少数派的游戏

——浅谈企业信息检索技术的现状与挑战

智能技术与系统国家重点实验室 智能检索组

刘奕群 2006年3月

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 企业信息检索与传统信息检索的主要差别
- 企业信息检索研究面临的挑战

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

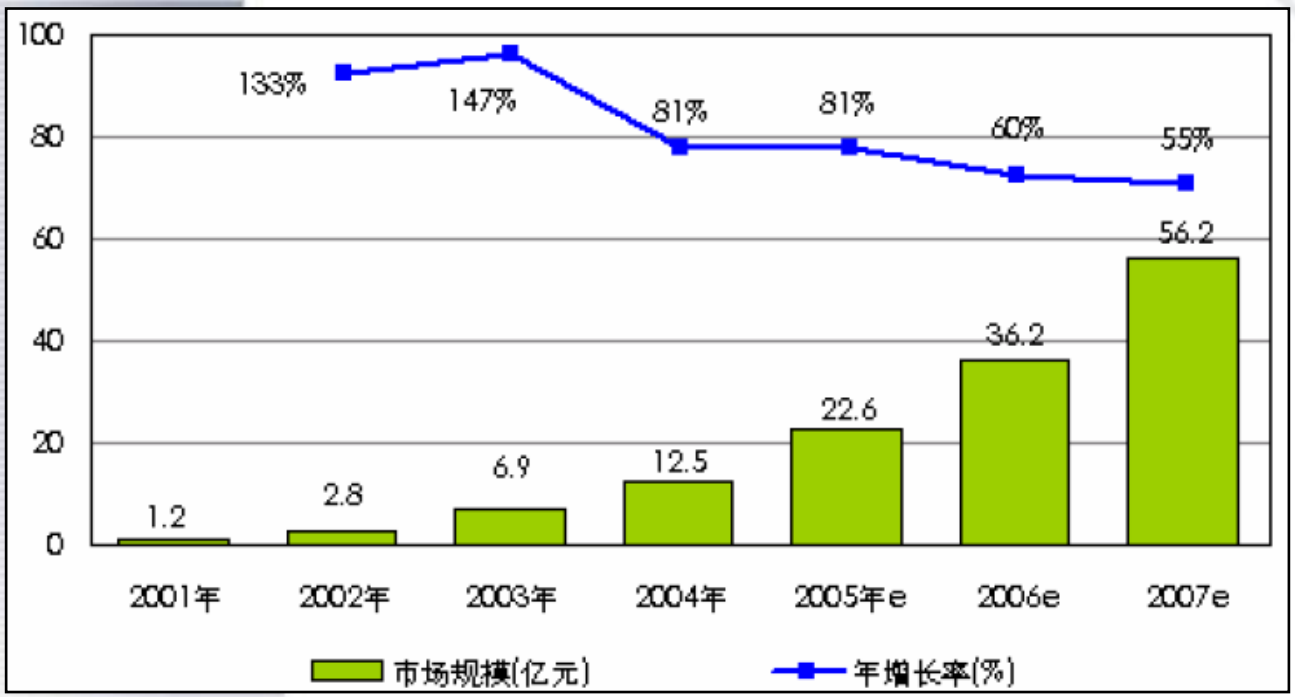
- 问题背景
- 企业信息检索与传统信息检索的主要差别
- 企业信息检索研究面临的挑战

问题背景

- Web的发展带来了什么？
  - 信息数量的急剧膨胀
  - 知识的获取空前简单与繁荣
    - Information is no longer a scarce resource - attention is.
      （纽约时报，2005年10月16日）
    - 在信息化时代，知识实际上已经不是资源，智慧才是资源。（经管学院魏杰教授）
  - 从网络中有效的获取知识正在成为人们生活与工作的必须技能
    - 高科技企业员工1/3的时间用于查找资料
    - 由于无法找到有效信息而浪费的产值占企业收入1/5

