北京大学

中国中文信息学会战略研讨会

需求和应用是中文信息处理事业发展的动力源泉

陈晓鸥　2012. 4. 13

北京大学计算机科学技术研究所

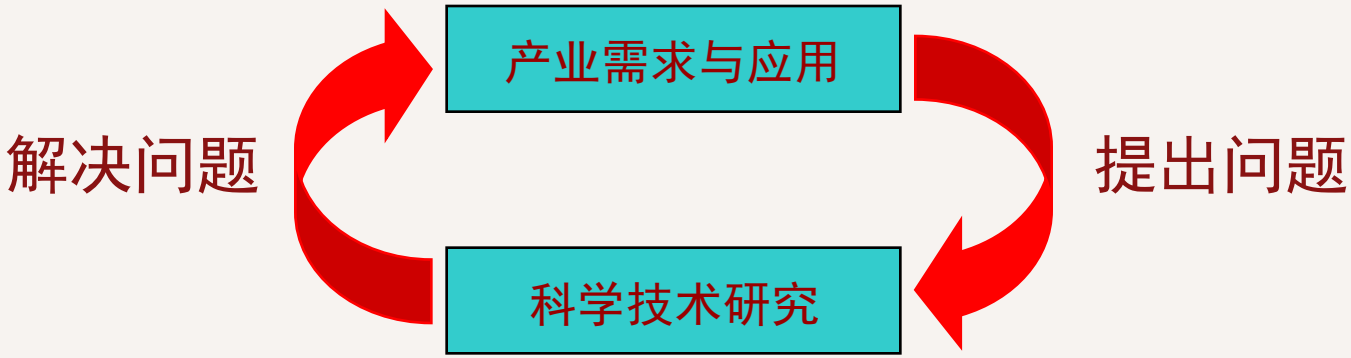
主要内容

一. 在新的历史条件下的定义和理解

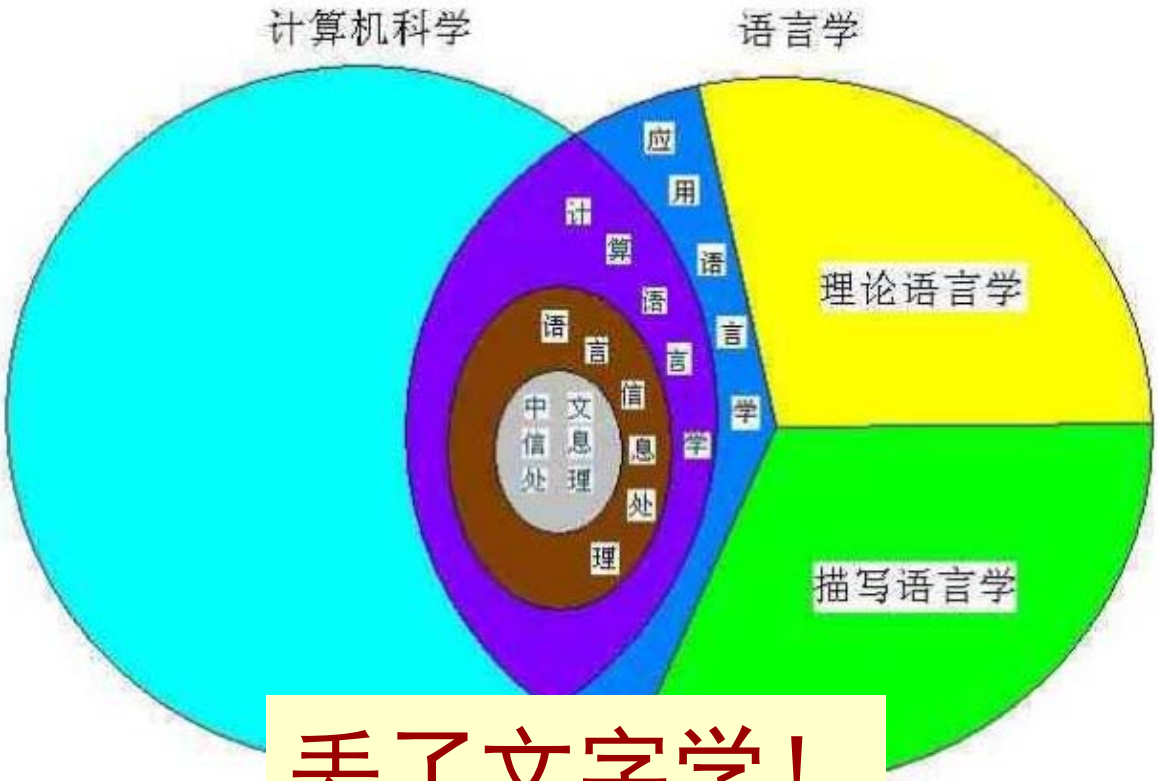
二. 产业发展方向及产业价值链分析

