搜狗信息流推荐算法综述

王东

提纲
Content

1 推荐系统架构

2 文章NLP

3 召回算法

4 个性化排序

推荐系统架构

提纲
Content

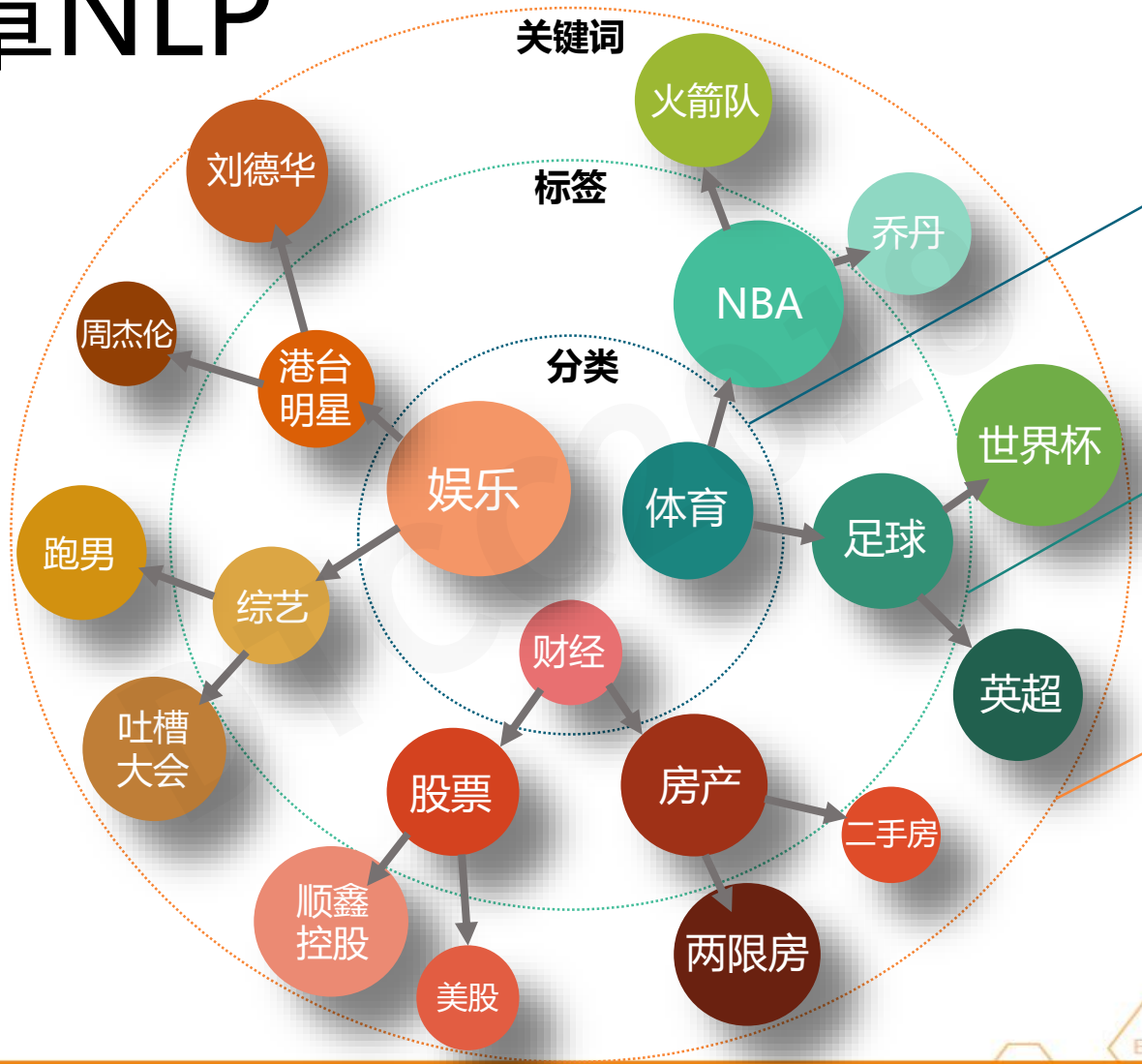
1 推荐系统架构

2 文章NLP

3 召回算法

4 个性化排序

文章NLP



聚焦主要领域
为用户提供信息导航

描述比较精确，但又属于抽象概念的语义

领域内热点人物、机构、作品、产品等实体内容

