基于可视化分析的多媒体语义挖掘

罗远哉

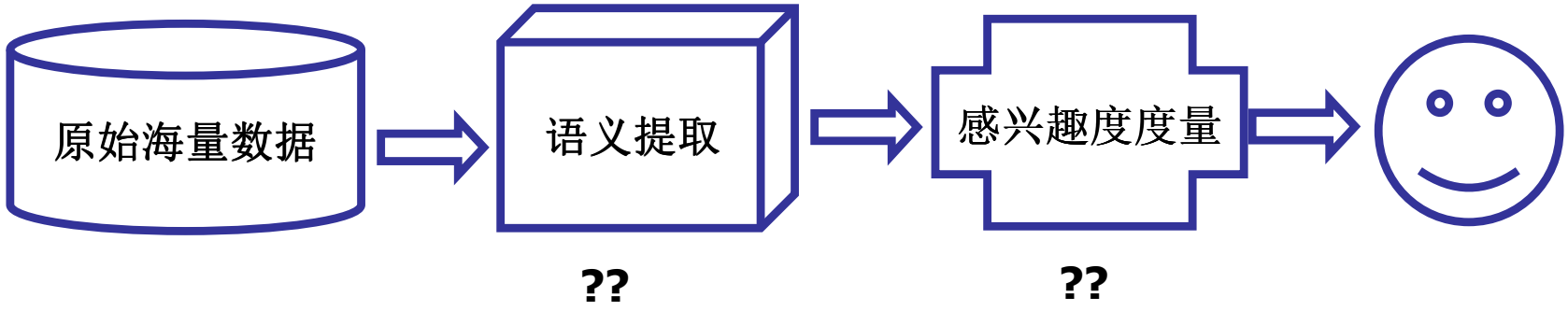
华东师范大学

研究动机

- “数据灾难”
  - 数据已经极大丰富
    - 中文WEB：>75TB；Internet: >520PB
      - 哪篇是今天最有意义的新闻？
    - 地质、海洋勘探数据很容超过1TB
      - 哪里会发生地震、赤潮？
    - 城市视频监控数据很容易超过1PB
      - 哪个是犯罪现场？
  - 但是我需要的在哪里？

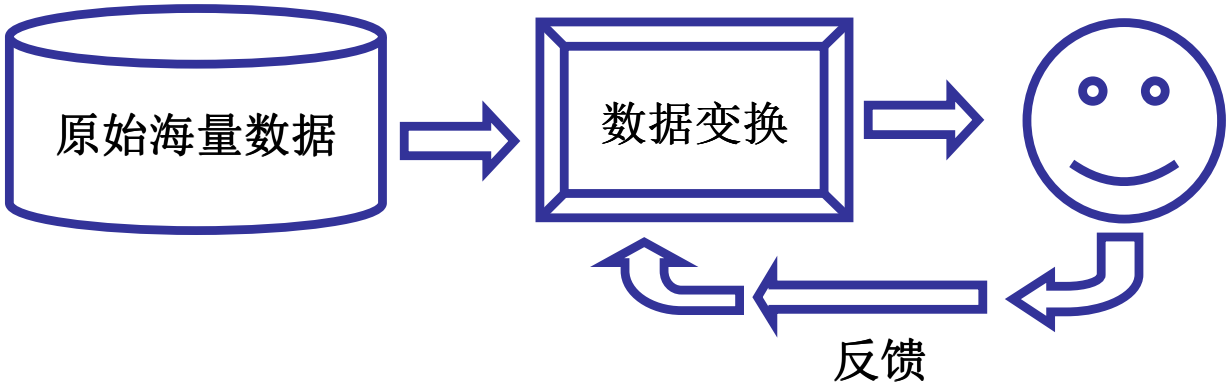
研究动机

- 直接提取语义并进行筛选
  - 使用复杂的算法，试图让计算机象人一样理解原始数据的语义，并按照人的喜好和需求进行过滤
    - 终极解决方案，但难度极大



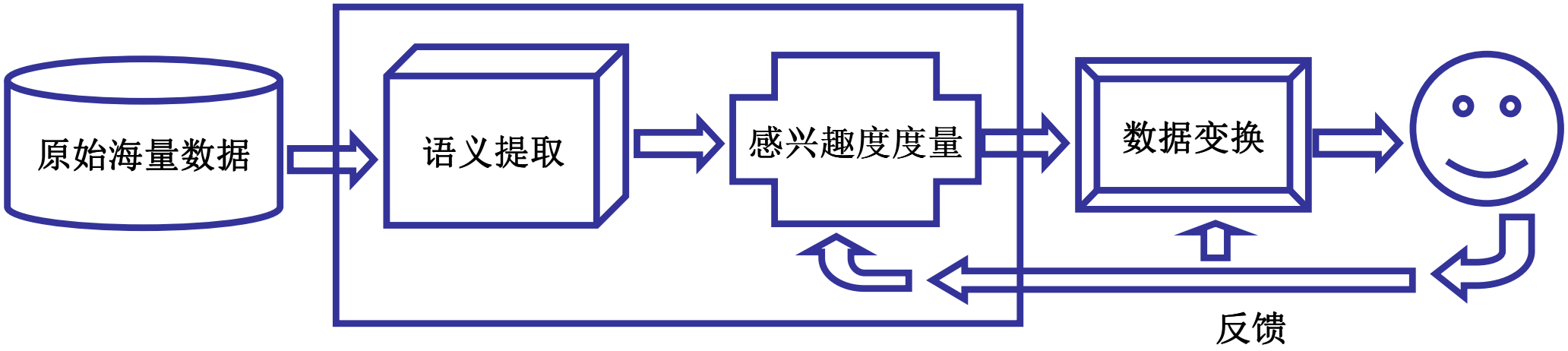
研究动机

- 可视化及可视化分析技术
  - 把原始数据变换成易于查看的图像、动画
    - 通过视觉这个最高速的信息交互通道与人交流
    - 很难找到好的数据变换公式
      - 最好的变换是语义提取



