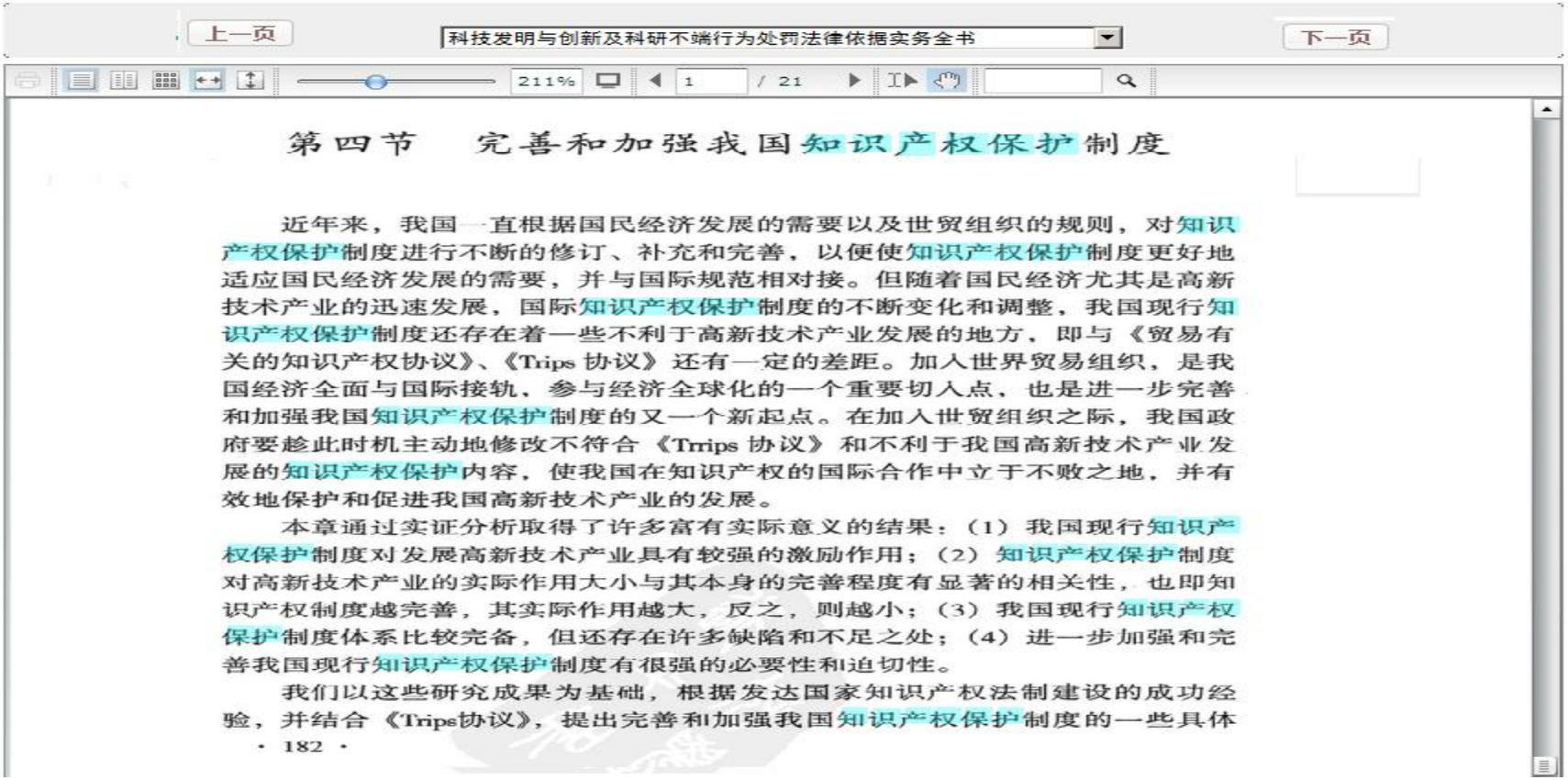
全文搜索试验的效果

- 实验环境下的效果



原文在线阅读效果

- 全文阅读（从命中页起始，按页读）



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

- 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 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	

