

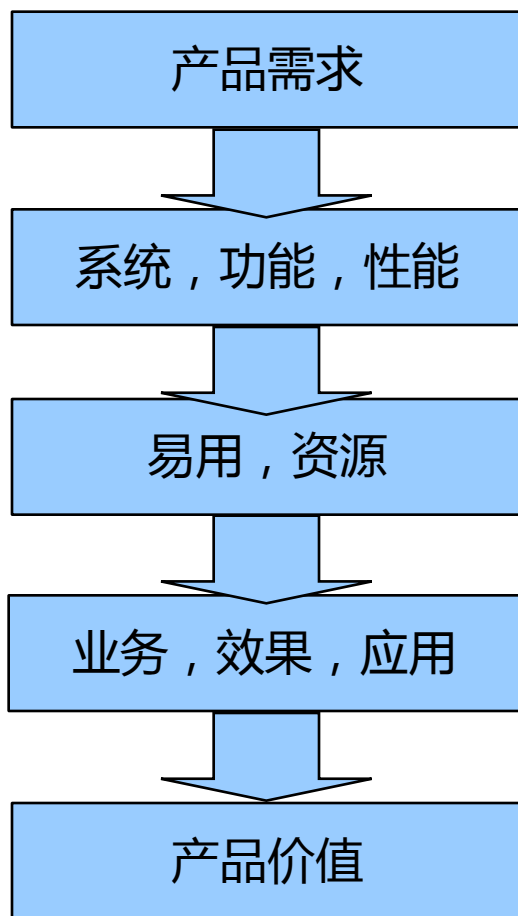
网易通用搜索优化之道 系统实现与数据分析

网易杭州研究院
吴一男
2013/08

大纲

- 介绍
- 通用搜索系统
- 搜索的云服务化
- 搜索数据分析与应用
- 产品应用
- 未来发展

发展过程



介绍：背景

- 背景与需求
 - 众多产品的搜索需求
 - 搜索引擎产品 vs. 面向产品的搜索服务
 - 通用搜索，定制搜索，搜索优化
- 相关产品
 - 开源产品：Solr/SolrCloud, ElasticSearch, IndexTank, Sensei
 - 云搜索：Amazon CloudSearch，阿里云搜索
 - 通用搜索：淘宝TSearcher