三. 主要的着眼点和着力点应在哪里

四. 领域的重要原始创新可能在哪里



一、在新的历史条件下的定义和理解



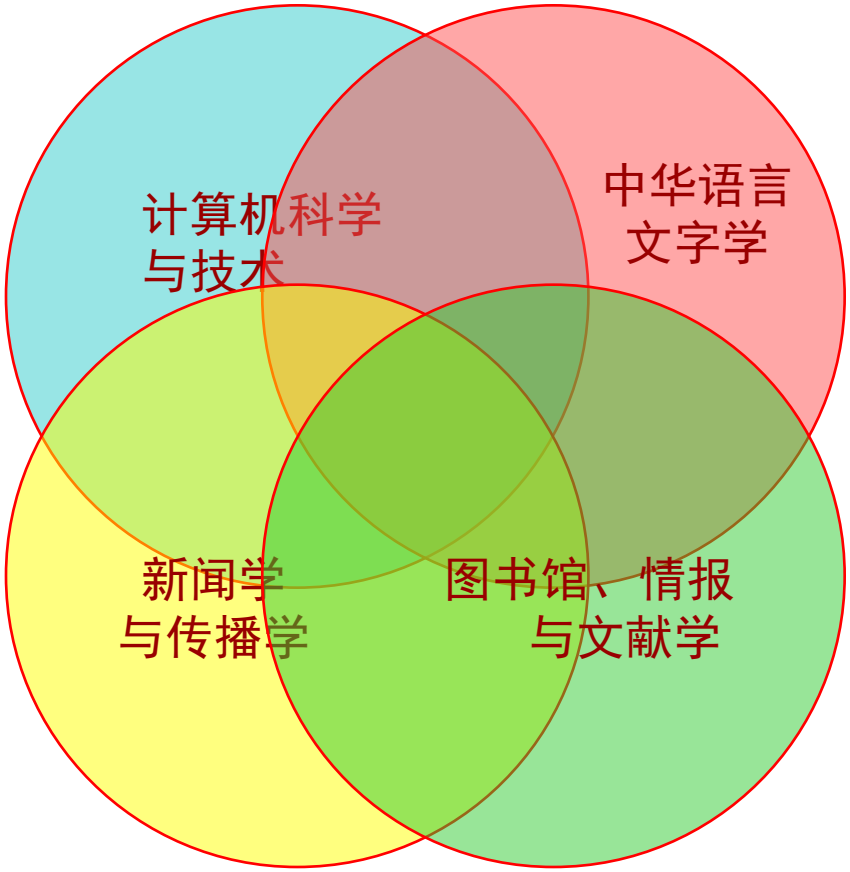
丢了文字学！

中文信息处理 ∈ 语言信息处理 ⊂ 计算语言学 = 计算机科学 ∩ 应用语言学

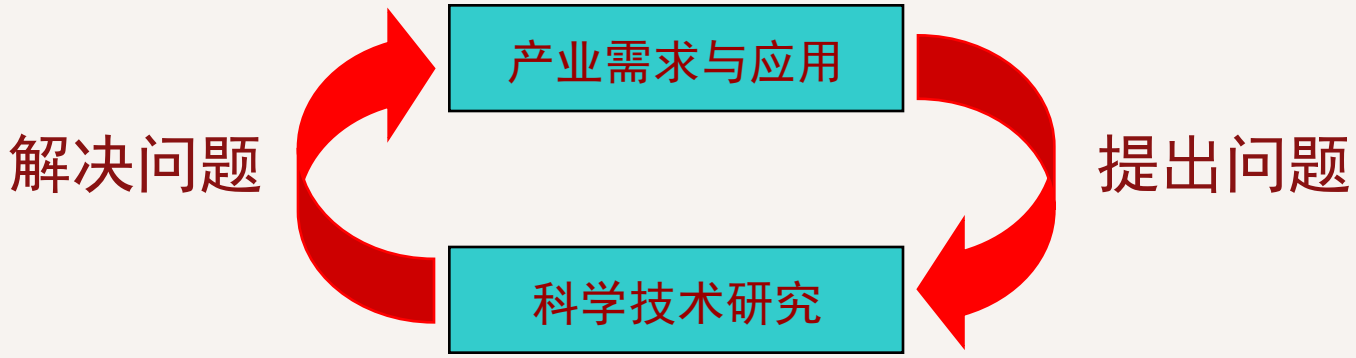
一、在新的历史条件下的定义和理解

- 如果用文
  中