问题背景

- 网络信息检索工具/搜索引擎的发展



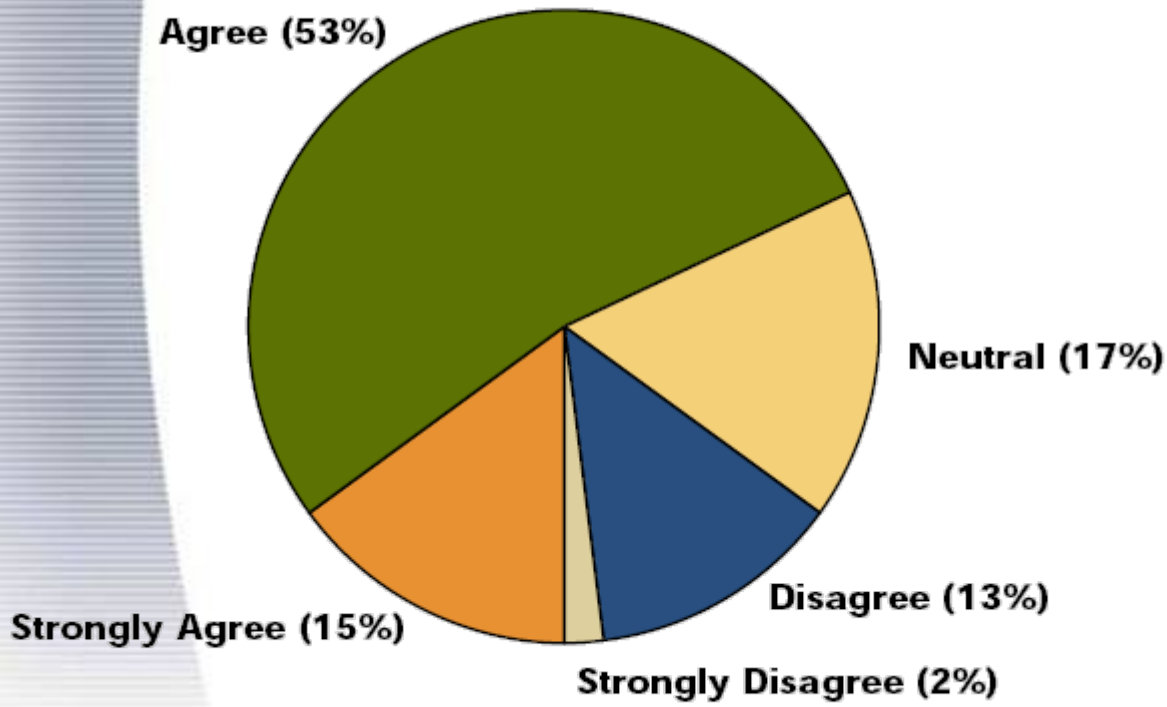
– 与Mp3的市场规模（50亿元）相当，但发展速度更快，利润值更大

问题背景

- 企业数据的增长情况
  - 每年200%的速度增长（超过Web增长）
  - 已有的数据量大大多于Web资源
    - 企业发布到互联网的信息只占到信息量的1%–2%，而98%以上的信息是存储在企业内部的。
    - 80%的数据以文件、邮件、图片等非结构化数据形式存储在局域网的计算机内
  - 相对较少受到研究人员关注