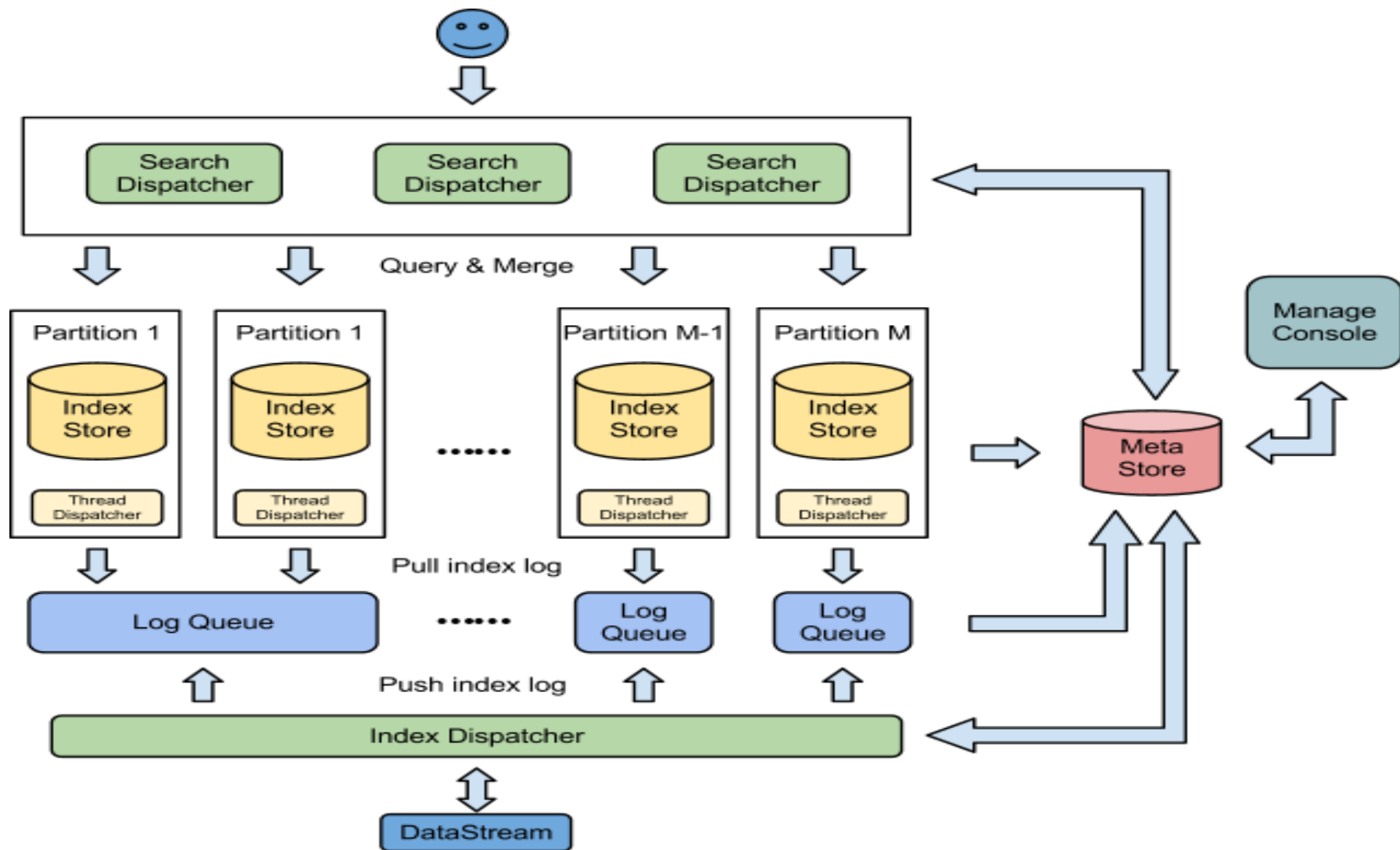
介绍：实现

- 通用搜索系统 (NDIR)
- 搜索云服务化 (NCS)
- 搜索数据分析与应用

通用搜索系统

- 系统架构
- 主要模块
- 整合与管理

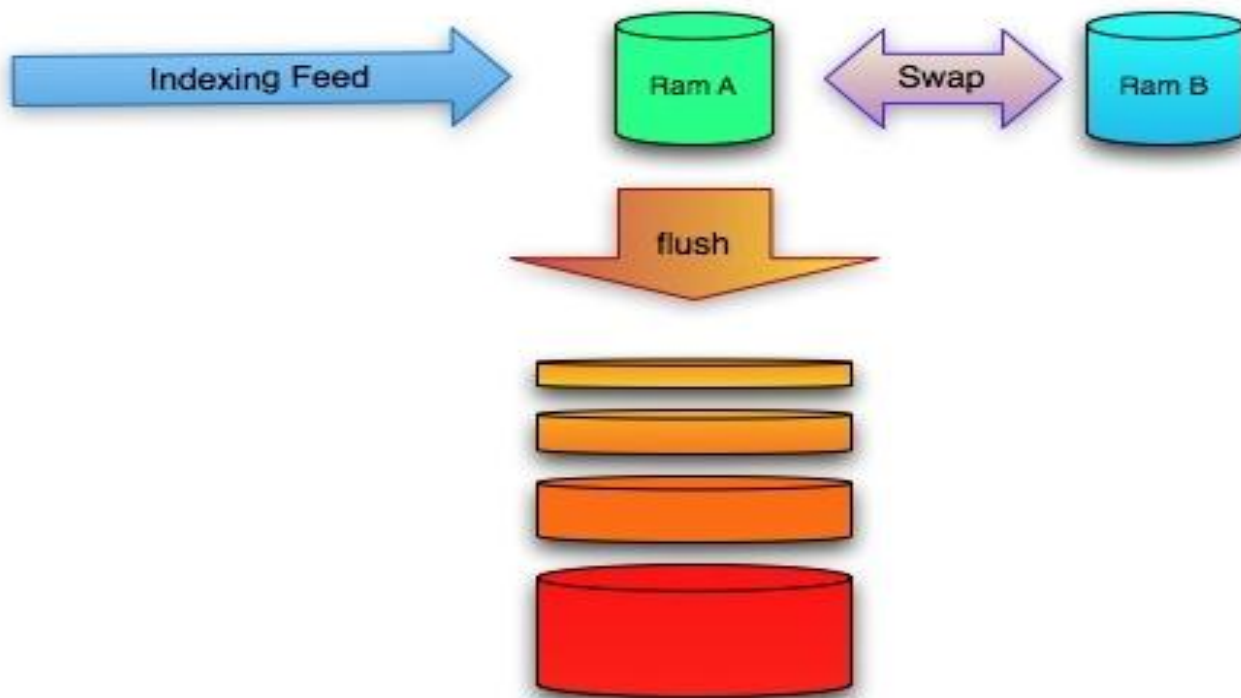
通用搜索系统：架构



通用搜索系统：模块

- 索引/检索引擎：Lucene
- 实时索引：Zoie
- 分布式系统：分区，镜像，主从，扩容
- 数据接入：DataStream同步产品数据库
- 服务接口：HTTP/REST API，Java SDK
- 配置管理：Zookeeper
- 定制化：插件化
- 监控：主机/进程，服务可用，应用状态，日志

通用搜索系统：Zoie实时索引



通用搜索系统：整合

```
{
  code: 200,
  result: {
    totalHit: 18,
    offset: 0,
    docs: [
      {
        score: 6.087317,
        fields: {
          Name: "烟花易冷",
          ArtistAlias: "Jay Chou",
          ArtistName: "周杰伦",
          ID: 185668,
          Genre: "",
          AlbumNameKeyword: "跨时代",
          DJProgram: true,
          AlbumAlias: "",
          Score: 842021
        }
      },
      {
        score: 6.05712,
        fields: {
          Name: "烟花易冷",
          ArtistAlias: "",
          ArtistName: "群星",
          ID: 25723157,
          Genre: "",
          AlbumNameKeyword: "我是歌手(第六期)",
          DJProgram: true,
          AlbumAlias: "",
          Score: 85160
        }
      },
      {
        score: 2.2625945,
```

Beta

网易云音乐

发现音乐 | 我的音乐 | 朋友 | 下载客户端 HOT

Q 单曲/歌手/专辑/歌单/用户

烟花易冷

Q

搜索“烟花易冷”，找到 18 首单曲

单曲	歌手	专辑	歌单	用户
▶ 烟花易冷	周杰伦	《跨时代》		
▶ 烟花易冷	林志炫	《我是歌手(第六期)		
▶ 烟花易冷	李维	《烟花问》		
▶ 烟花易冷	邓涛	《为我的男人唱情!		
▶ 烟花易冷	群星	《铂金试音室①号:		
▶ 烟花易冷	群星	《在乎你每分每秒)		
▶ 烟花易冷	群星	《网络榜中榜·时尚		
▶ 烟花易冷	钟明秋	《为你钟情》		

通用搜索系统：配置管理

全文检索系统NDIR管理平台

app-30.photo.163.org:9999/admin-web/controller.do?controllerName=Login&methodName=login

Google - hkGoogleworktmp其他书签

全文检索系统管理平台

Add IndexApp: qa.itestLogout: neteaseHelp

Manager Cluster
IndexDispatcher
SearchDispatcher
LogQueue
IndexStore
HotQuery
Index Configuration
IndexSchema
PartitionMapping
LogQueueMapping
IndexStoreMapping
Runtime Statues
Replication
TokenizerShow
ServerStatus