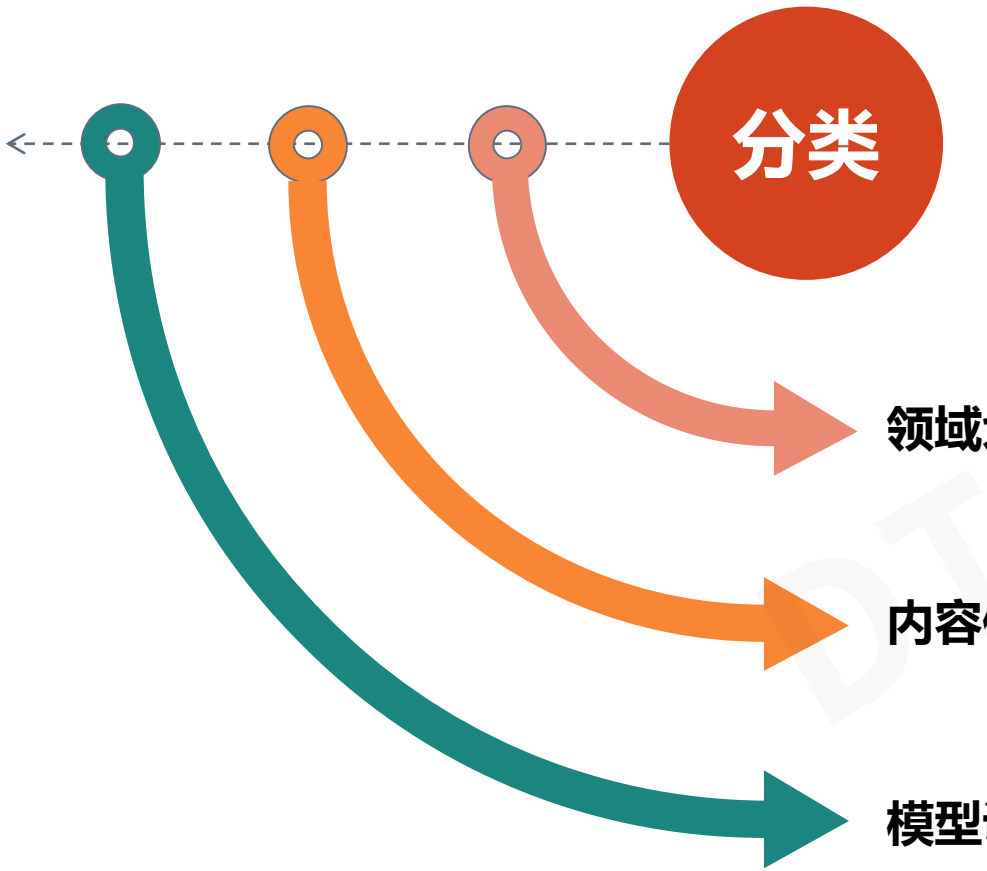
文章NLP

关键词

标签

分类

分类

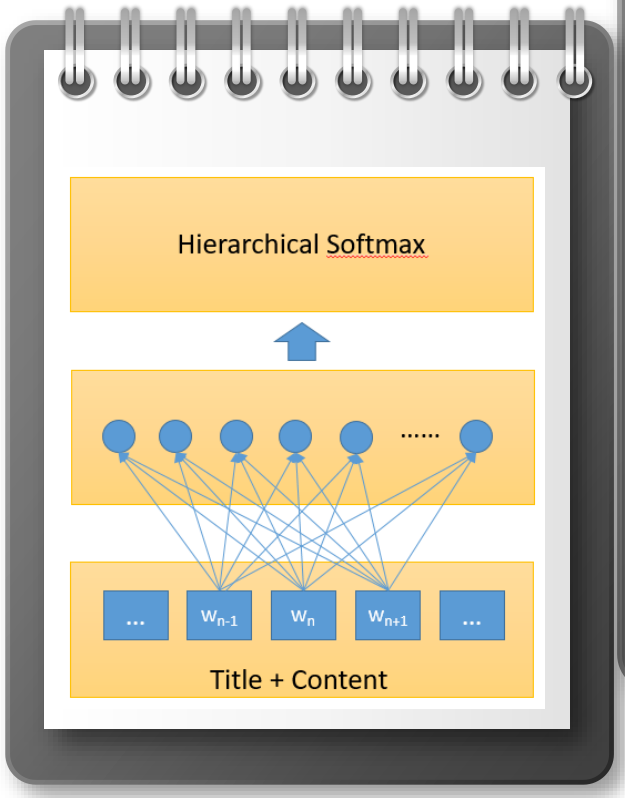


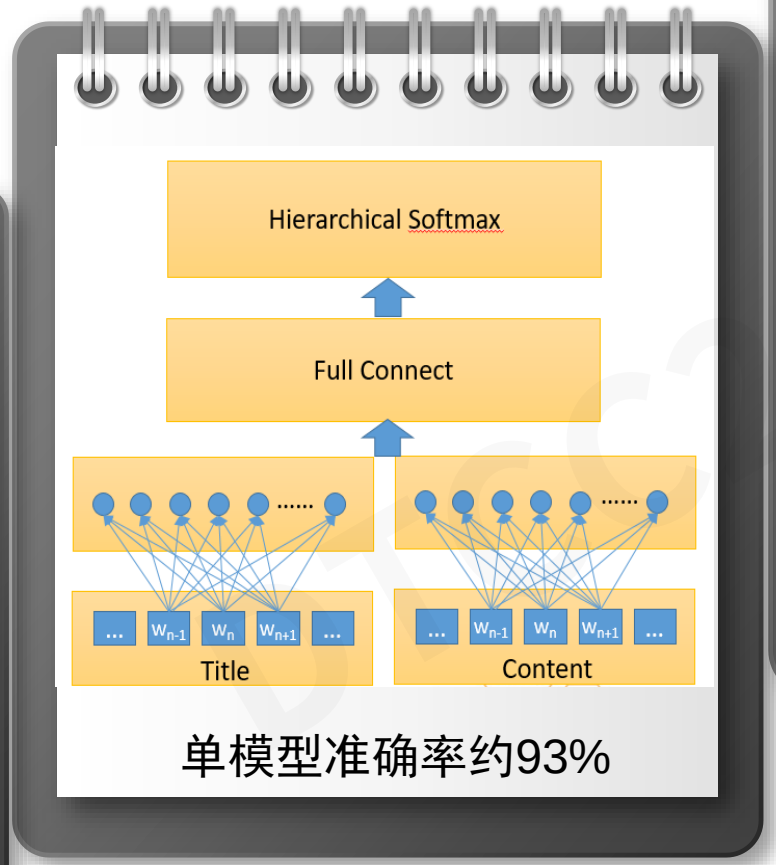
领域划分
- 内容覆盖尽量全面
- 有一定用户受众，同时有一定文章量

内容体系
- 娱乐, 情感, 军事, 体育, 健康, 美食, 汽车, 星座, 游戏, 时尚, 财经, ……