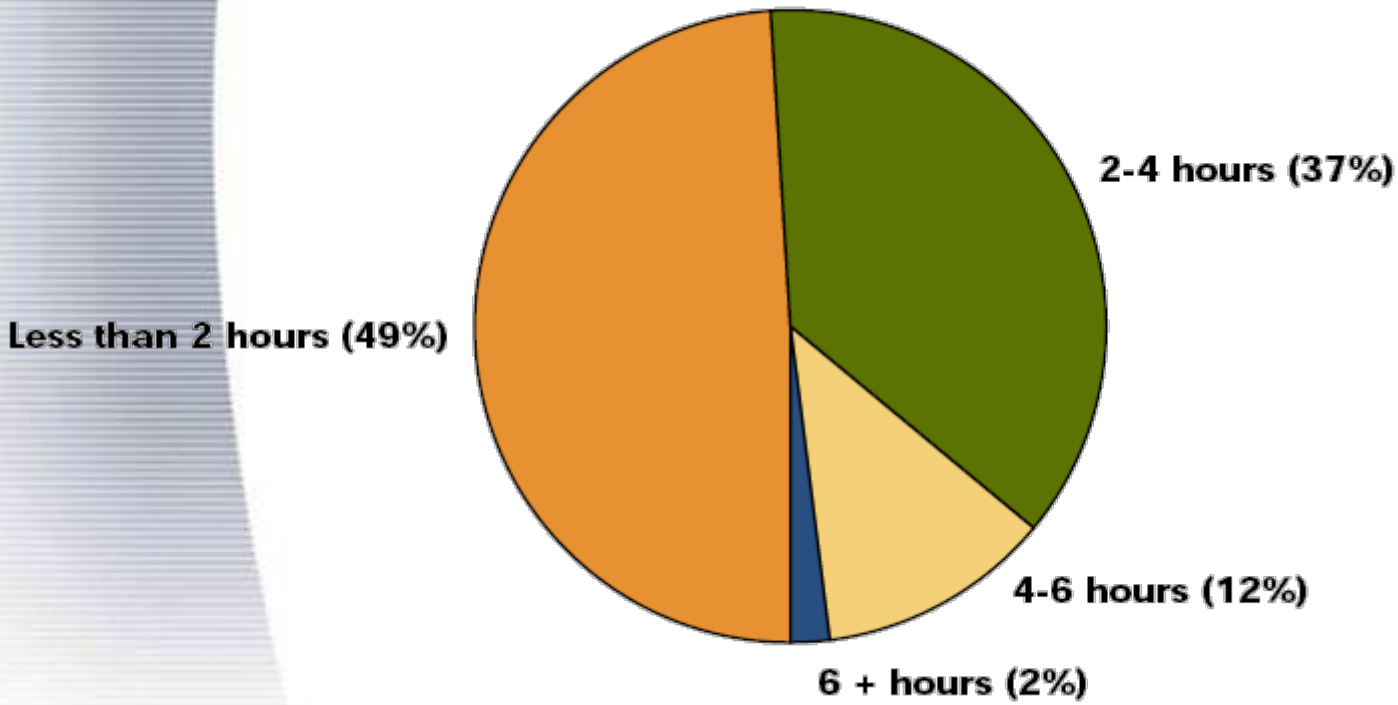
问题背景

- 查找与工作相关的关键信息是困难的



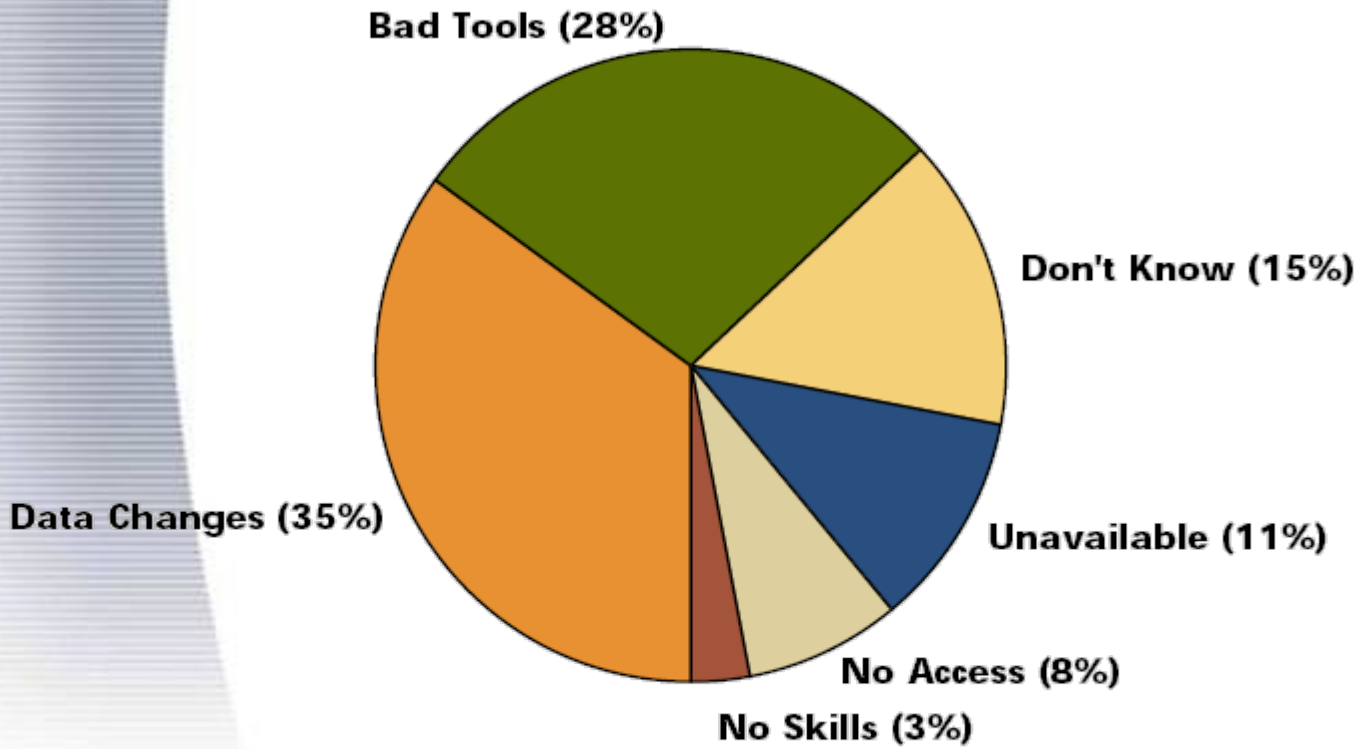
问题背景

- 商业人士每天花费在查找资料上的时间



问题背景

- 无法查找到所需信息的原因



问题背景

- 企业检索市场的情况
  - 2004年美国企业检索市场约为6.2亿美元，增长率约20%
  - 绝大部分市场份额集中在少数公司
    - 国外：Verity，Fast，Autonomy等
    - 国内：TRS，Baidu等
  - 比较欧美而言，国内尚有非常大的发展空间
  - 除企业之外，政府机关、科研机构和媒体都是可能的用户群体

问题背景

- 企业检索的主要产品
  - Autonomy: IDOL server
  - Convera: RetrievalWare
  - FAST: Enterprise search platform
  - Verity: K2 Enterprise and Ultraseek
  - 被称为“the big four”，占据国际市场80%以上的份额
  - TRS: 国内最大的企业信息检索服务供应商，主打产品为TRS Database Server （70%以上占有率）
  - 百度公司：2000年开始涉足企业检索产品，主要产品包括网事通产品系列、企业竞争情报产品系列、数据库检索系统

问题背景

- 软件巨头虎视眈眈
  - IBM: 2005年1月，推出基于WebSphere平台的OmniFind系统
  - Microsoft: 基于Sharepoint Portal和Exchange Server的一系列企业检索产品
  - Oracle: 基于其数据库产品的Secure Enterprise Search(SES) 10g
  - 更多的是在争取传统大客户的企业检索份额，而非一个独立的企业检索产品

问题背景

- 传统搜索引擎供应商跃跃欲试
  - Google:
    - Google企业搜索所使用的技术都是首先为Google公司本身的使用而开发的，随后才推向市场
    - Google mini 
    - Google toolbar for enterprise
    - Google desktop for enterprise
    - Google earth for enterprise
    - Google Search Appliance 
  - AltaVista：与FAST合并
  - Yahoo：与Verity合作推出新一代ES产品

问题背景

- 企业检索的定义

  For any organization with text content in electronic form, enterprise search includes:

  – Search of the organization's external website;
  – Search of the organization's internal website (intranet);
  – Search of other electronic text held by the organization in the form of E-mail, database records, documents on fileshares and the like.