  科，是
  、识别、
  论、方
  学的交叉学
  、传输、输出
  处理问题的理
- 抛出两个
  - 要不要
  - 要不要

  比如加入人类
  信息学，
  这两个都是一
  级学科）



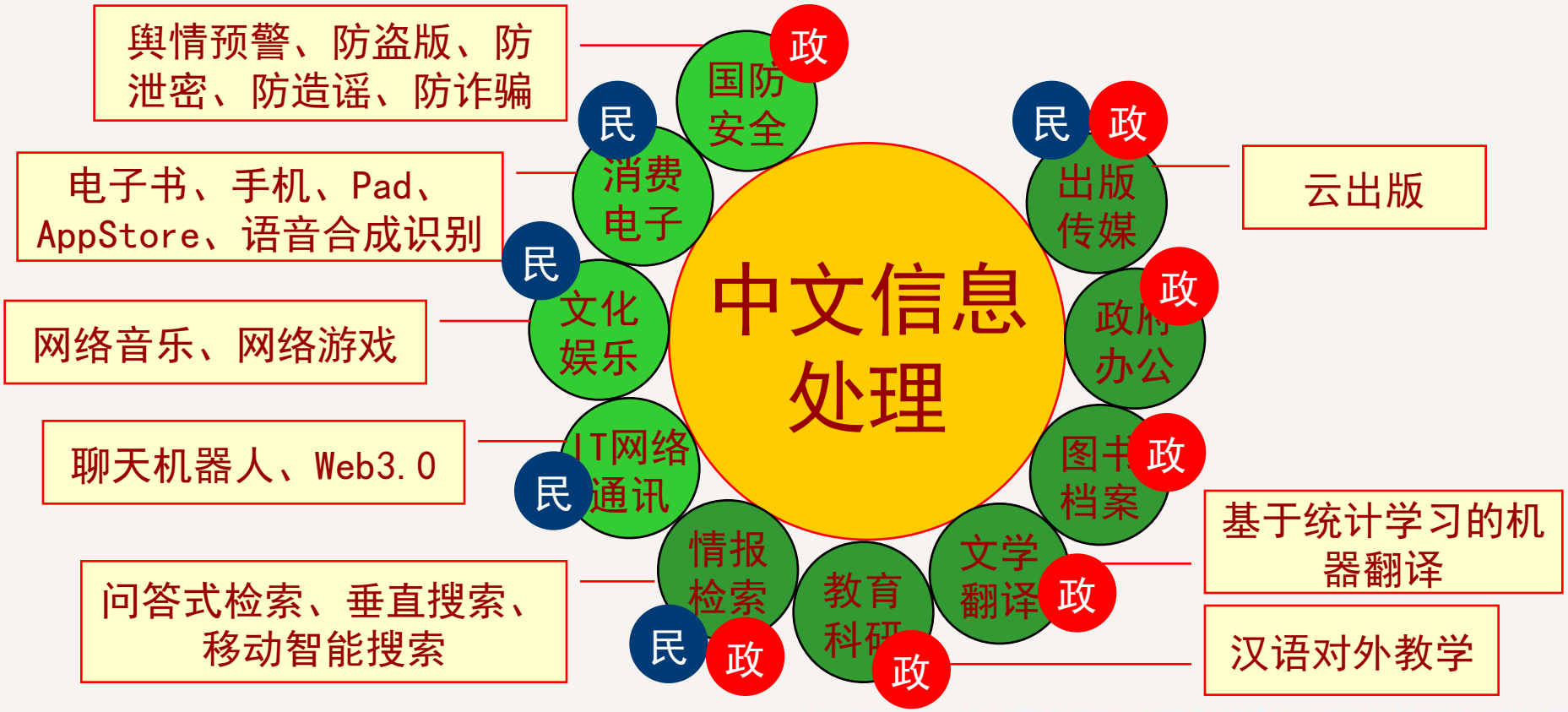
- 新的历史条件下中文信息处理面临的问题
  - 下一步的发展和研究失去强劲动力和重量级的话题
    - 理想的研究环境是