我的主页IndexSchema x

新建导入

	索引名称	分区映射状态	创建日期	操作
<input type="checkbox"/>	1	正常! 扩容	2012-08-10 14:35:35	详细 修改 删除 导出
<input type="checkbox"/>	HotQuery_Test	正常! 扩容	2012-09-04 11:37:19	详细 修改 删除 导出
<input type="checkbox"/>	Index_change	正常! 扩容	2012-11-28 15:29:21	详细 修改 删除 导出
<input type="checkbox"/>	MultiTest	正常! 扩容	2012-06-19 22:38:44	详细 修改 删除 导出
<input type="checkbox"/>	MultiTest2	暂无! 新建	2012-06-21 16:50:53	详细 修改 删除 导出
<input type="checkbox"/>	SortOne_TEST	正常! 扩容	2012-08-30 14:00:53	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER	正常! 扩容	2012-03-27 10:50:59	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER10	正常! 扩容	2012-07-03 11:12:38	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER11	正常! 扩容	2012-07-20 11:35:53	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER12	正常! 扩容	2012-08-20 17:43:14	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER13	正常! 扩容	2012-08-23 15:24:48	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER14	正常! 扩容	2012-08-27 16:52:33	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER15	正常! 扩容	2012-09-03 16:54:57	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER16	正常! 扩容	2012-09-07 14:18:14	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER17	正常! 扩容	2012-10-09 10:33:57	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER18	正常! 扩容	2012-10-22 11:29:42	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER2	正常! 扩容	2012-05-16 13:26:53	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER3	正常! 扩容	2012-05-22 13:46:03	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER4	正常! 扩容	2012-06-08 16:23:31	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER5	正常! 扩容	2012-06-11 14:42:59	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER6	正常! 扩容	2012-06-18 13:57:41	详细 修改 删除 导出
<input type="checkbox"/>	TEST_USER7	正常! 扩容	2012-06-21 17:28:12	详细 修改 删除 导出
<input type="checkbox"/>	----	----	----	----

Copyright © 1997-2012 NetEase(Hangzhou)

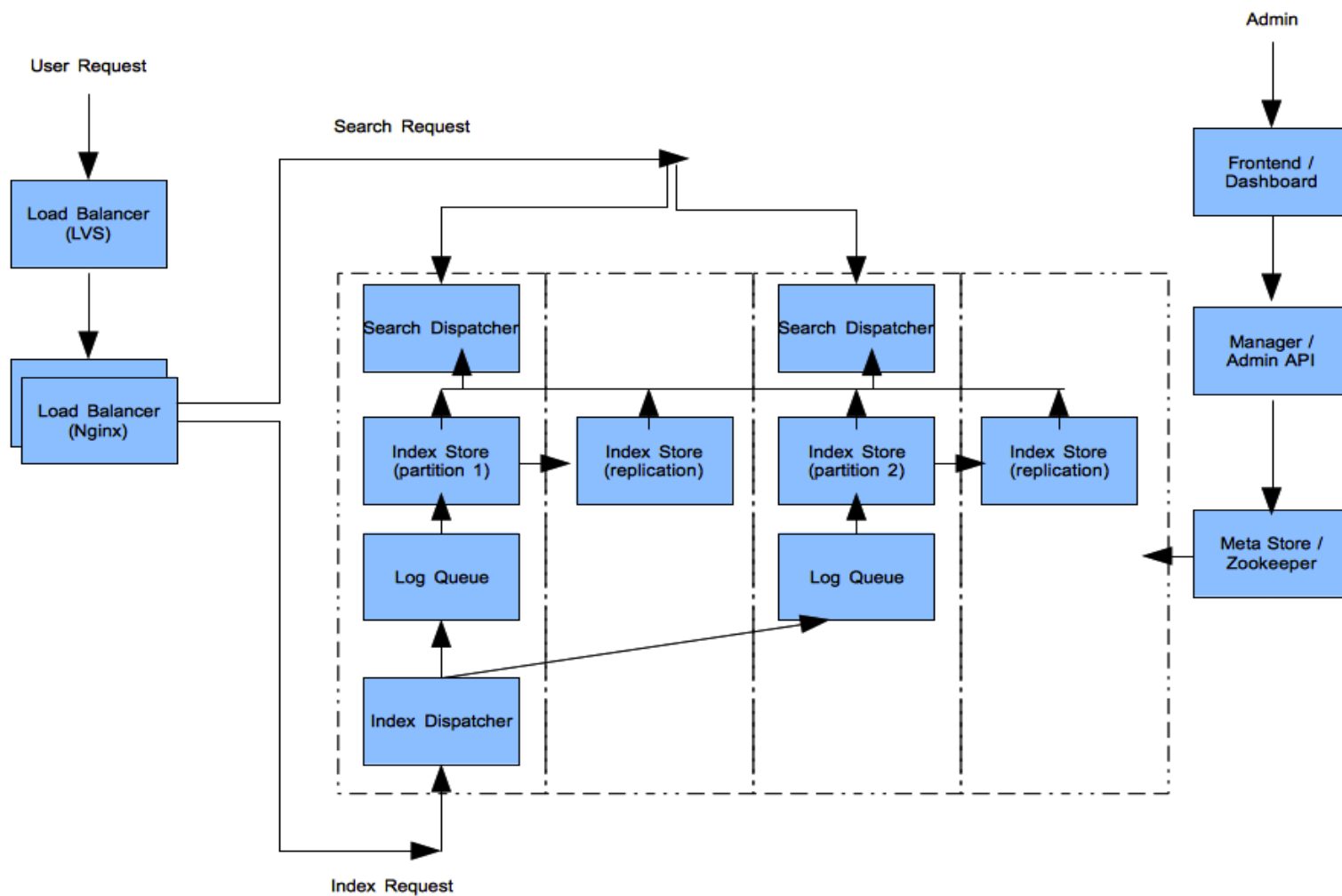
搜索的云服务化

- 云搜索架构
- 主要模块
- 系统后续改进

搜索的云服务化：目标

- 面向人员：产品开发者（轻量级运维）
- 使用与运维：简单，低成本
- 资源利用：弹性，共享，高效
- 服务质量：高可用，数据可靠性，性能合理，系统稳定

搜索的云服务化：架构



搜索的云服务化：模块

- 云搜索实现：
 - 通用搜索系统NDIR
 - 云平台资源：云主机，云硬盘，网络
 - 管理服务器 + 管理前端
 - 数据服务接口：兼容
- 云平台资源：
 - 云主机NVS：弹性计算资源
 - 云硬盘NBS：弹性存储资源
 - 网络：浮动IP资源

搜索的云服务化：管理界面

 网易云 beta
cloud.163.com

晚上好, cloudsearch | 安全退出 | 帮助

云主机云硬盘对象存储关系型数据库分布式数据库云监控云搜索

云搜索首页

集群管理

▼ beta

索引管理

集群管理 > beta > 索引管理

创建索引

导入索引定义

修改索引

删除索引

导出索引

操作后请点击此"刷新"按钮手动刷新! 我知道了»