研究动机

- 结合语义提取与可视化分析
  - 提取较容易获得的初等语义，适当过滤
    - 以适合可视化为目的
  - 在初等语义上可视化，并提供反馈手段
    - 可以反馈修改感兴趣度度量算法



研究动机

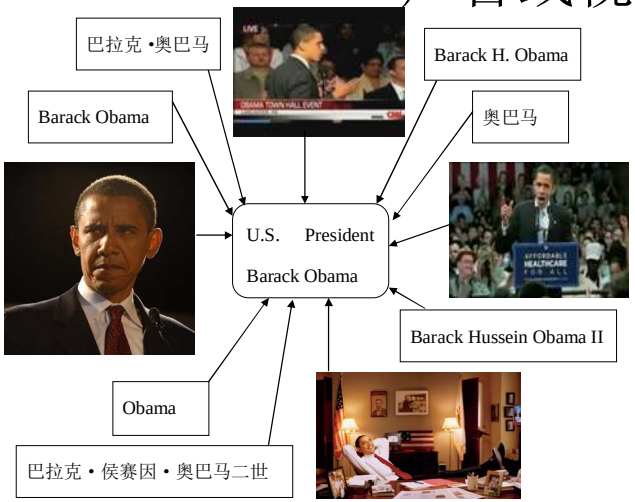
- 需要解决的问题
  - 如何表示和提取适合可视化的语义？
    - 语义模型及提取算法
  - 如何过滤语义，获得用户感兴趣的部分？
    - 感兴趣度度量算法
  - 使用什么样的可视化技术？
- 不同领域的数据可能需要完全不同的算法
  - 多媒体新闻分析与可视化

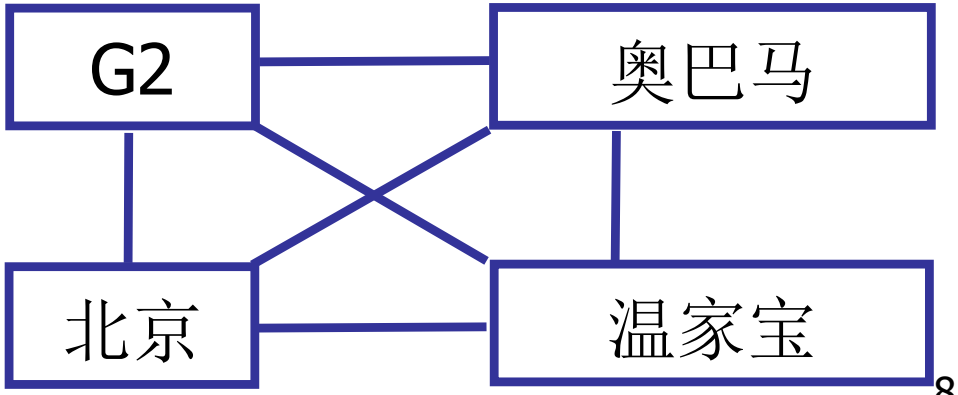
多媒体语义提取

- 传统方法
  - 语义分类
    - 由于存在语义鸿沟，极其困难
  - 视觉词袋(Bag of Words)模型
    - 把视觉特征聚类成码书，用码书项作为视觉关键字，再使用文字检索类似的技术
    - 码书项并不直接表示语义
    - 适合于某些视觉检索

多媒体语义提取

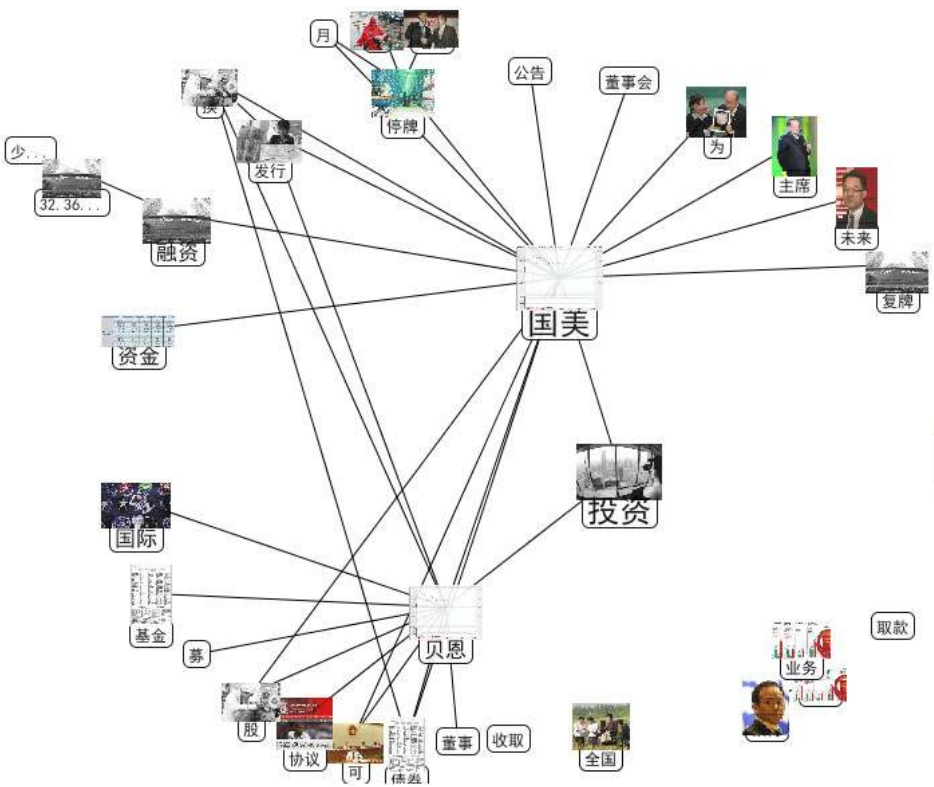
- 语义实体-关系模型
  - 以语义实体为节点、实体间的关系为边的图结构
    - 语义实体：具有固定语义的实体
      - 根据需要，语义实体可能被表示成不同的文字、图像、声音或视频