模型训练
- FastText文本分类模型

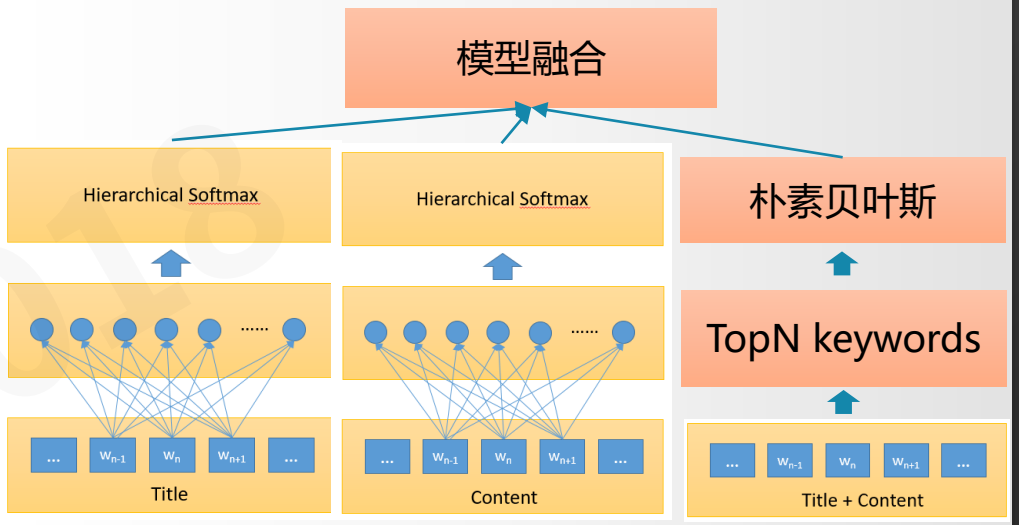
分类





单模型准确率约93%

多个模型融合准确率达到96%



标签

- 描述比较精确，同时又属于抽象概念的语义

领域划分

分类	标签
娱乐	电视剧，明星八卦，真人秀，港台娱乐，综艺，韩娱...
军事	武器，海军，陆军，空军，中东局势，环球军事，中国军情...
健康	疾病，保健品，男性健康，女性健康，养生，营养学，食疗，中医，中药，饮食健康...