刷新

<input type="checkbox"/>	索引名称 ▾	索引状态	索引创建时间
<input type="checkbox"/>	pctest	✔ 正常	2013-06-18 18:28:39 CST

« < 第1页,共1页 > »

索引详情

关键字段

占用磁盘空间大小:

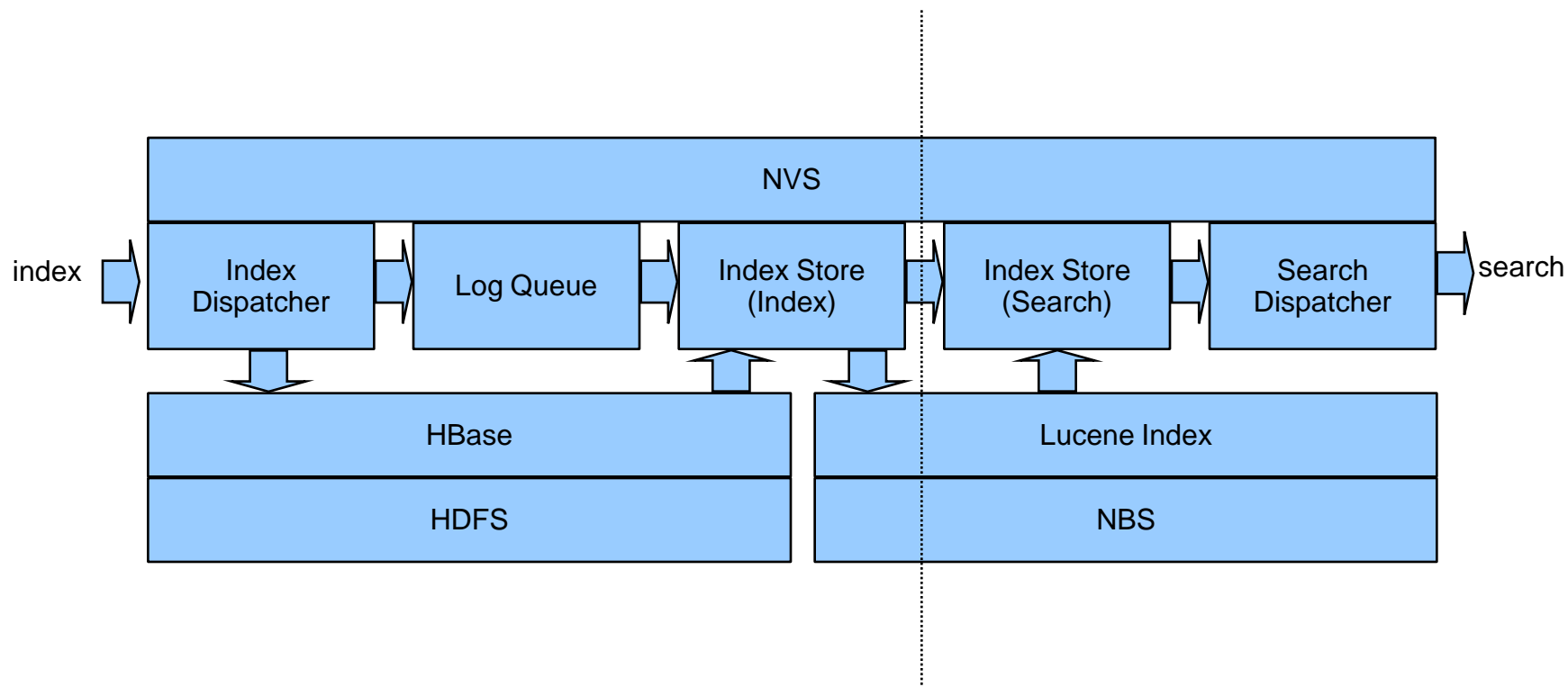
数据条数:

name	type	store	index	sort	analyze	norm	indexTokenizer	queryTokenizer
------	------	-------	-------	------	---------	------	----------------	----------------

系统后续改进

- 资源分离
- 计算资源：云主机NVS（高可用）
- 存储资源：云硬盘NBS + HDFS（数据可靠性）
- 索引资源：Lucene + HBase（索引独立）