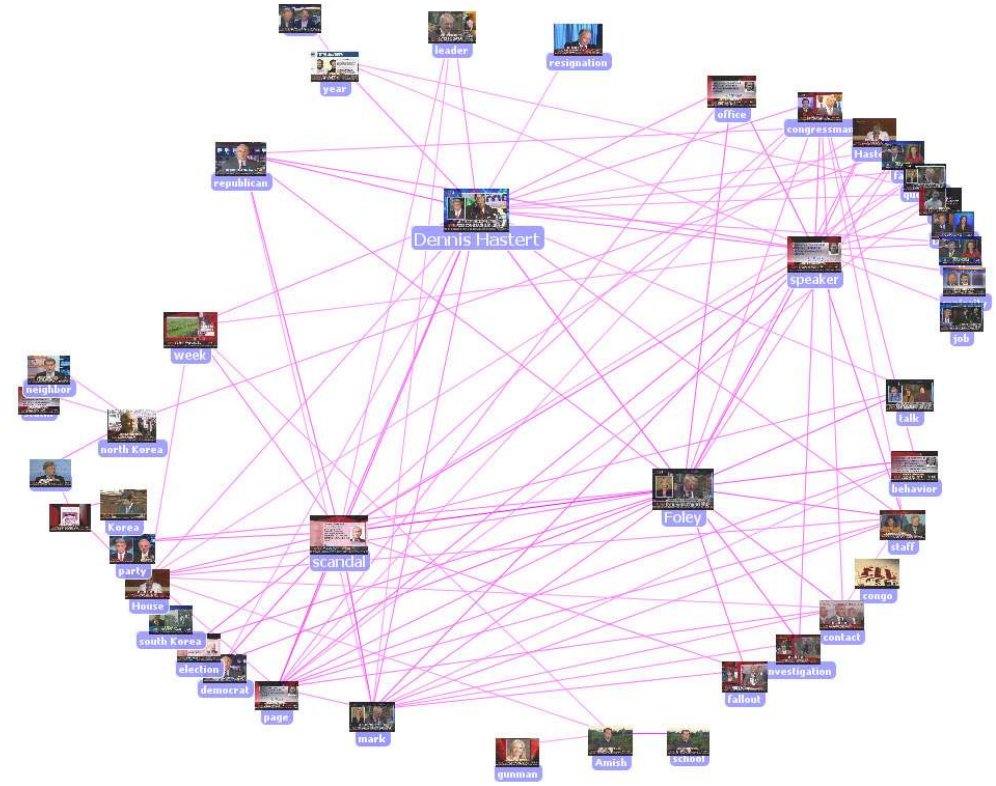




多媒体语义提取

- 语义实体关系模型例





多媒体语义提取

- 语义实体关系模型优点
  - 保留更多信息
    - 多媒体信息
    - 关系信息
  - 可表示一个文档集合的摘要
    - 保留跨文档信息
  - 大量成熟的图算法可以直接使用
  - 如何提取?

多媒体语义提取

- 语义实体关系模型提取
  - 第一步：跨媒体关联
    - 把多媒体文档和文字内容进行关联
      - 语义提取，OCR，ASR
        - 虽然只能提取少量语义，但因为直接来自视觉信息，仍然有很大帮助
      - 视觉相似性聚类
      - 元文字
        - 文件名、链接文字、临近文字……

多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 序列标注算法 (CRF, HMM)
      - 新实体检测准确率较低
        - 从来没有在训练集中遇到
    - 在序列标注算法中集成统计特征
      - 实体文字串的统计特征与非实体串显著不同
        - 内部紧密，外部松散，头尾特别

