机器学习在搜索中的若干应用

百度 王海峰

2010年11月6日

MLA10，南京

内容提要

- 百度搜索技术概况
- 机器学习在需求分析中的应用
  - 特征表示
  - Term重要性计算
  - 实体属性识别
  - Query分类
- 其它应用
- 结束语

内容提要

- 百度搜索技术概况
- 机器学习在需求分析中的应用
  - 特征表示
  - Term重要性计算
  - 实体属性识别
  - Query分类
- 其他应用
- 结束语

百度产品

用户产品

商业产品

Baidu 百度

百度用户产品

用户通过百度主页，可以瞬间找到相关的搜索结果， 这些结果来自于百度超过百亿的中文网页数据库。

网页搜索

提供更加完善的搜索体验，满足的多样化的搜索需求

Baidu MP3 Baidu 图片 Baidu 视频 Baidu 地图 测试版

垂直搜索

搜索技术

提供表达和交流思想的自由网络空间

Baidu 贴吧 Baidu 知道 Baidu 空间 Baidu 百科 Baidu 文库 测试版

社区产品

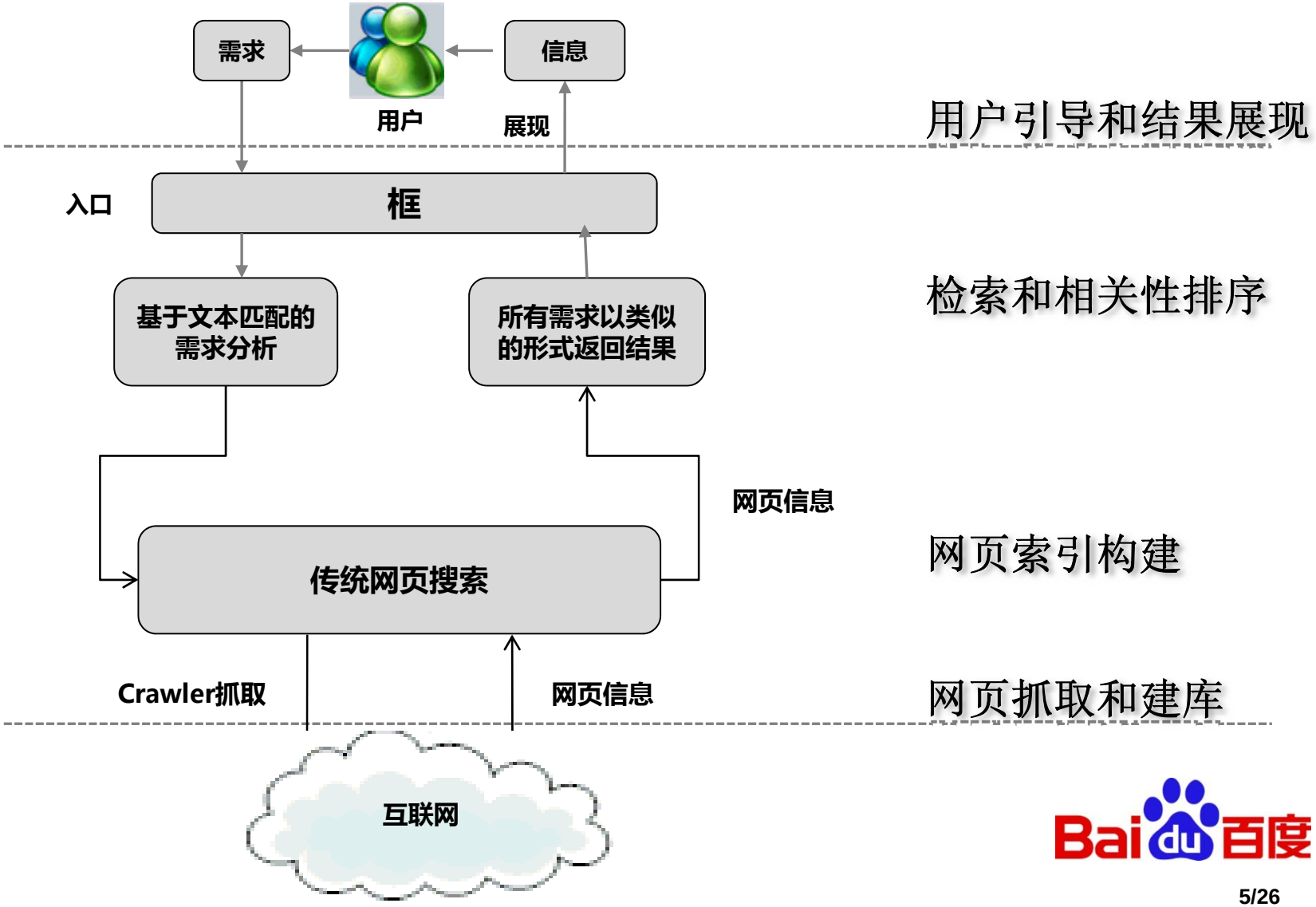
突破性实现网络交易和网络社区的无缝结合；通过与品牌商的合作，为电子商务营销提供创新的模式和全新的环境