科技	手机、软件、人工智能、移动互联网、通信、移动支付、穿戴设备、网络安全、大数据...
教育	家庭教育，留学，小学，资格考试，研究生，中考，大学，幼儿园，高考...
.....	.....

内容体系

- TextCNN文本分类模型

模型训练

标签

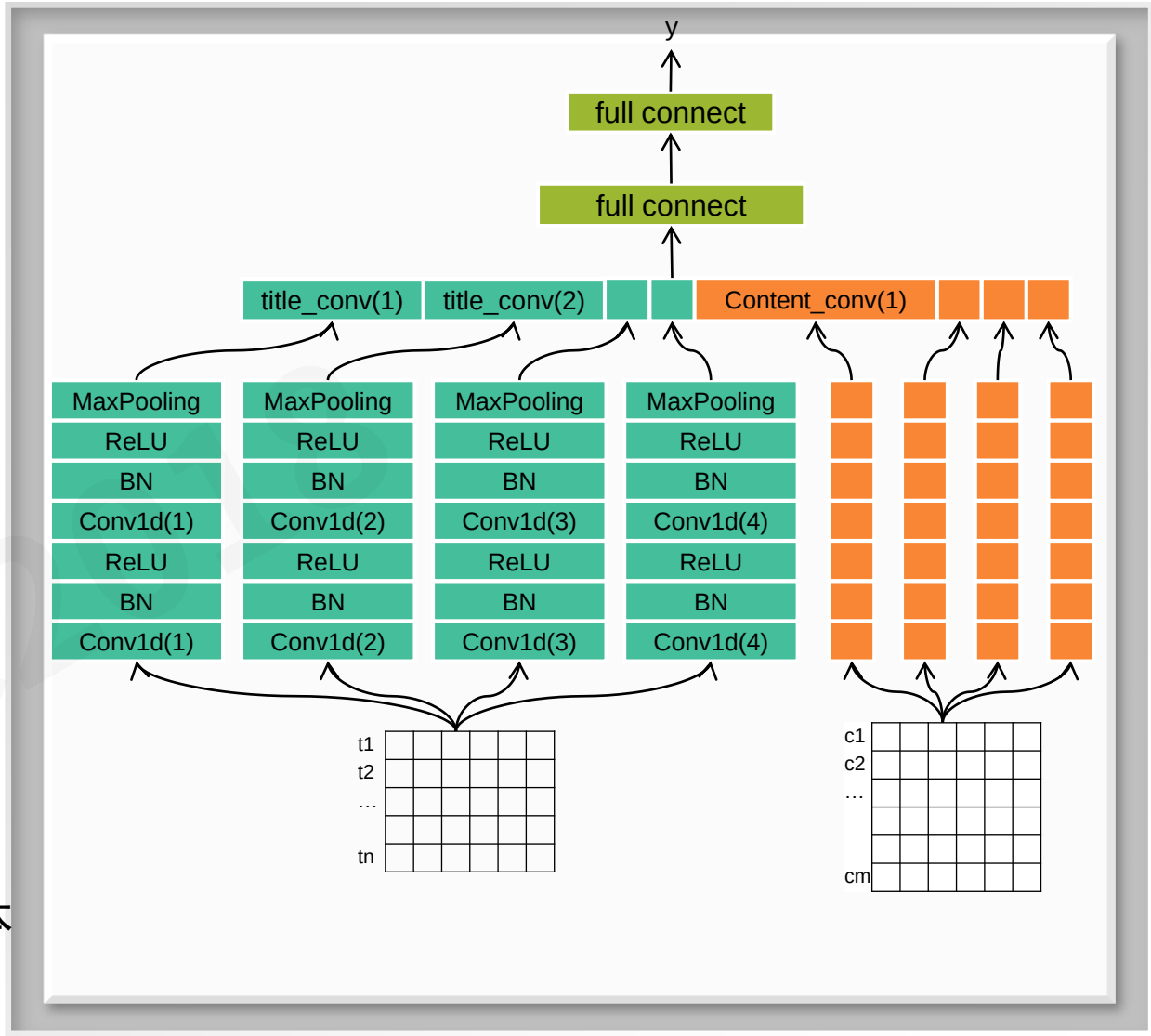
- 对标题和正文分别进行卷积计算
- 使用两层卷积
- 使用BatchNorm
- 使用两层全连接完成分类计算