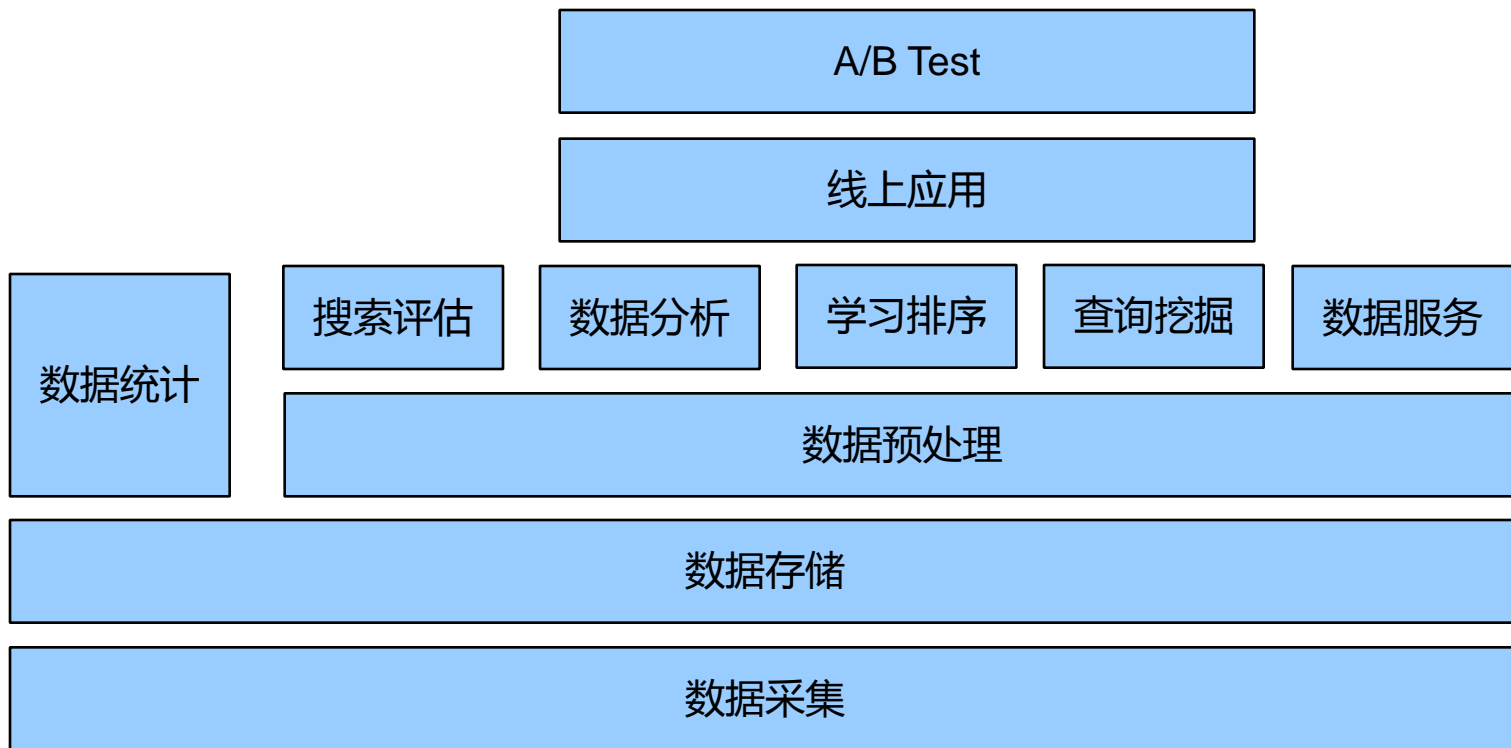
系统后续改进：架构



搜索数据分析与应用

- 数据处理流程
- 主要模块
- 搜索优化流程

搜索数据分析与应用：数据处理流程



搜索数据分析与应用：模块

- 数据采集，数据存储，数据预处理，数据统计
- 搜索评估，数据分析
- 学习排序，查询日志挖掘，A/B test
- 数据服务

数据整合与统计

- 搜索日志数据：搜索，展示，点击
- 采集：日志 → DataStream → HDFS
- 预处理：搜索session分析
- 统计（ MapReduce ）， 查询（ Hive ）

数据整合：搜索统计



数据分析

- 搜索效果评估/比较：Precision/MAP/NDCG
- 搜索数据分布：2/8 vs. 长尾
- 热门搜索/badcase分析
- 具体case分析

搜索数据分析：效果评估/比较

全局统计评分信息比较结果

api-0的全局统计信息: MAP 0.651401594026	api-1的全局统计信息: MAP 0.586686957287
NDCG 0.613529291135	NDCG 0.500906305235
NDCG_ClickRate 0.420175737283	Online Precision 0.213971550004
Online Precision 0.260497760037	Precision 0.567763995195
Precision 0.623971276436	

搜索数据分析：热门搜索

热门搜索列表 - (api-0, Song, 1000)

wuhun 21198 101 点击率: 0.004764600434 分数: [0.985134, 0.883689, 0.5, 0.0045759, 0.5]

儿歌 19681 6677 点击率: 0.339261216402 分数: [0.635351, 0.116341, 1.0, 0.0519283, 1.0]

陈奕迅 16776 2355 点击率: 0.140379113019 分数: [0.697513, 0.0846602, 1.0, 0.0239628, 1.0]

周杰伦 16684 3777 点击率: 0.226384560058 分数: [0.831534, 0.143095, 1.0, 0.0443539, 1.0]

汪峰 15842 5902 点击率: 0.372553970458 分数: [0.872425, 0.745802, 1.0, 0.156862, 1.0]

董小姐 15179 9051 点击率: 0.596284340207 分数: [0.874869, 0.719806, 0.666667, 0.485671, 1.0]

张学友 14216 3733 点击率: 0.262591446258 分数: [0.948226, 0.134952, 1.0, 0.06148, 1.0]

天下3 11666 6805 点击率: 0.583319046803 分数: [0.823695, 0.12433, 1.0, 0.121807, 1.0]

大话西游2 10504 9043 点击率: 0.860910129474 分数: [0.661906, 0.225214, 1.0, 0.222011, 1.0]

王菲 9734 2642 点击率: 0.271419765769 分数: [0.690224, 0.0424615, 1.0, 0.0460242, 1.0]

张惠妹 9668 2683 点击率: 0.277513446421 分数: [0.914249, 0.378663, 1.0, 0.0829541, 1.0]

轻音乐 9175 4140 点击率: 0.451226158038 分数: [0.807, 0.621292, 1.0, 0.122507, 1.0]

大话西游 8648 3960 点击率: 0.457909343201 分数: [0.611633, 0.268272, 1.0, 0.114015, 1.0]

搜索数据分析：Badcase列表

[网易大话西游2](#) 7130 47 点击率: 0.00659186535764 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[wuhu](#) 3254 0 点击率: 0.0 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[wuh](#) 2763 0 点击率: 0.0 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[网易大话西游](#) 1905 8 点击率: 0.00419947506562 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[网易大话](#) 1821 2 点击率: 0.00109829763866 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[艾薇儿](#) 1420 85 点击率: 0.0598591549296 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[宋祖英](#) 1257 323 点击率: 0.256961018298 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[中国好声音第二季](#) 1116 145 点击率: 0.129928315412 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

[梦之声](#) 928 324 点击率: 0.349137931034 分数: [0.0752266, 0.107158, 0.35, 0.006]

[江南s](#) 926 89 点击率: 0.0961123110151 分数: [0.0485626, 0.00704336, 0.405556]

[懂小姐](#) 832 0 点击率: 0.0 分数: [0.0, 0.0, 0.0, 0.0, 0.0]

case分析：搜索/点击数据

关键词：烟花易冷

点击数：997

查询数：1334

NDCG: 0.917962

NDCG_ClickRate: 0.916993

MAP: 1.0

Online_Precision: 0.681409

Precision: 1.0

搜索结果 - 烟花易冷 (api-0, Song)