美称黄海军演尚在计划

美航母赴黄海军演对中国来说到底是喜还是忧

美韩在黄海军演是个不智之举

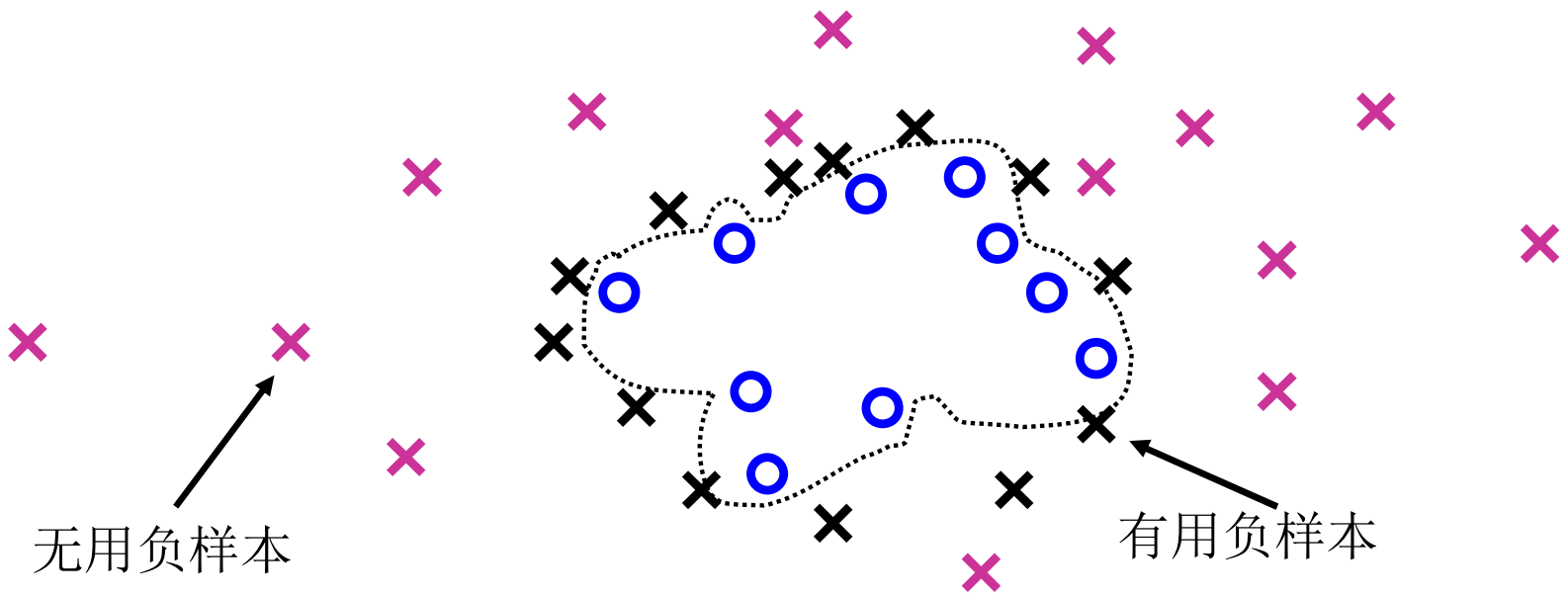
美国在黄海军演要吓唬谁

多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 提取统计特征运算量很大
      - 需要对大量数据进行统计
    - 绝大多数文字串不是实体
      - 6词窗口
        - 人民日报：0.4%
        - BNC XML：0.02%

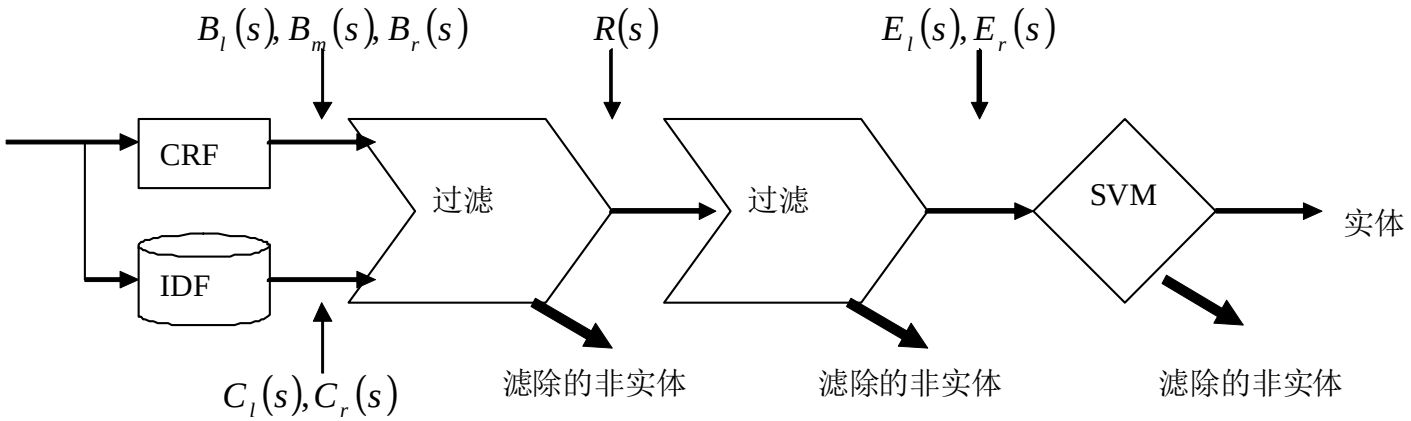
多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 负样本过滤
      - 绝大多数负样本对分类训练没有好处



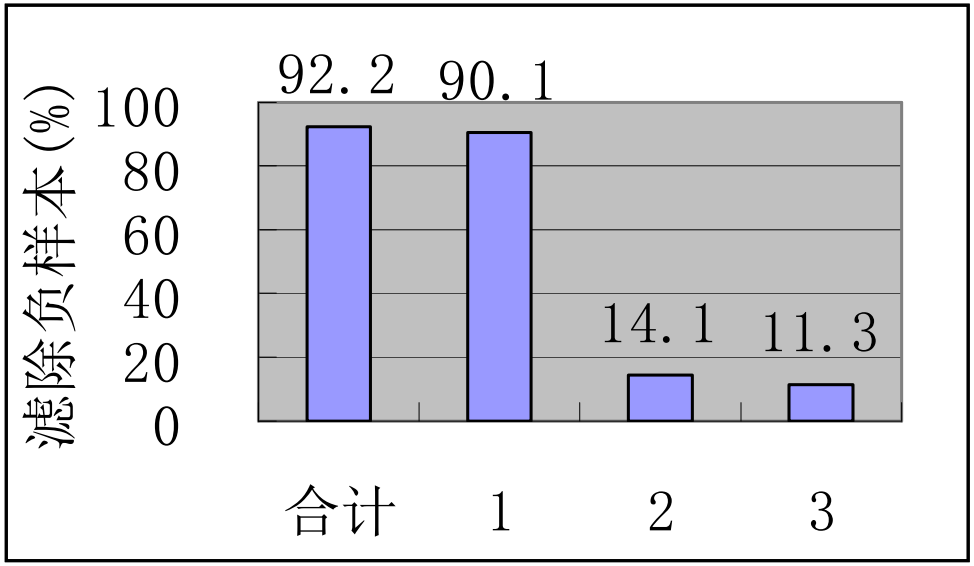
多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 分步过滤
      - 先使用计算量小的特征，用高召回率算法过滤去除大部分负样本，后续步骤只处理过滤剩下的样本