有啊 beta 、B2C

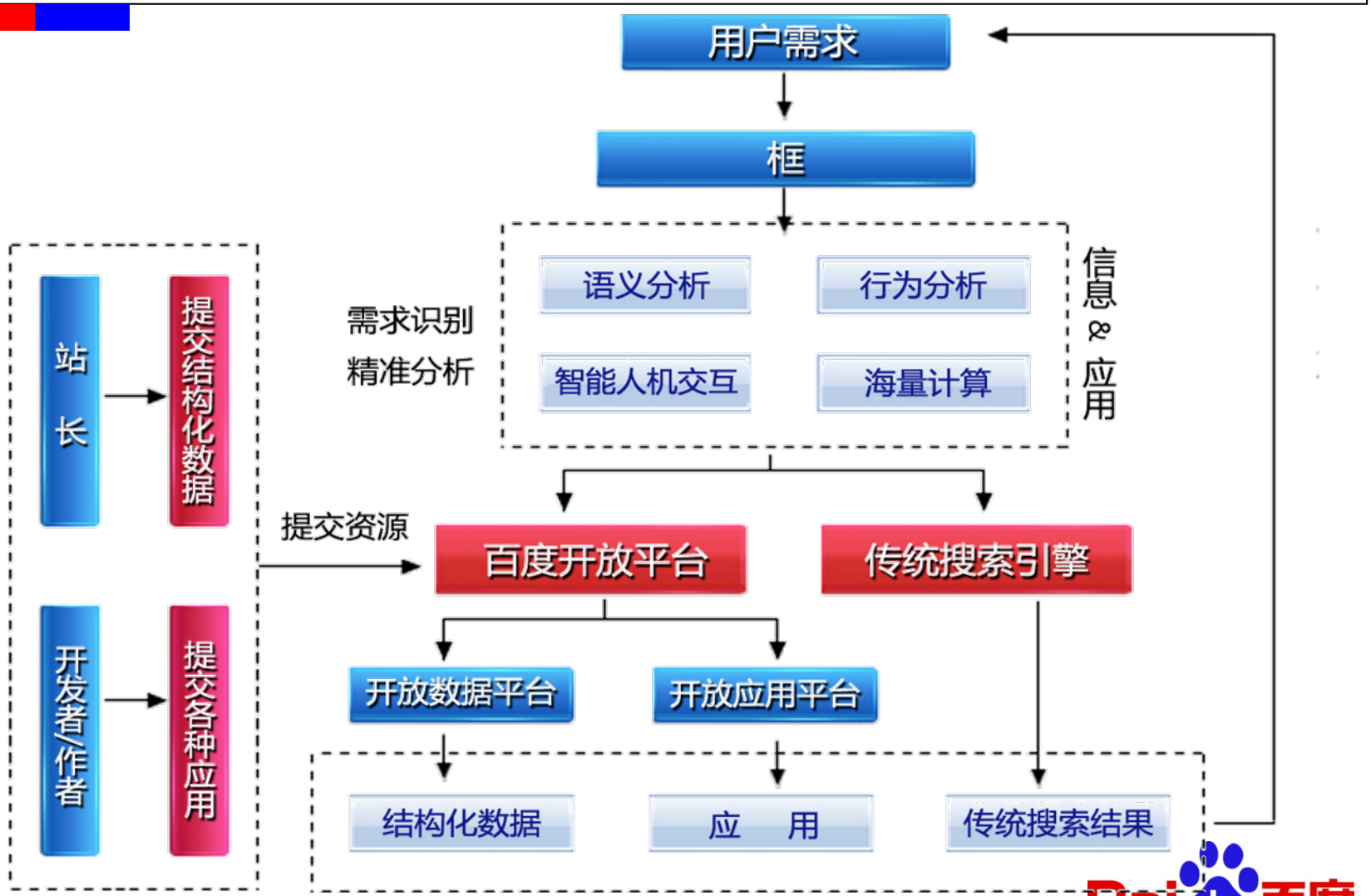
电子商务

Baidu 百度

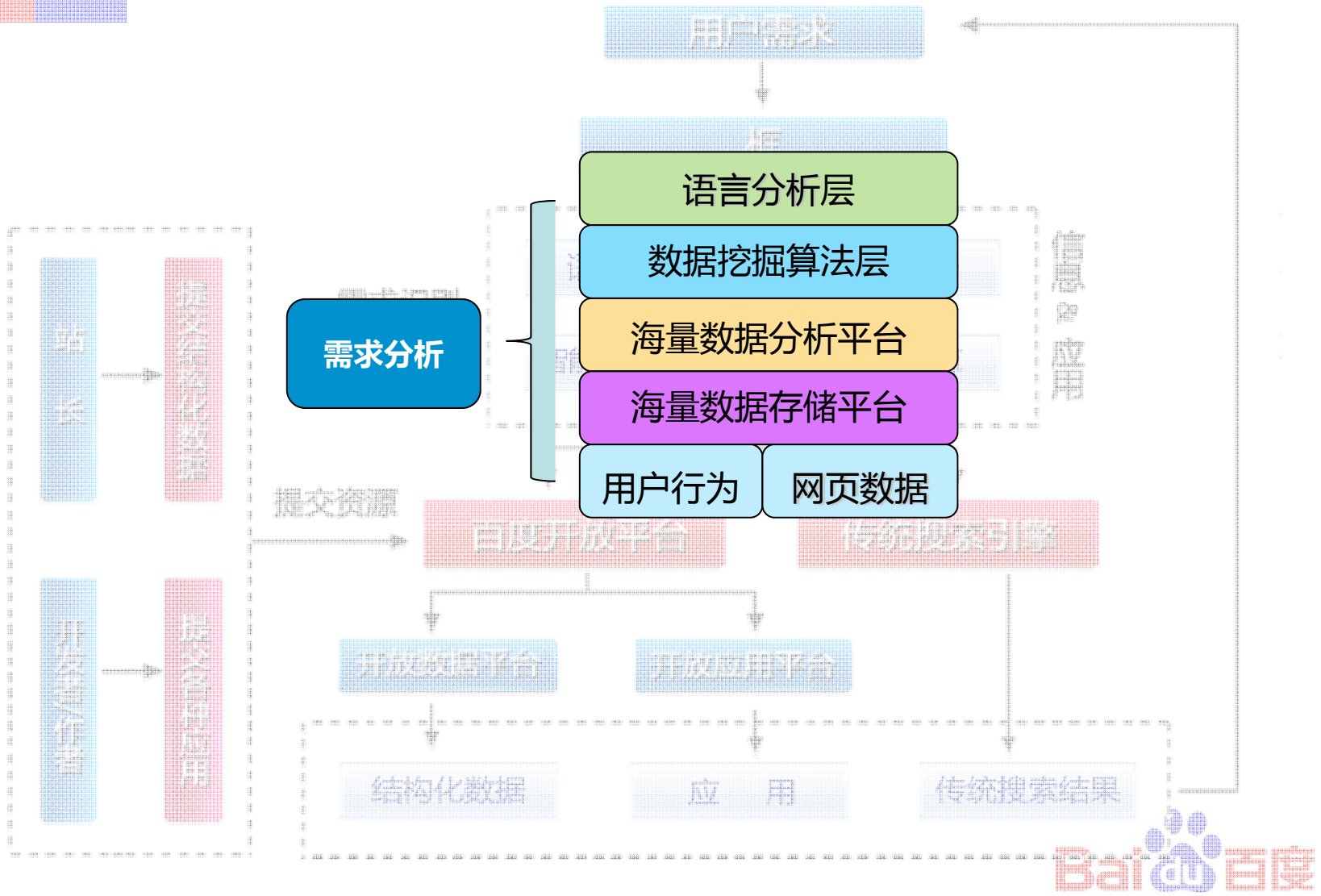
通常的搜索引擎


需求
信息
用户
展现
用户引导和结果展现
入口
框
基于文本匹配的
需求分析
所有需求以类似
的形式返回结果
检索和相关性排序
网页信息
传统网页搜索
网页索引构建
Crawler抓取
网页信息
网页抓取和建库
互联网

基于框计算的新一代搜索引擎



需求分析



用户需求实例

听起来快乐的歌曲	百度一下
令人心情愉快的图片	百度一下
现在几点了	百度一下
电脑中毒了怎么办	百度一下
哪能买到漂亮衣服	百度一下
北京哪能找到女朋友	百度一下

内容提要

- 百度搜索技术概况
- 机器学习在需求分析中的应用
  - 特征表示
  - Term重要性计算
  - 实体属性识别
  - Query分类
- 其他应用
- 结束语

Query特征表示

- 问题定义
  - 用一组特征向量来表示query，需要体现出query语义方面的信息
- 应用点
  - 语义相关性计算
  - 语义相关query触发
  - Query分类

Query特征表示 – Topic Model

- Query term 扩展
  - 基于query的检索结果摘要
  - 基于同session内的query集合