    - 而中文信息处理的研究环境
      - 产业需求与应用的推动力比较弱
      - 新的研究成果在应用上表现不彰

二、产业发展方向及产业价值链分析

- 中文信息处理的产业需求和应用方是谁?
  - 产业发展方向、相关热点话题及投资来源的梳理



二、产业发展方向及产业价值链分析

- 中文信息处理产业价值链分析

二、产业发展方向及产业价值链分析

- 中文信息处理产业价值链分析——结论
  - 对中文音和形的产业发展，已经走完了价值链的全过程
  - 对中文义的产业发展，中间有断层
    - 对“义”的产业发展滞后于音和形的产业发展
    - 义的产业化应用，不需要芯片、板卡、终端这些产业化过程
  - 我认为重要方向或方面应该在“义”的领域，特别是对Web2.0、3.0相关的产业

	算法、引擎、标准	
音：语音识别、语音合成	义：搜索、语法语义分析	形：排版、字库、识别
	芯片、板卡	
音：语音识别、合成芯片		形：RIP、汉卡、图形卡
	终端设备	
音：专用设备、移动终端		形:打印机、电子书、PAD
	技术服务	
音：解决方案、集成	义：搜索门户、应用软件	形：排版软件、集成

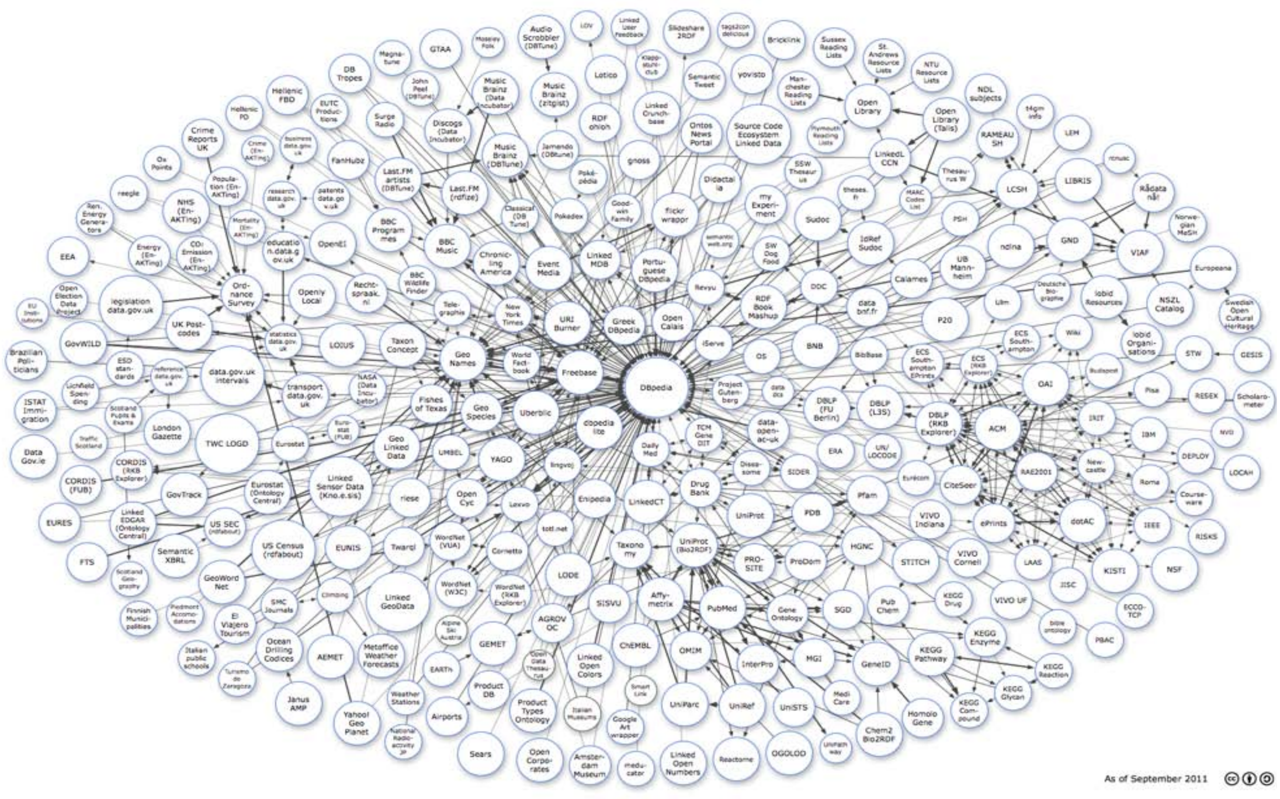
三、主要的着眼点和着力点应在哪里

- 前三十年着力于——如何让中文进入计算机
- 后三十年着力与——如何让计算机“理解”中

	算法引擎	
音：语音识别、语音合成	义：搜索、语法语义分析	形：排版、字库、识别
	芯片、板卡	
音：语音识别、语音合成		形：RIP、汉卡、图形卡
	终端设备	
音：专用设备、移动终端		形:打印机、电子书、PAD
	技术服务	
音：解决方案、集成	义：搜索引擎、应用软件	形：排版软件、集成

三、主要的着眼点和着力点应在哪里

- 推荐两个着力点——之一：语义网及关联数据网（云）
  - 发展迅速的Linked Open Data （LOD）