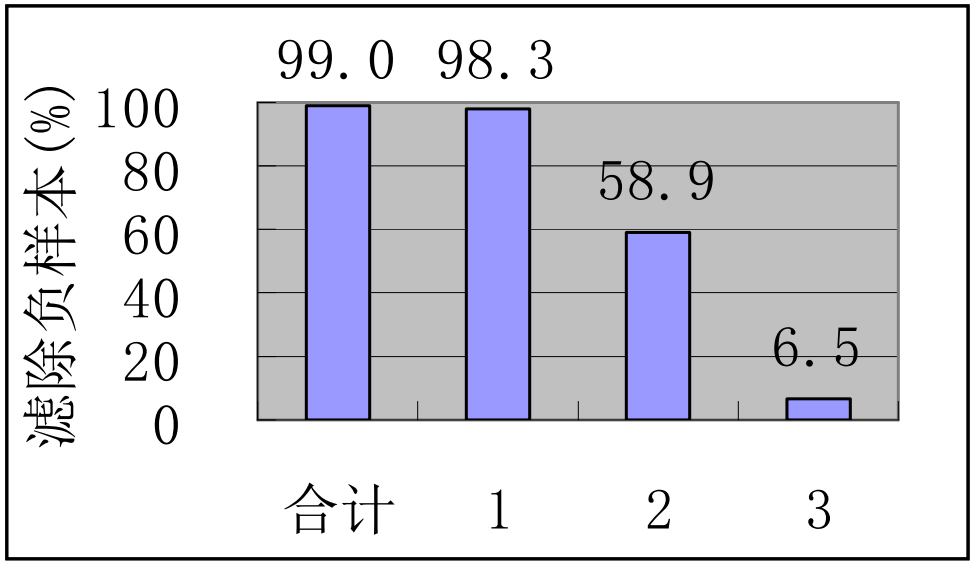


多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 分步过滤



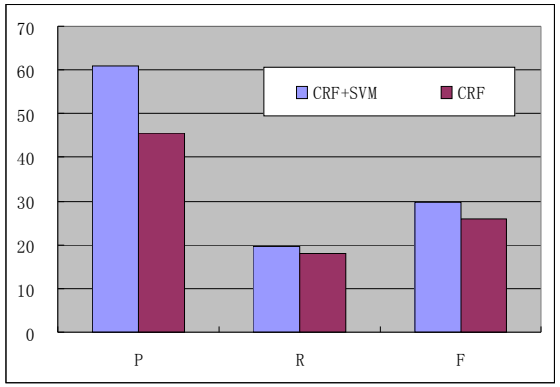
人民日报



BNC XML

多媒体语义提取

- 语义实体关系模型提取
  - 第二步：文字实体提取
    - 集成CRF和统计特征进行SVM分类
    - ACM Multimedia Grand Challenge 2009: “Extracting Informative Images from Web News Pages via Imbalanced Classification”



BNC XML

多媒体语义提取

- 语义实体关系模型提取
  - 第三步：名义实体聚类
    - 一个实体可能有不同名字
      - US, USA, Washington Administration
    - 一个名字可能表示不同实体
      - Clinton: Hilary Clinton, Bill Clinton
    - 多模EM聚类

文字　　关系　　视觉

$$s(e_i,e_j)=\sum_{l=1}^{n} w_l \alpha_l(e_i,e_j)+w_{n+1}\beta(C(e_i),C(e_j))+w_{n+2}\gamma(V(e_i),V(e_j))$$

$$\sum_{l=1}^{n+2} w_l = 1$$

多媒体语义提取

- 语义实体关系模型提取
  - 第四步：关系提取
    - “同时出现”的实体对即为有关系
      - 在同一个句子中
      - 在同一个视频镜头中
      - 在同一个图像中
      - ……

多媒体语义提取

- 语义实体关系模型提取
  - 问题：所有的实体/关系都一样吸引人吗？
    - 舟曲<>汶川 2010年8月
      - 舟曲：高得多的词频
      - →把词频作为实体和关系的权重
        - 加权图 (原始语义实体关系模型)
    - 舟曲<>北京 2010年8月
      - 词频似乎不太可靠？

如何度量用户的内在感兴趣度?