ID: 185668 歌名: 烟花易冷 歌手: 周杰伦 点击数: 330 展示数: 1088 点击率: 0.303308823529

ID: 25723157 歌名: 烟花易冷 歌手: 群星 点击数: 484 展示数: 1082 点击率: 0.447319778189

ID: 121693 歌名: 烟花易冷 歌手: 李维 点击数: 38 展示数: 1088 点击率: 0.0349264705882

ID: 227089 歌名: 烟花易冷 歌手: 邓涛 点击数: 27 展示数: 1087 点击率: 0.0248390064397

ID: 5234865 歌名: 烟花易冷 歌手: 群星 点击数: 17 展示数: 1088 点击率: 0.015625

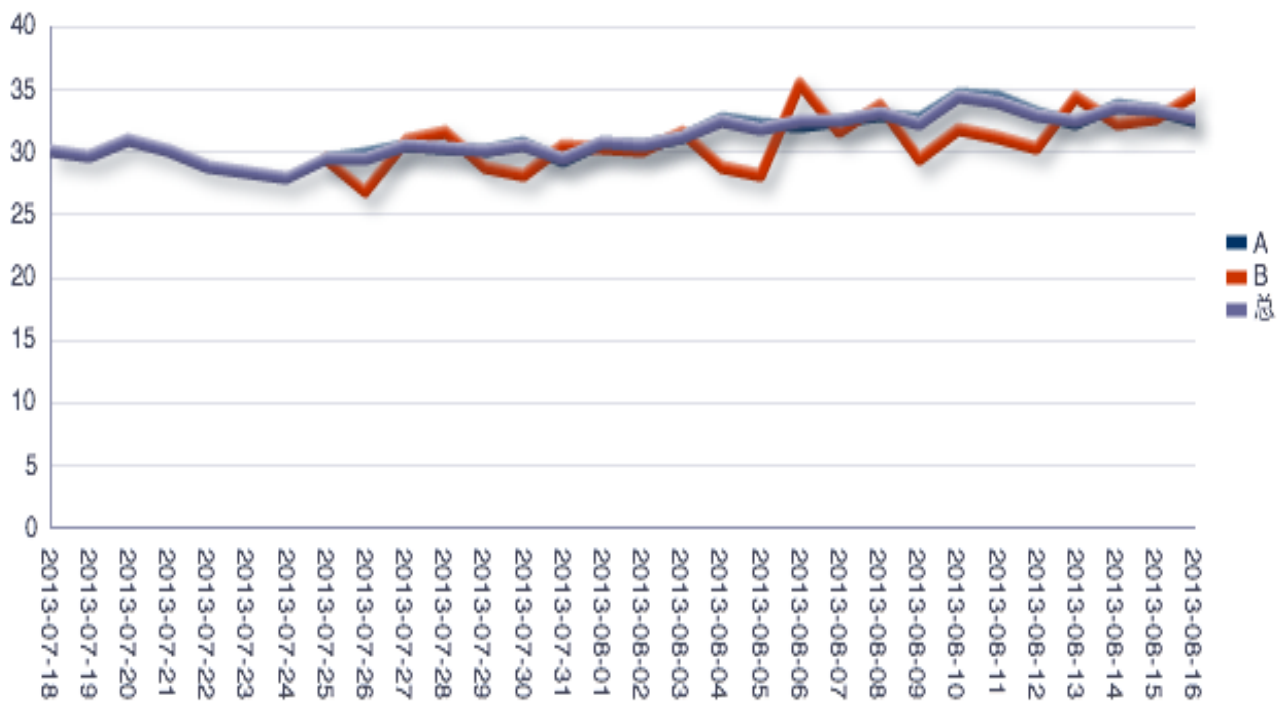
基于数据的优化与挖掘

- 排序优化：学习排序/LTR
- 日志挖掘：查询纠错，意图识别，相关查询

学习排序

- 训练数据：搜索日志 → 查询-结果对
- 数据标注：CTR → score
- 数据特征：文本相关特征，产品热度特征，点击数据
- 算法：
 - Pair-wise LTR，线性 / 非线性
 - RankingSVM，RankBoost，GBDT，etc.
- 评估指标：NDCG，A/B test

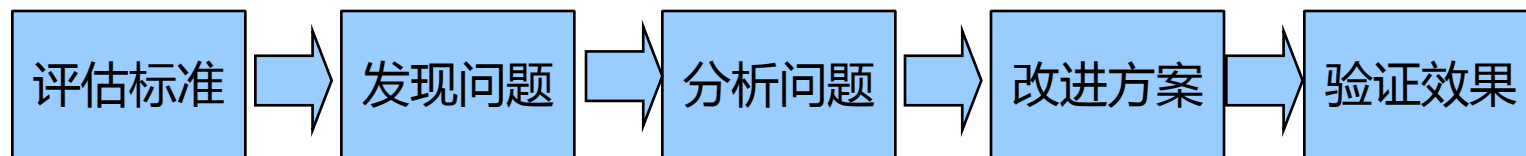
A/B test



数据服务

- 搜索数据：搜索词热度，物品点击热度 → 产品应用
- 个性化数据：用户搜索/点击偏好 → 个性化推荐

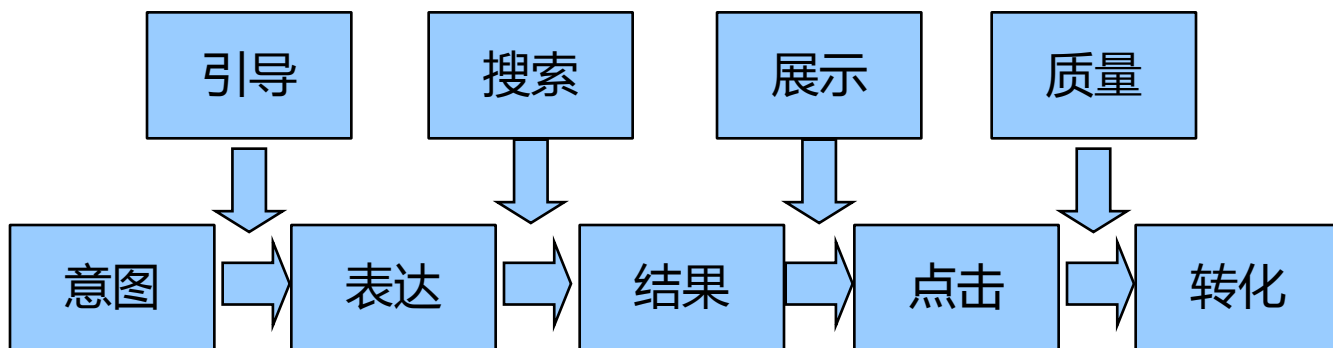
搜索改进流程



搜索改进流程

- 建立评估标准：线上点击率，离线评估指标
- 发现问题：产品反馈，用户反馈，数据分析
- 分析问题：内容，检索，排序
- 改进方案：规则，算法，运营/编辑
- 验证结果：人工判定，离线评估，线上A/B test