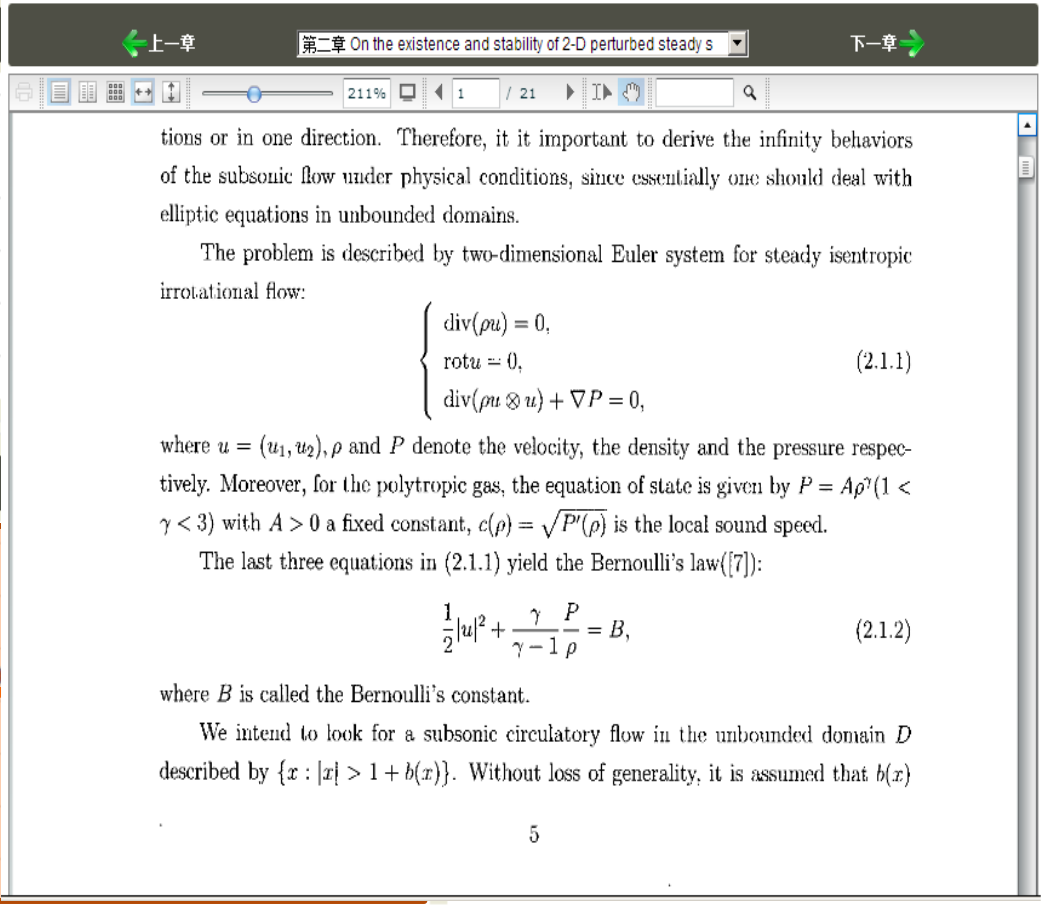
隐藏

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 一类非线形薛定谔方程约束态的存在性和集中性
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;非线形薛定谔方程;若Hanark不等式;Poincare不等式;越山引哩;集中'紧致性引理
英文关键词	subsonic circulatory folw;Euler equations;steady isentropic irrotational flow;generalized mass-flux condition;weighted Holder estimates;Bound state;nonlinear Schrondinger 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



连续的全文阅读

- 在博硕士论文系统中的全文阅读

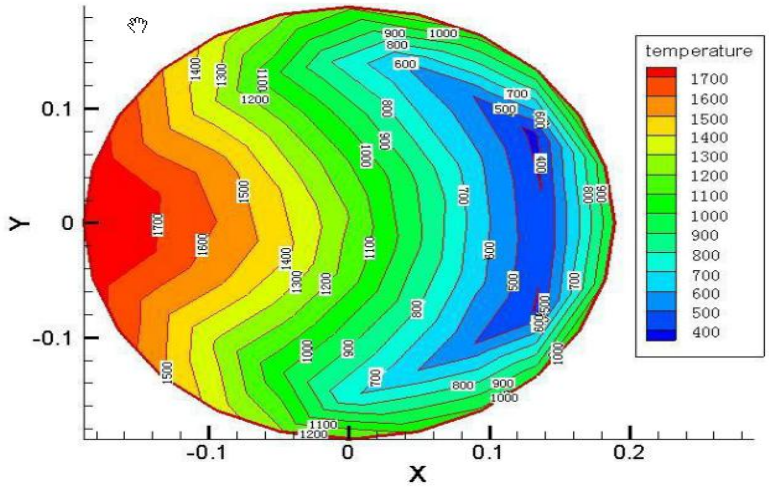
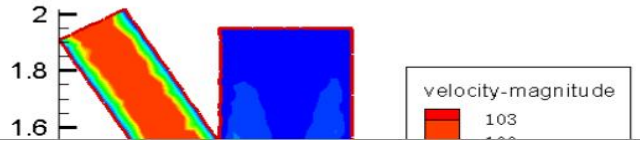


图 4.12（e） 出口截面处温度场（c=0.2kg/kg; v=28m/s; p=0.4MPa）



谢谢大家

欢迎会后交流！