感兴趣度度量

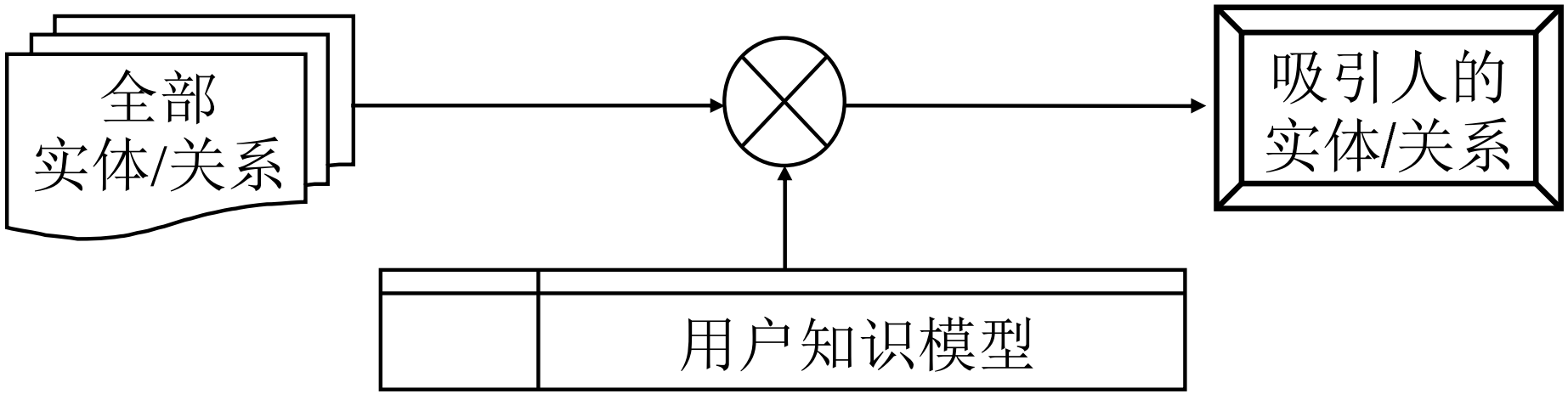
- 究竟哪些实体/关系更有吸引力?
  - 用户不知道的
    - “狗咬人不是新闻，人咬狗才是”
    - 正常<>异常
      - 风险管理
      - 安全监控
      - 企业管理
      - ……