用户搜索过程



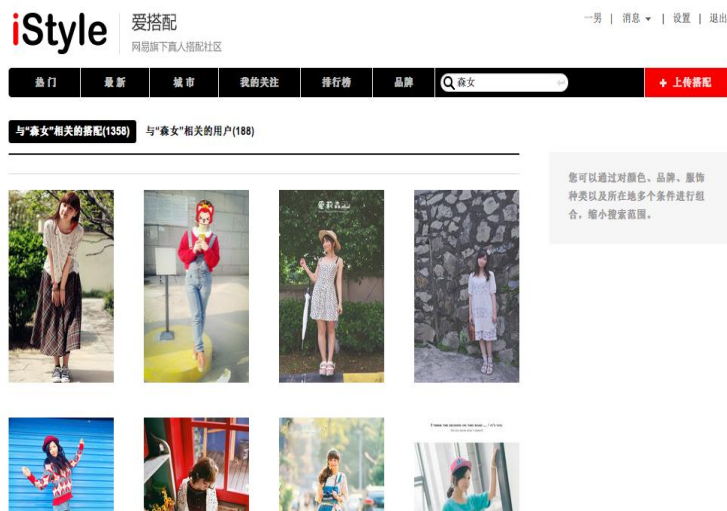
用户搜索过程

- 意图：自发，触发
- 表达：提示，纠错，推荐
- 结果：检索，排序
- 点击：展示优化
- 转化：内容质量

产品应用

- Lofter
- 博客相关产品
- 云阅读
- 云课堂
- 云音乐
- 相册/摄影
- 邮箱
- 网易看游戏
- 内部反垃圾
- 推荐应用

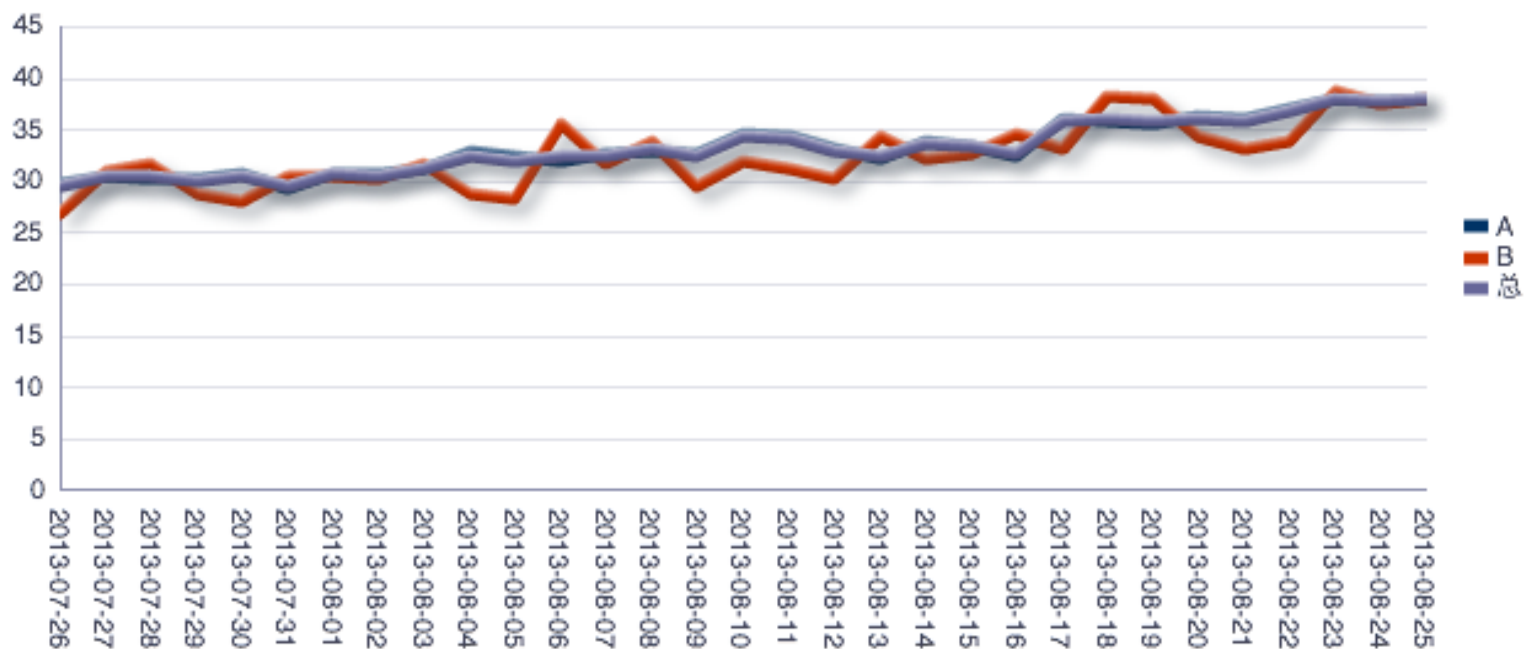
产品应用



产品案例：网易云音乐

- 多端：iPhone/Android，Web，PC
- 搜索场景：资源分类搜索
- 问题与改进：
 - 内容：曲库完善
 - 检索：检索字段，分词/检索调整
 - 排序：相似度/热度计算，排序模型优化
 - 查询处理：查询纠错，意图识别
 - 产品形态：混合搜索提示
- 数据支持：统计，分析，评估，线上验证
- 效果：搜索点击率提升