01 TextCNN

标 签

02 数据

- 数据增强：单篇文章拆分多个样本
- 多段文本预测结果拟合



关键词

1 领域划分

- 各分类中的热点人物、机构、作品、产品等实体内容

2 内容体系

- 周杰伦, 搜狗, 吐槽大会, OPPO, 奥巴马, 印度, 皇家马德里, 灌篮高手,……

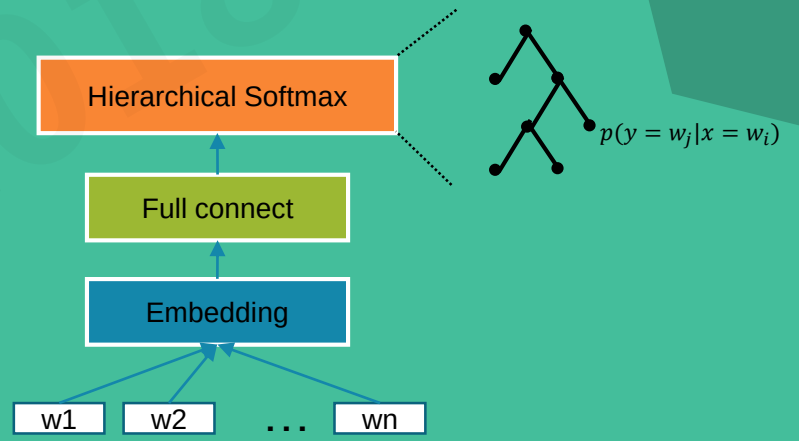
3 模型

- 相似度模型：Tf-idf、lda、word2vec
- 概率模型： Skip-Gram + 层次Softmax

关键词

概率模型

- 问题定义
  - 对于文本S，条件概率$p(S|w_i)$ $w_i \in S$, 表示通过w_i能够猜测出文本S大意的可能性
  - $p(S|w_i)$ 值越高，则w更加适合最为这段文本的关键词
  - 使用朴素贝叶斯假设，$p(S|w_i) = p(w_1, w_2, \dots, w_n|w_i) = \prod_{k=1}^{n} p(w_k|w_i)$
- 对$p(w_k|w_i)$建模
  - 选取模型：Skip-Gram + 层次Softmax
  - 词向量训练方法，预测概率 $P(Context|w_t)$
- 算法优势
  - 基于$p(s|w_i)$较高者为关键词的定义，逻辑上清晰严谨
  - 训练速度快
  - 提取准确率89%