  - 根tf*idf等信息完成term赋权
  - 选取term构成query的扩展特征向量
- 基于term 扩展训练topic model
- 基于topic向量表示query

Query中Term重要性

- 问题描述
  - 将query中的term根据重要程度分成主干、强限定、弱限定和冗余等成分。
- 应用点
  - Query term赋权
  - Query term省略

Query中term重要性计算

- 问题抽象
  - —Query中term重要程度分类
- 特征集合
  - —点特征：term表义能力、iqf、词性、实体属性等
  - —边特征：ngram、互信息等
- 分类方式
  - —两级分类实现

命名实体识别

- 在query中自动识别出人名、地名、机构、品牌、商品等实体
- 应用点
  - 信息抽取
  - 检索粒度分析
  - Query term赋权

命名实体识别方法

- 问题抽象
  - 基于字粒度的query序列标注
- 标记集合
  - 4词位（词首、词中、词尾、单字词）与各类实体的组合。例如人名词首、机构词中、地名词尾等。
- 特征集合
  - 上下文文本特征及特征组合
- 模型及优化

Query分类

- Query的分类方式
  - Query领域分类
  - 主题分类（如：百度知道）
  - 频道query分类（如：视频、图片）
- 应用点
  - Query流量分析
  - 分类别的检索需求满足
  - 分类别的推荐

Query分类方法

- 类别体系
- 数据
  - 数据标注
  - 数据扩充
- 特征集合
  - N-gram、位置、扩展、关键词、关键词标签
- 分类模型
  - 多分类器

内容提要

- 百度搜索技术概况
- 机器学习在需求分析中的应用
  - 特征表示
  - Term重要性计算
  - 实体属性识别
  - Query分类
- 其他应用
- 结束语

机器学习在网页分析中的应用

- 网页分类
  - 问题描述
    - 根据主题类别或者结构类别对网页进行分类
  - 应用点
    - 页面赋权策略
    - 网页筛选和过滤

机器学习在网页分析中的应用

- 特征集合
  - 结构特征
    - 页面url pattern、页面重复性子结构、区域信息位置等
  - 语义特征
    - 标题关键词、正文关键词等
  - 特征处理
    - 特征选择、连续特征离散化等
- 分类模型

机器学习在搜索结果评估中的应用

- Query满足度评估
  - —问题抽象
    - 基于用户行为信息，实现query满足情况分类
  - —特征集合
  - —分类模型

结束语

- 数据
  - 网页数据、人工数据、用户数据 ……
- 目标函数
  - 面向应用
- 表示
  - 假设的表示
  - 数据的表示
- 学习算法
  - 大规模、并行化、高效率