网易云音乐：搜索效果提升



网易云音乐：搜索效果调整



小时代

搜索“小时代”，找到 14 首单曲

单曲	歌手	专辑	歌单	用户
▶ 小时代曲 (国)	罗力威	《LLV II》	04:04	
▶ 我好想你	苏打绿	《我好想你》	05:24	

- ▶ 热雪
- ▶ 时间煮雨
- ▶ 我好想你
- ▶ 残忍的缠绵
- ▶ 热雪
- ▶ 时髦
- ▶ 小小时代



小时代


搜索“小时代”，找到 14 首单曲

单曲	歌手	专辑	歌单	用户
▶ 我好想你	苏打绿	《我好想你》	05:24	
▶ 热雪	魏晨	《热雪》	03:28	
▶ 时间煮雨	郁可唯	《小时代 电影插曲集》	04:58	
▶ 残忍的缠绵	刘忻	《小时代 电影插曲集》	03:44	
▶ 我好想你	苏打绿	《小时代 电影插曲集》	05:24	
▶ 热雪	魏晨	《小时代 电影插曲集》	03:28	
▶ 小小时代	郭采洁 柯震东 杨...	《小时代 电影插曲集》	03:55	
▶ 雨	付梓郁	《小时代 电影插曲集》	03:51	

产品案例：网易云课堂

- 场景：分类搜索，混合提示
- 问题：
 - 内容：质量好，提高数量与覆盖
 - 检索/排序：结果含badcase，提高精度/降低召回
 - 产品形态：搜索结果分类显示，匹配原因不足
 - 数据统计/评估：不足
- 改进：
 - 整合数据统计评估
 - 产品形态：混合搜索结果，增加高亮摘要
 - 调整检索与排序


网易云课堂：搜索形态调整

 网易云课堂 BETA

全部分类

搜索

我的云课堂 1



★★★★★ (1份评价) 98人在学


简介：本专题与连词、简单句、并列句、动词时态紧密联系，因此，我们在复习时一定要结合已学内容，反复训练，达到

目录：课时1 祈使句，and/or+陈述句的基本用法是什么？ 课时2 引导时间状语从句的

课程分类：中小学 | 来源：清大学习吧 | 讲师：清大学习吧

查看更多相关的课程

相关的计划




斯坦福大学公开课：机器学习课程

★★★★★ (7份评价) 1683人在学

简介：目前还没有任何曙光。但是，机器学习无疑是最有希望实现这个目标的方向之一。斯坦福大学的

目录：任务1 机器学习的动机与应用 任务2 监督学习应用.梯度下降 任务3 欠拟合与过

计划分类：IT与互联网 | 来源：网易公开课



机器学习入门

★★★★★ (2份评价) 255人在学

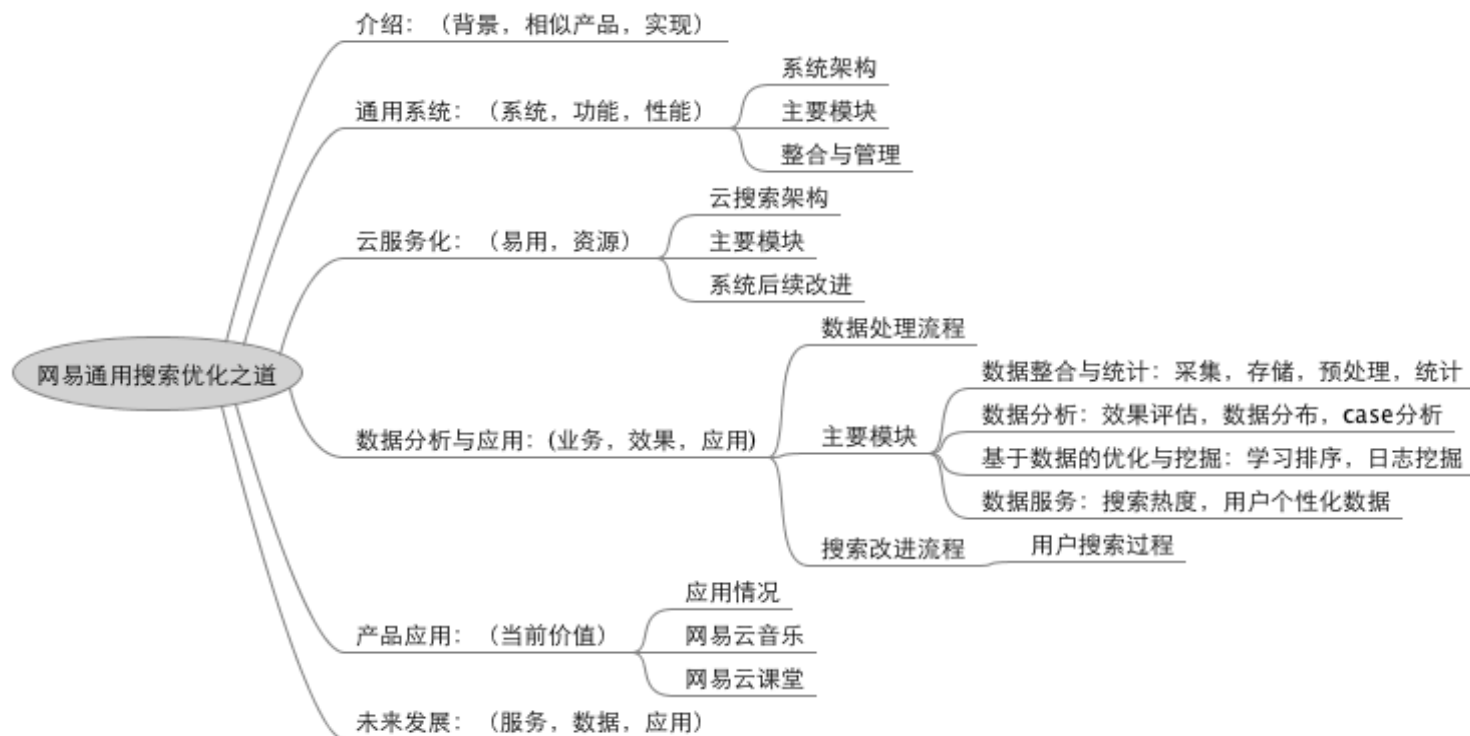
简介：给初学机器学习的人，一个初步的意见和建议。当你完成本计划的同时，你应该可以应用机器学习领域的基本

目录：任务1 机器学习入门和课程的学习 任务2 机器学习课程的学习 任务3 机器学习课程的学习

未来发展

- 完善服务化/云平台建设
- 数据可视化，基于数据驱动的改进
- 搜索个性化，与用户数据的结合
- 加强移动端搜索应用
- 扩展应用领域

总结



团队

- 通用搜索系统/ 云服务
- 数据与算法
- 个性化推荐
- 数据平台

联系：hzwuyinan@corp.netease.com

谢谢！

Q&A