提纲
Content

1 推荐系统架构

2 文章NLP

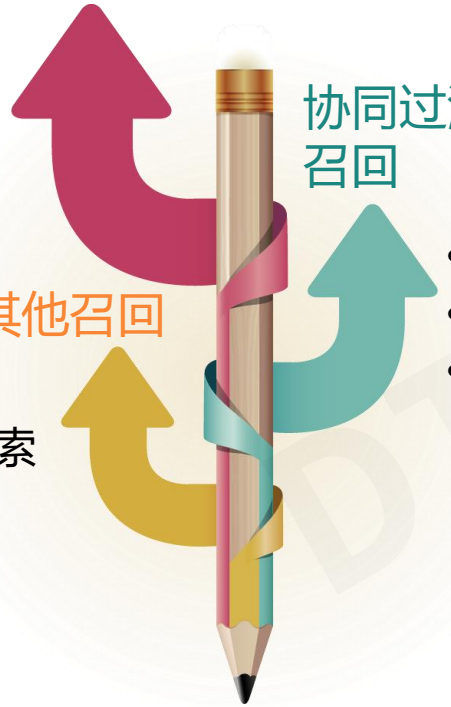
3 召回算法

4 个性化排序

召回算法

基于内容（CB）召回

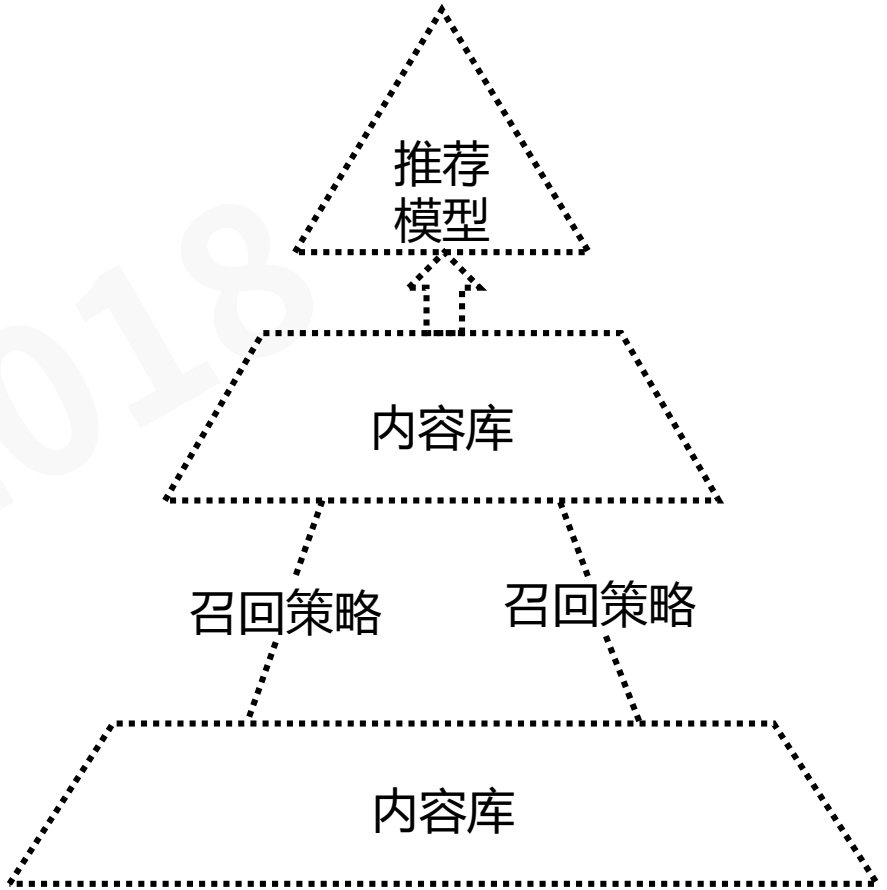
- 显式分类，标签，关键词，隐式分类
- 地域、人口属性、搜索历史、订阅



协同过滤（CF）召回

- Item-based
- LFM
- NCF

其他召回



基于内容（CB）召回


用户兴趣
根据兴趣拉取相应文章并rank，获取top结果
离线更新倒排索引
军事
0.5
手机
0.4
电影
0.1
贾冰
0.1
军事
手机
电影
贾冰

基于内容（CB）召回

基于规则排序

$$final\_core = account\_score * hot\_score * time\_score$$

基于模型排序

将问题简化为预测ctr，结合相关性

问题抽象

排序的主要目标

基于内容（CB）召回

- 文章样式：视频？图文？单图？多图？
- Title：长度？包含关键词？特殊标点符号？
- 内容：topic，tag，keyword
- 账号：等级，来源，地域
- 入库时间