2011年的LOD

三、主要的着眼点和着力点应在哪里

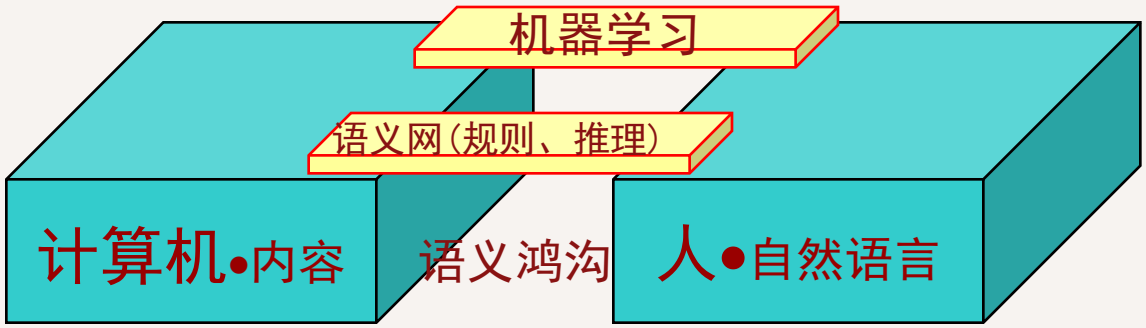
- 推荐两个着力点——之一：语义网及关联数据网（云）
  - 中文的Linked Open Data （LOD）在那里？

表 4　DBpedia的跨语言摘要[16]

序号	语言	摘要个数
1	英语	3 144 000
2	德语	503 000
3	法语	545 000
4	波兰语	430 000
5	荷兰语	392 000
6	意大利语	381 000
7	西班牙语	362 000
8	日语	275 000
9	葡萄牙语	367 000
10	瑞典语	213 000
11	汉语	179 000

三、主要的着眼点和着力点应在哪里

- 推荐两个着力点——之二：基于自然语言的精准检索（问答式检索）
  - 全文检索动辄几十万、几千万的检索结果，人们需要更精准的检索
  - 除了文本内容以外，越来越多的非文本内容需要检索，如图片、视频、音乐、语音等等
    - 对这些内容的最佳检索条件，非自然语言莫数
    - 实现基于自然语言的检索，必须解决语义鸿沟问题
    - 跨越语义鸿沟的常用手段是：机器学习、语义网



四、领域的重要原始创新可能在哪里

- 重要的原始创新可能在云技术与语义网相结合的领域里
  - 回顾一下历史——每一次新的计算平台出现，都对中文信息处理提出新的挑战，带来创新的高潮
    - 小型机计算时代——使汉字进入计算机
    - 微型机计算时代——使中文计算机出版走向实用
    - 互联网计算时代——使中文全文检索得以普及
    - 移动云计算时代——？使中文自然语言检索得以实现
  - 重要的理论原始创新可能在
    - 基于中文自然语言的信息检索
      - 搜索引擎升级、框计算、聊天机器人
    - 互联网中文云安全
      - 舆情预警、防盗版、防泄密、防造谣、防诈骗

北京大学

谢　谢　！