  （David Hawking, 2004）

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 企业信息检索与传统信息检索的主要差别
- 企业信息检索研究面临的挑战

企业信息检索与传统信息检索的主要差别

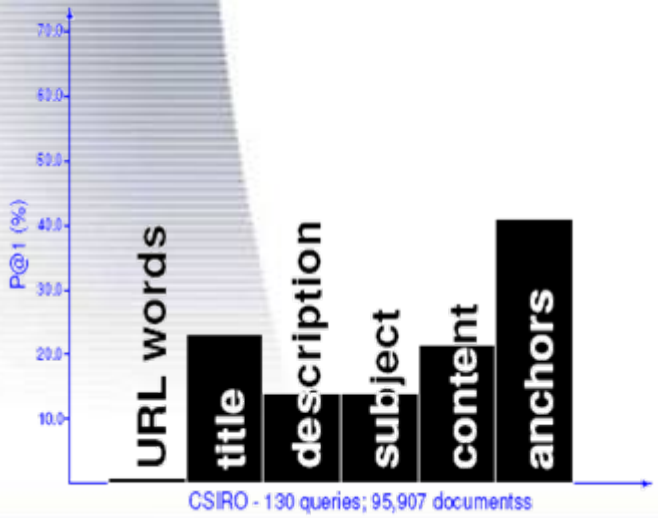
清华大学 Tsinghua University

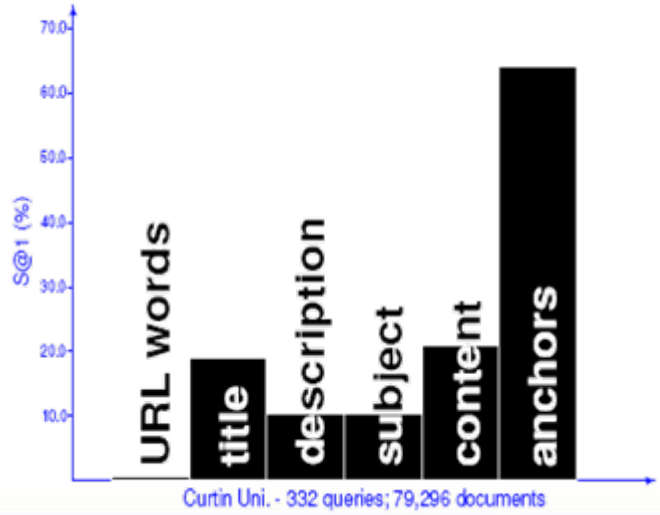
- 企业网络中数据的特点
  - 企业数据仅仅是出于传播信息的需要，而非吸引某些特定用户的目的
  - 大量的查询可能只有个别（甚至是唯一）的正确答案
  - 企业网络中没有作弊信息（Spam-free）
  - 大部分企业网络节点并不利于搜索引擎的数据收集（not search-engine-friendly）

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索需求分析
  - 应用场景1：企业外部人员对企业Web站点的检索。
  - 类似于普通Web检索。
  - 就检索任务而言，大部分是Navigational类型的查询。
  - In-link Anchor text可能在检索中发挥较大的作用。

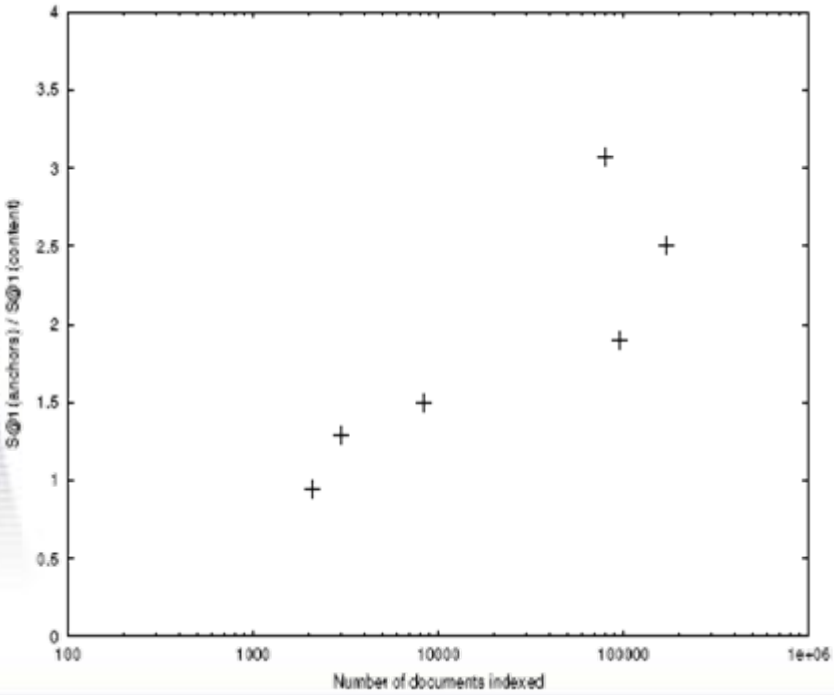




企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

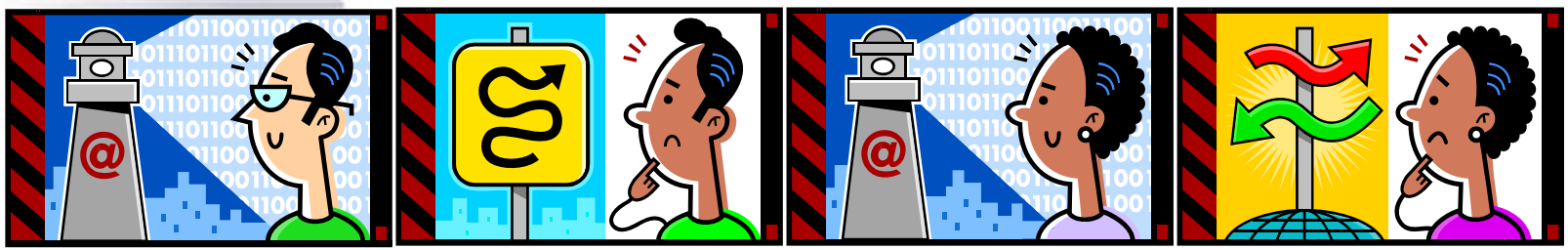
- 企业信息检索需求分析
  - 随着企业Web站点页面规模的增加
    - 主体内容检索的作用降低（扰乱项增多）
    - Anchor的作用变大（可参考项增加）