已知

未知

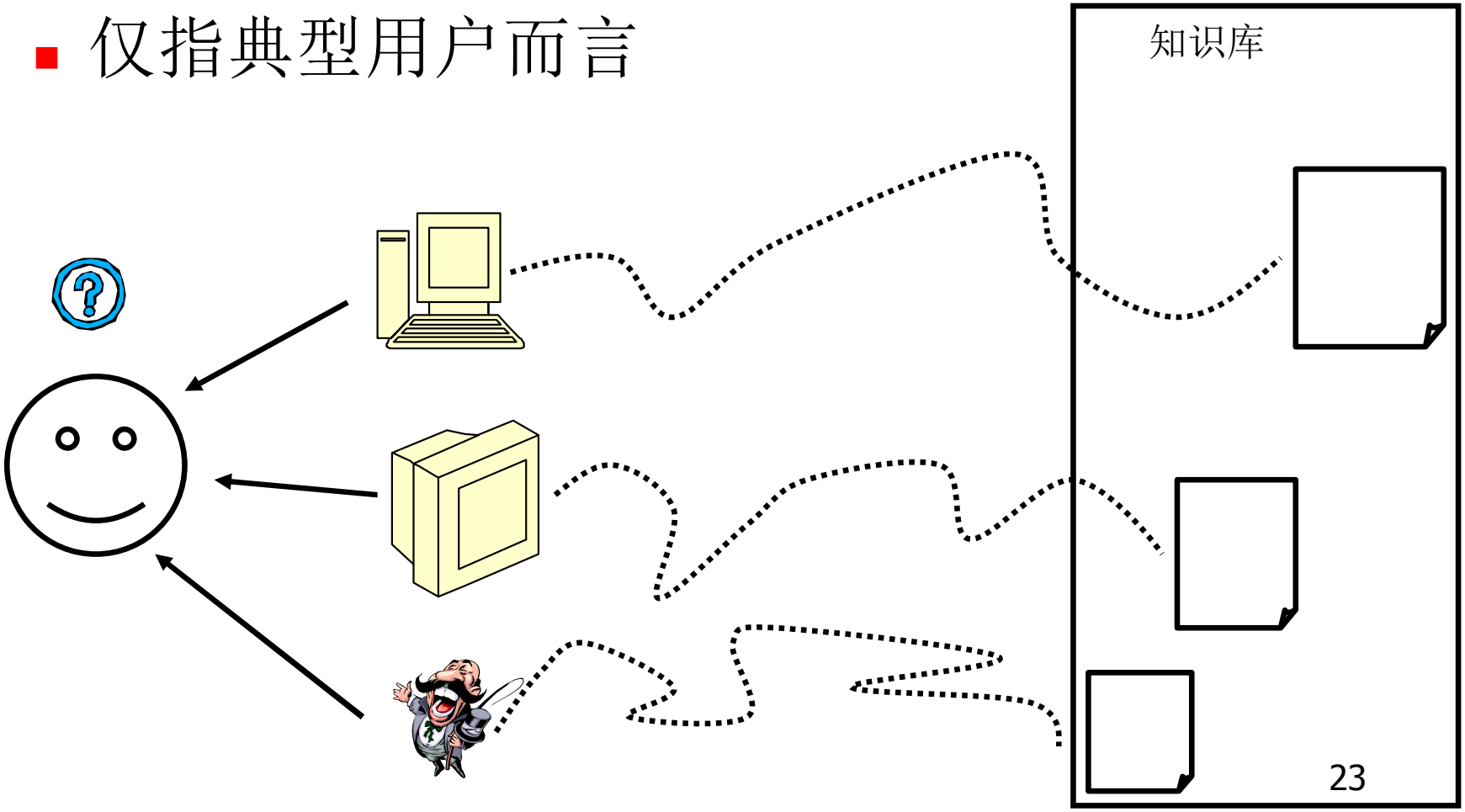
感兴趣度度量

- 究竟哪些实体/关系更有吸引力?
  - 用用户知识模型来"预测"
    - 难以预测中的更吸引人
    - 前提：准确提取用户知识模型



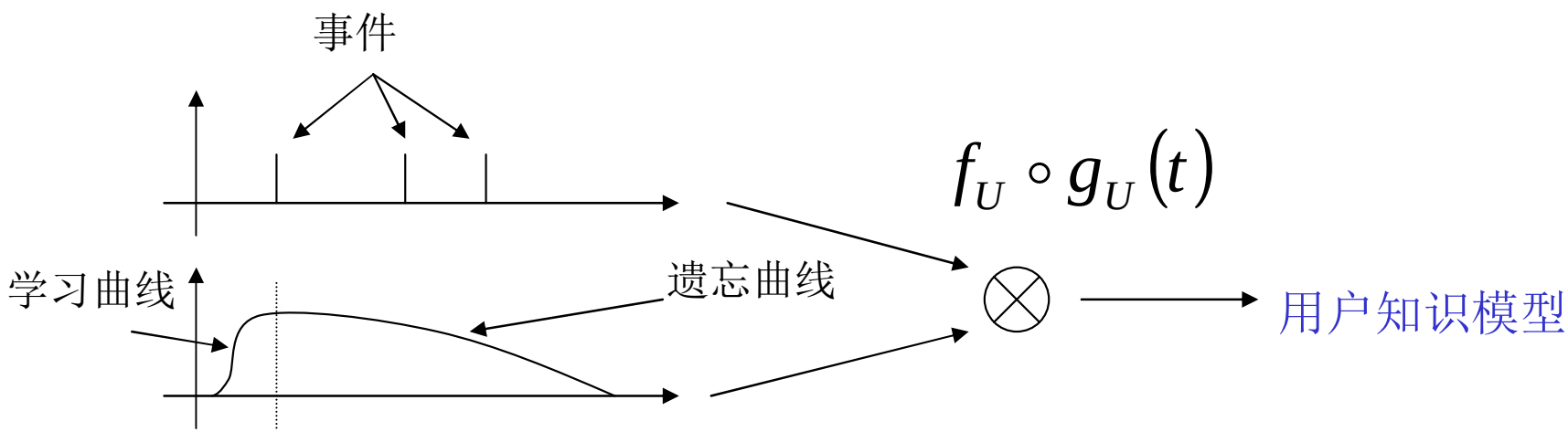
感兴趣度度量

- 用户知识模型可以用所有公开数据近似
  - 仅指典型用户而言



感兴趣度度量

- 用户知识模型
  - 事件和学习遗忘曲线的卷积
    - VAST 2007: “Analyzing Large-Scale News Video Databases to Support Knowledge Visualization and Intuitive Retrieval”



感兴趣度度量

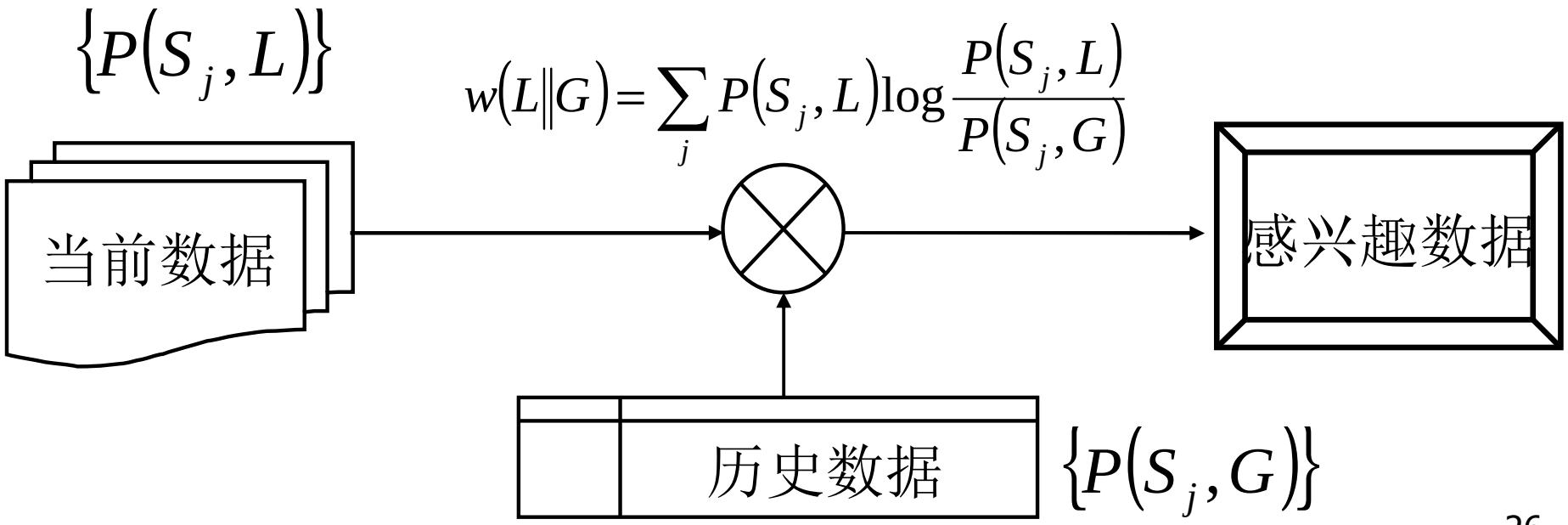
- 用户知识模型
  - 个性化
    - 个性化学习遗忘曲线
    - 个性化频道影响因子
    - 基于语义实体和关系的检索
      - 兴趣传递

$$K'_D(s_I) = \left\{ \left(k_i, w_U(k_i) \times c^{\max\{r(s_a, s_I), r(s_b, s_I)\}} \right) \right\}$$