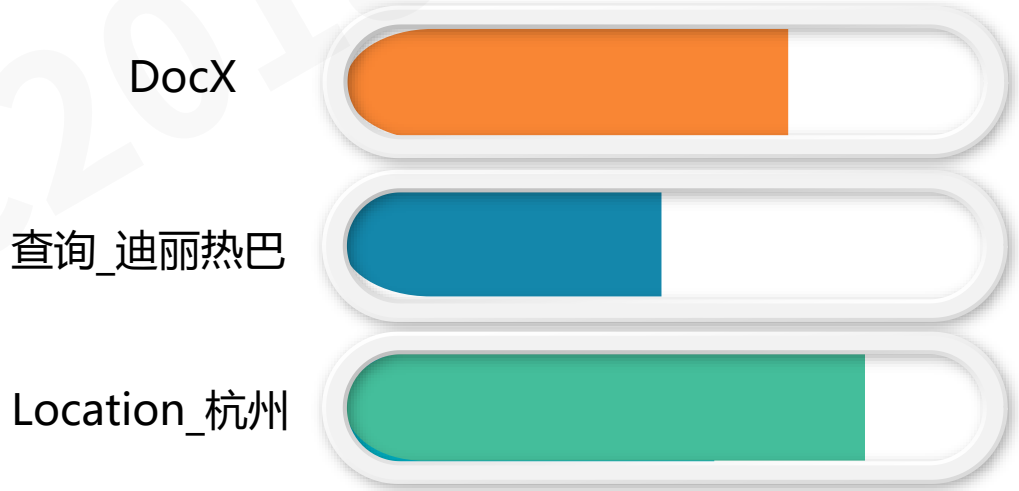
- 召回词的位置
- 召回词的向量化与其他关键词的夹角

- 文章热度：展现，点击，分享，收藏，不喜欢
- 文章-召回词热度
- 账号热度

协同过滤（CF）召回

Item-based

	U1	U2	U3	……	UN
Doc1					
……					
DocN					
Query					
Location					



协同过滤（CF）召回

隐语义模型（LFM）

- model-based协同过滤
- 通过降维的方法将评分矩阵补全
- 核心思想是通过隐含特征联系用户和物品
- 对物品和用户进行兴趣分类，对某个用户，先得到他的兴趣分类，确定他对各类物品的喜欢程度，再在这个类里挑选他可能喜欢的物品
- 采取基于用户行为统计的自动聚类

	item 1	item 2	item 3	item 4
user 1	R11	R12	R13	R14
user 2	R21	R22	R23	R24
user 3	R31	R32	R33	R34

R

=

	class 1	class 2	class 3
user 1	P11	P12	P13
user 2	P21	P22	P23
user 3	P31	P32	P33

P

×

	item 1	item 2	item 3	item 4
class 1	Q11	Q12	Q13	Q14
class 2	Q21	Q22	Q23	Q24
class 3	Q31	Q32	Q33	Q34