企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索需求分析
  - 应用场景2：企业局域网的导航类检索
  - 企业内部存在大量的信息与管理资源
    - 各种文档类型
    - 各种基于Web的服务（报销申报、密码更新等）
  - 查询与文档的不匹配现象较少出现
  - 安全性问题是需要考虑的重点



企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索需求分析
  - 应用场景3：企业局域网中的信息类检索
  - 企业信息资源包含的各种格式的信息中相关的部分。
  - 例：某电脑配件公司计划承建清华大学教学实验室基建项目，需要调查本公司之前与清华大学交流的情况
    - 各种工作业绩报告
    - 财务报表
    - 电子邮件
  - 安全性问题
  - 数据异构问题：一致、合理、有效的结果排序

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索需求分析
  - 应用场景4：文档之外的检索需求
  - 除Web页面，各种形式、各种结构化程度的文档之外，企业员工还有其他的检索需求
    - 人员（包括其联系方式）检索
    - 工作组检索
    - 产品检索
  - 与普通的Web检索有较大的差别
  - 安全性问题
  - 数据异构问题：多源、多元数据的融合问题

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业人员检索系统的例子：P@NOPTIC
  - Designed by CSIRO
  - The system automatically identifies experts in an area, based on the documents already published on an organization's intranet.
  - Can be queried like a standard Web search engine

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- P@NOPTIC原理
  - Staff list
    - Up to date （List之内的人员方可被检索）
    - Include employee contact details and home page URLs （结构化信息，反馈给用户）
  - Employee document
    - One employee is corresponding to one document
    - All document described the employee is put into the document
    - Simply the concatenated text of the related documents

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- P@NOPTIC的启示
  - 异构数据结果反馈方式
    - Document, staff home page
    - Contact information
  - 异构信息集成方式
    - 以查找目标为中心的信息组织形式
  - 用户使用方式
    - 尽量接近现有的Web检索（simple query）
  - 真实商品化：http://funnelback.com/

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索需求分析
  - 应用场景5：依据法律内容进行的相关搜索
    - 专利诉讼
    - 债务情况分析
    - 破产/坏账原因调查
  - 强调查全率的检索
  - 检索文档为主，但方法上与传统的网络信息检索（强调准确率）有较大不同

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业检索一个综合的应用场景
  - M公司的销售经理小A计划向T公司销售一批产品，则一个完善的企业检索系统应当向他提供如下信息：
    - 从Web上搜集的关于T公司的概要情况
    - T公司的联系方式，以及当权人物的列表
    - 与T公司相关的产业分析与股票市场报告
    - M公司与T公司相关的交易情况
    - M公司与T公司相关的电子邮件往来情况
    - M公司当前与T公司有关的员工列表
  - 一个完善的企业检索系统，应该是企业信息管理的核心，与决策的重要辅助

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索与传统网络信息检索方式的差别
  1. 多数据源
     - Storage: File share systems, Web servers, Lotus Notes, Microsoft Exchange Servers, Database systems
     - Format: document, presentation, pdf, ps, html
  - 要求：
     - 多重数据收集接口
     - 分布式检索架构
     - 数据格式转化
     - 归一化的索引结构设计

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索与传统网络信息检索方式的差别

  2. 安全性与个人化需求

  – 安全性需求是企业信息检索系统必须具备的功能
    - Google: secure APIs
    - Oracle: secure enterprise search (SES)
  – 利用安全性管理的契机，可以为不同用户提供个人化检索服务
    - 根据用户个人信息和以往检索行为
    - 比Web信息检索更有条件实现
  – 要求：
    - 索引、用户交互等各个级别支持的权限管理
    - 用户行为的挖掘与反馈

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索与传统网络信息检索方式的差别

  3. 不同结构化数据的分析与集成

  – Google mini: The appliance connects with Oracle's database, IBM's db2, Microsoft's SQL, the open-source mySQL and Sybase's database.

  – 不同层次的集成：

    - 结果层次：针对某个特殊的检索词，把各种结构化数据相对应的检索结果进行集成
    - 检索对象层次：以某些特殊的对象为中心，将各种结构化数据组织成对这个对象的描述

  – 要求：充分利用不同结构化数据检索的已有研究成果

企业信息检索与传统信息检索的主要差别

清华大学 Tsinghua University

- 企业信息检索与传统网络信息检索方式的差别
  - 已有的研究结果大都从企业管理运作以及工程实现的角度来给出企业检索系统的定义
    - Stenmark (1999), Abrol (2001)
    - Infonortics Search Engine Meeting
  - 如何从IR的角度，对企业信息检索（或称企业级信息检索）面临的挑战进行分析？
    - 企业检索的目标，是以某种形式反馈给用户（企业允许范围内）对他最有用的检索结果。
    - 第三代搜索技术最可能的发源地
      - 满足用户“查询背后的需求”
      - 数据质量、天然的个性化检索需求

网络信息检索技术的现状与挑战

清华大学
Tsinghua University

- 问题背景
- 企业信息检索与传统信息检索的主要差别
- 企业信息检索研究面临的挑战

企业信息检索研究面临的挑战

- David Hawking在2004年的Challenges in Enterprise Search一文中提出了7项企业检索领域涉及的IR关键问题
- 对这些问题进行综合，结合企业检索中出现的新问题，我们提出如下几点企业检索研究中可能的关键挑战，并对其进行一些分析
  - 企业检索研究数据的获取问题
  - 企业检索系统的结构设计问题
  - 企业检索用户模型的建立问题
  - 多元数据的检索与评价问题