      - Signal Processing: Image Communication (doi:10.1016/j.image.2008.04.014): Integrating multi-modal content analysis and hyperbolic visualization for large-scale news video retrieval and exploration

感兴趣度度量

- 用户知识模型
  - 预测: KL-divergence
    - VAST 2006: “Exploring Large-Scale Video News via Interactive Visualization”

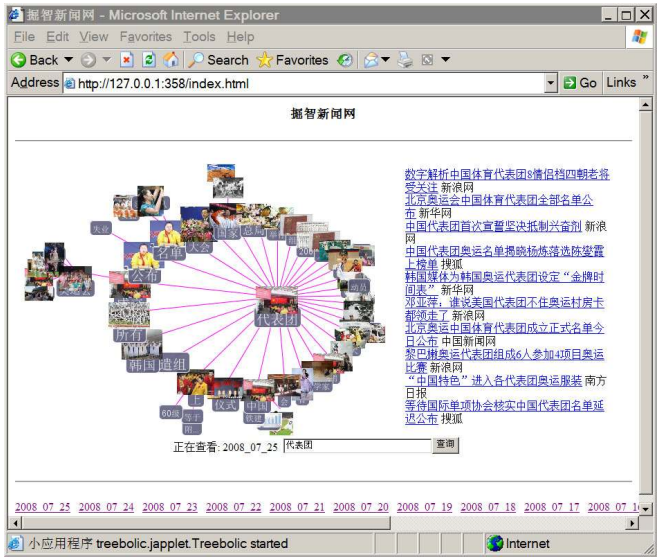


可视化推荐

- 呈现语义实体关系模型给用户
  - 浏览与检索
    - VISUAL 2007: “Mining Large-Scale News Video Database via Knowledge Visualization”

$$\hat{A}(k_i) = \arg\max_{A(k_i)} \left\{ \sum_{k_i} \left(A(k_i) \times w_U(k_i) \right) \right\}$$

$$\Rightarrow \begin{cases} \hat{A}(k_i) = \begin{cases} B(k_i) & i \le \gamma \\ 0 & i > \gamma \end{cases} \\ \forall (i < j): w_U(k_i) < w_U(k_j) \\ \sum_{i=1}^{\gamma} B(k_i) \le c, \sum_{i=1}^{\gamma+1} B(k_i) > c \end{cases}$$



计算架构

- 海量计算
  - 分类器训练
  - 语义分类
  - 跨媒体关联
  - 语义实体关联、聚类
  - 用户知识模型提取
  - ……

计算架构

- 海量计算
  - 特点
    - 需要处理海量数据
    - 数据可以分割成海量的可并行运算单元
    - “读数据，写数据”
      - 计算节点之间不交换数据
  - 高性能计算模型(HPC)不适合
    - 太先进，太昂贵
    - 缺少海量数据存储与管理工具

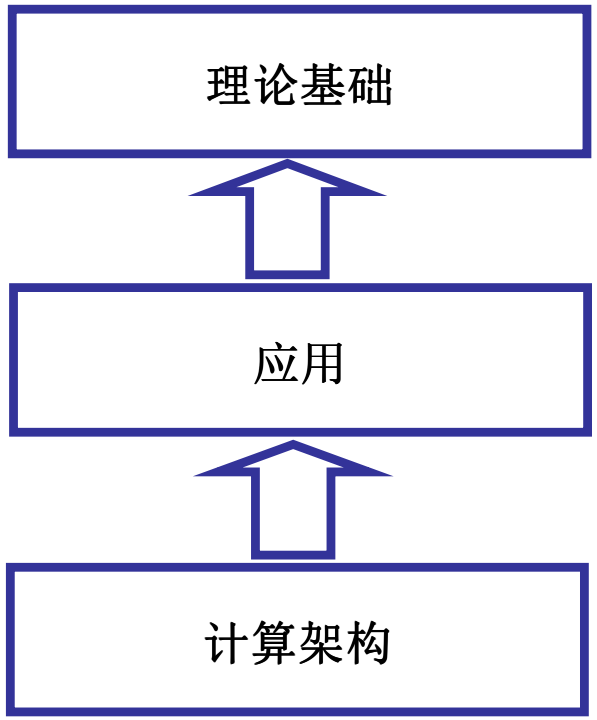
计算架构

- 现有海量计算模型
  - Google (较适合类似文字索引任务)
    - GFS, MapReduce, Bigtable
  - Yahoo (与Google模型类似)
    - Hadoop, HBase
  - Amazon AWS (更象大型网站开发工具)
    - EC2, S3, SQS, SimpleDB, Alexa, ……
  - Wikipedia, Flickr, Facebook……
    - 针对相应业务流程优化，难以运行更复杂任务

计算架构

- ParaQueue
  - 并行、分布式队列
    - 高吞吐率
    - 高可靠性
  - 队列上的集合操作
    - “读队列，写队列”
  - 部分数据库功能
    - 索引、搜索

总结



机器学习: **CRF-SVM, McEM**
语义提取: 语义实体关系模型

可视化推荐

ParaQueue云计算平台

系统演示

谢谢!

请提问