企业信息检索研究面临的挑战

- 企业检索研究数据的获取问题
  - 实验数据平台的重要性和目的：调整算法，考察产品
  - 数据建设的需求：
    - 对相当数量企业所管理的数据现状的考察
    - 对企业用户完整的信息需求的了解
  - 不仅仅是多种类型数据的堆积
    - 数据间需要包含内容上的联系
    - 基于数据能够满足真实的企业检索需求
    - 涉及到多种法律、经济方面的困难，获取某些破产公司的信息可能更为现实

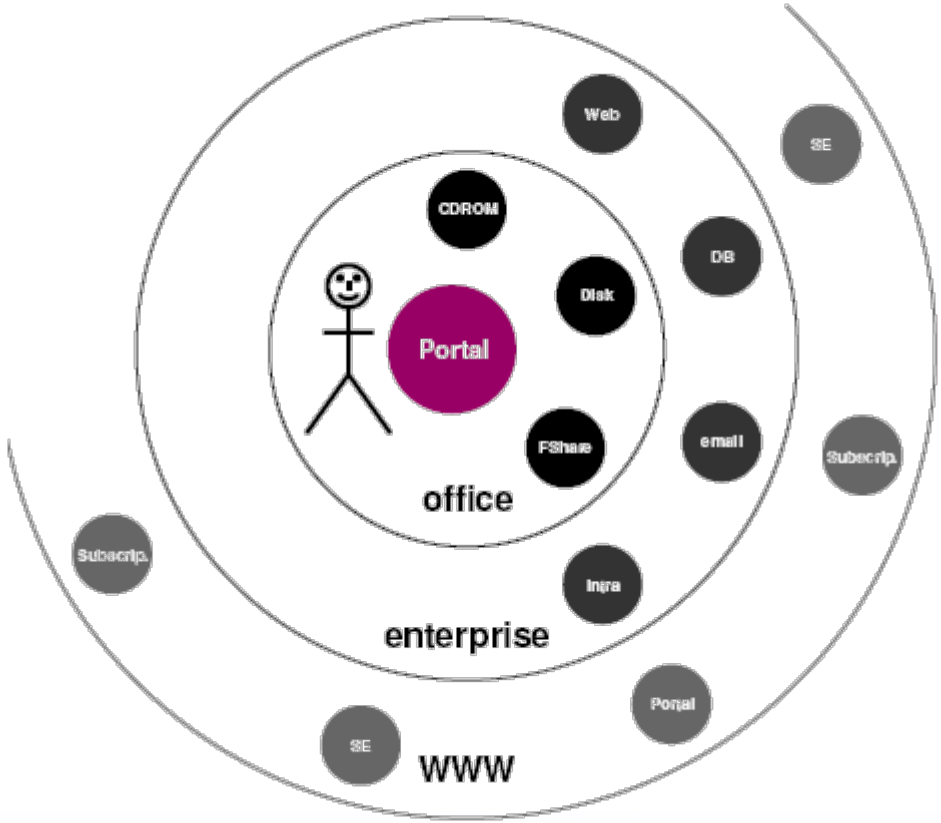
企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 企业检索研究数据的获取问题
  - 理想的构成：
    - External web sites
    - Internal web sites
    - XML document extracted from database, email, word processing, presentation, spreadsheet…
    - Realistic (or ideal) enterprise information/service needs
    - 以上都需要覆盖某个企业绝大部分的真实数据
    - 每类数据包含10G以内的规模量级

企业信息检索研究面临的挑战

- 企业检索系统的结构设计问题
  - 企业用户眼中的信息层次

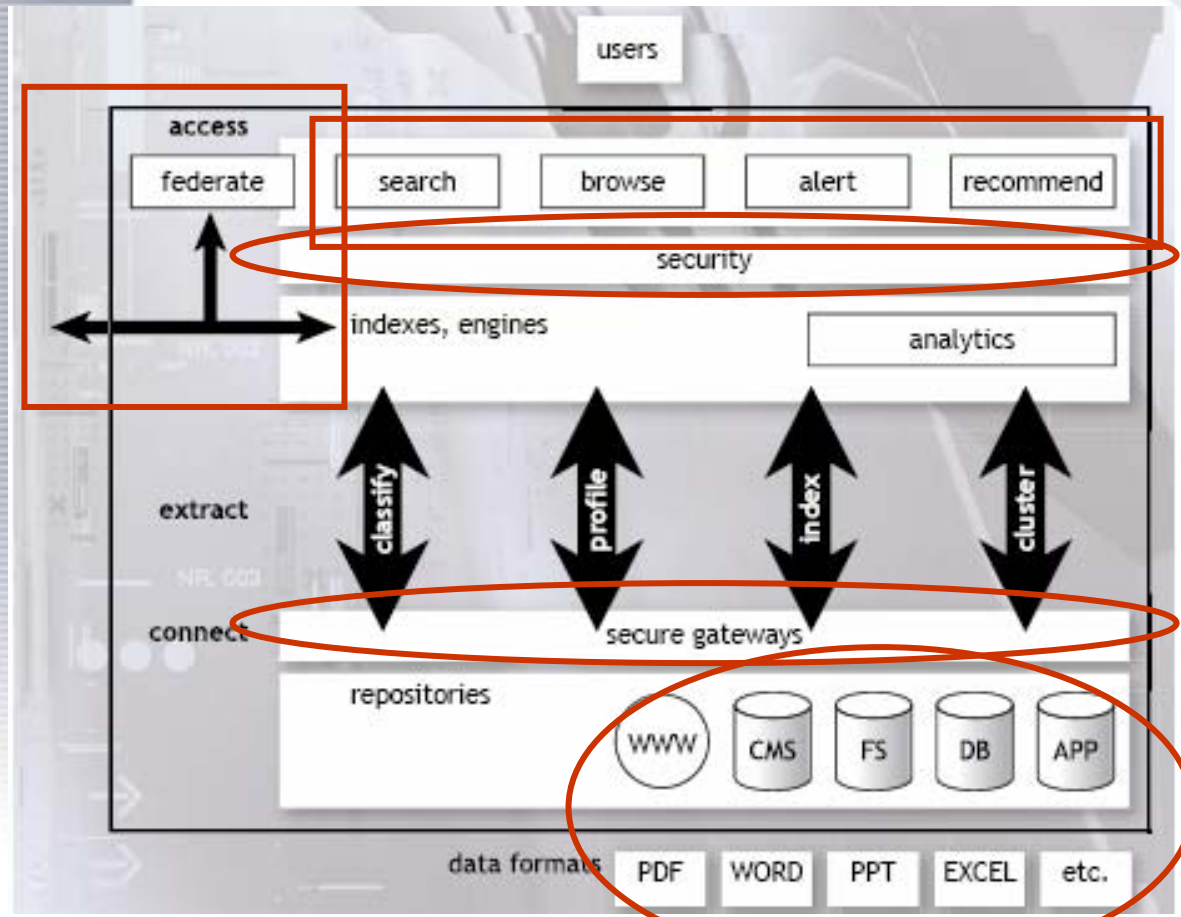


企业信息检索研究面临的挑战

- 企业检索系统的结构设计问题
  - 理想：员工的个人化信息入口（personalized employee portal）
    - 分布式检索的问题
    - 多元数据检索的问题
    - 检索安全性的问题
    - 个人化检索的问题

企业信息检索研究面临的挑战

- 典型的企业检索系统设计



企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 企业检索用户模型的建立问题
  - 企业检索模型与普通检索模型的相异处
    - 检索任务的差别
      - 导航类检索（大部分只有唯一答案，较少存在查询与文档的不匹配现象）
      - 查全性检索（法律需要，避免重复劳动的需要等）
    - 个人化带来的差别
      - 检索“上下文”利用的可能性
      - 背景知识：地理位置、用户档案、检索日志等
      - 如何集成这些相关知识？

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 企业检索用户模型的建立问题
  - 检索模型构建的信息来源
    - 搜索日志可能提供的有用信息比较有限：取决于当前的企业检索能否很好的满足用户的需求
    - 即使能够较好的满足用户需求，但用户行为也存在着歧义性
  - 用户行为的歧义性
    - 用户不会在进行某次检索之后对这次检索的满意程度进行评价
    - 用户点击了某个网页，就代表他满意这条检索结果么？

企业信息检索研究面临的挑战

- 企业检索用户模型的建立问题

  可能的步骤：
  - 明确企业用户的信息需求
  - 获取有参考价值的实验数据
  - 建立与用户信息需求相对应的检索性能评估方式和标准评测平台
  - 学习可以投入企业检索应用的有效算法和模型

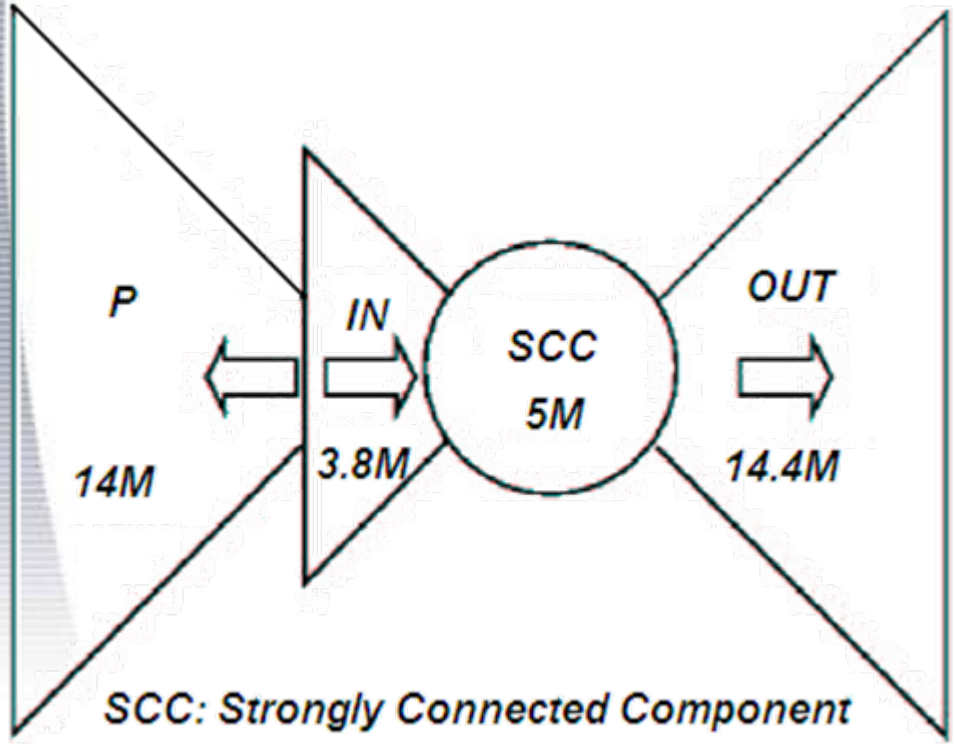
企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 多元数据的检索与评价问题
  - 数据质量评估问题
    - 企业中大部分数据缺乏网络数据固有的链接关系，导致网络信息检索中成熟的链接分析算法部分失效。
      - 即使具有链接关系的部分，其网络拓扑结构也有较大改变，主要源于企业检索中较少的互链接关系。
    - 事实上，大部分可以基于非内容特征挖掘的方法在企业检索中均部分失去了效用
    - 即使是企业数据中的Web文档，他们之间的链接关系与万维网页面也存在着较大的差别

企业信息检索研究面临的挑战

- 多元数据的检索与评价问题
  - 互联网与局域网的链接分布情况



企业信息检索研究面临的挑战

- 多元数据的检索与评价问题
  - 网络检索的成功经验：search and browse
    - 依赖于用户对于信息的再加工
    - 缺乏链接的数据无法支持这种信息获取模式
  - 链接关系自动构建
    - 使用内容管理系统（数据库系统，个人信息管理软件等）提供的信息自动生成链接
    - 是否可行？
    - 如果可行，链接分析算法在多大程度上可以使用？

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 多元数据的检索与评价问题
  - 数据相关性评估问题
    - 企业数据面临的多元化数据
      - 纯文本数据：文本，代码等
      - 半结构化数据：HTML, XML等
      - 结构化数据：表格，数据库数据，企业中的Deep Web等
      - 二进制数据：文本数据的二进制表示，流媒体等
    - 各种数据类型具有独特的处理方式，如纯文本信息检索，网络信息检索，数据库检索等
    - 但不同结构化数据之间的相关性如何比较？

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 示例1：如何将各种结构数据结合以向用户反馈结果
  - 应用假设：购买《纳尼亚传奇》
    - 各个购物站点的报价、版本信息（结构化），某些论坛的二手货相关帖子
    - 作者刘易斯的个人介绍、书籍的介绍（半结构化）
    - 与《纳尼亚传奇》相关的电影信息、其他书籍信息等（半结构化及结构化）
  - 如何组织这些信息，又应该是一个怎样的顺序反馈给用户？

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 示例2：对于二进制数据文件检索而言，如何将不同结构化的信息结合成对某个文件的统一描述
  - 应用假设：MP3检索“光良《约定》”
    - Mp3文件本身记录的元信息，艺术家，歌名乃至歌词等（结构化）
    - Mp3文件所在网页的信息（半结构化）
    - Mp3文件所在站点的信息（结构化）
  - 如何统一利用这些信息对这个Mp3文件给出合适的顺序？

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 多元数据的检索与评价问题
  - 对于检索而言，数据多元性的表现：
    - 结构是否清晰
    - 查询对象长度分布
      - 数据的条目长度相对恒定，而word文档长度则差别巨大
    - 是否存在链接关系
    - 文档间的相互关系
      - 站点主页与其他页面的关系，文档间的时序关系等
    - 文档中的内容重复问题
    - 语言使用问题
      - Xls表头的文字风格肯定异于正常的文字

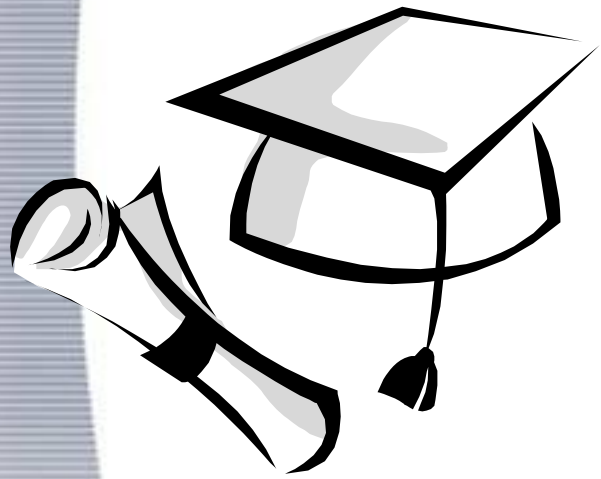
企业信息检索研究面临的挑战

- 多元数据的检索与评价问题
  - 可能的解决方式
    - 寻找各种数据形式的统一检索方式
      - 针对诸如文档长度分布不统一这样单独的问题，已经有一定的研究成果（Singhal, 1995），但对于如此复杂的多源数据处理问题，还缺少相应的处理办法
    - 各种数据形式分别处理，寻找一种合适的结果综合方式
      - 结果合并是一个困难的问题，而且很难达到一个统一检索方式所能达到的效果（Voorhees, 1995）
      - 多种不同形式的检索结果进行合并更是如此

企业信息检索研究面临的挑战

清华大学 Tsinghua University

- 总结
  - 相比网络搜索引擎的巨大成功而言，企业信息检索现阶段的发展差强人意
  - 尽管我们指出诸多的不利之处，但我们同样发现企业检索具有其相对有利的方面
    - 相对小规模的数据可能更加有利于一些理论上相对成熟技术的应用。如NLP技术（如自动分类技术、自动文摘技术、机器翻译技术）
    - 企业相对固定的用户群带来的个人化检索的可能
  - 企业检索可能正面临着一个跳跃发展的时期



Thank you!

Questions or comments?