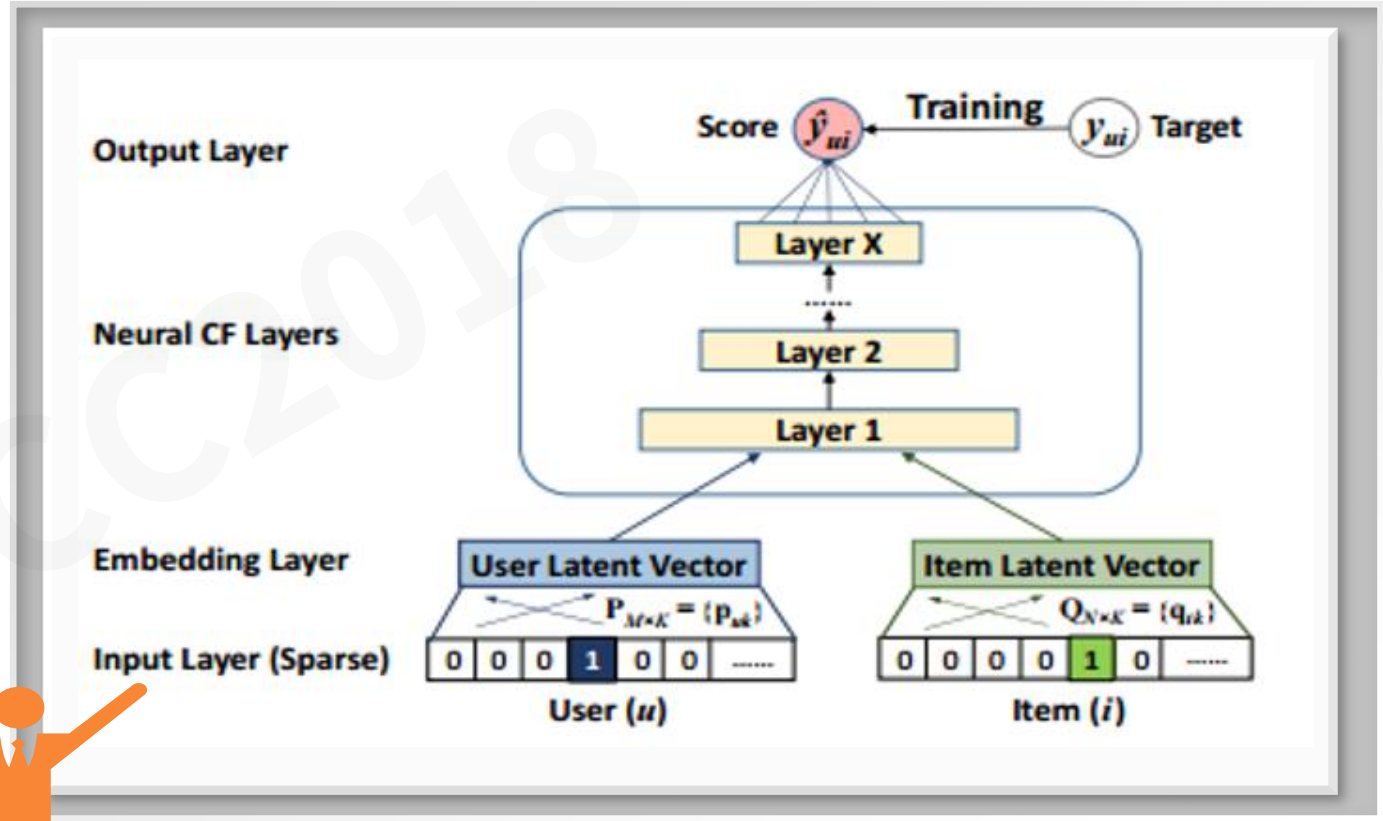
Q

$$r_{ui} = p_u^T q_i = \sum_{f=1}^{F} p_{u,f} q_{i,f}$$

协同过滤（CF）召回

神经网络协同过滤（NCF）

- 基于神经网络的技术，来解决在含有隐性反馈的基础上进行推荐的关键问题。



提纲
Content

1 推荐系统架构

2 文章NLP

3 召回算法

4 个性化排序

个性化排序

LR

- 最常用的点击率预估模型，速度快，效果好
- 与人工规则相比，效果提升显著

GBDT+LR

- 速度受限，对比LR优化效果不明显

FTRL

- 在LR的基础上，效果提升明显

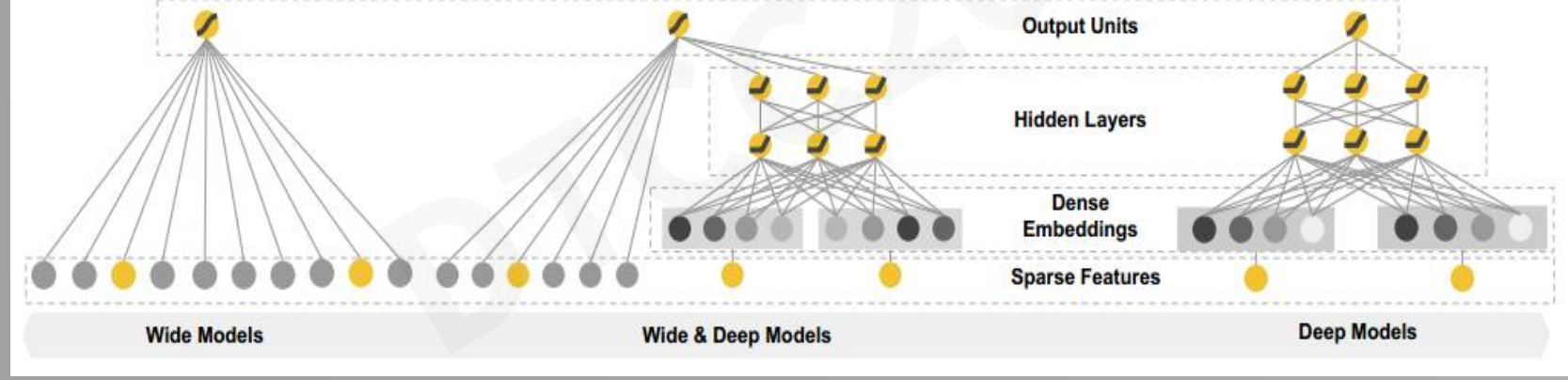
Wide & deep

- 在FTRL的基础上，效果再次提升

个性化排序

Wide&deep learning

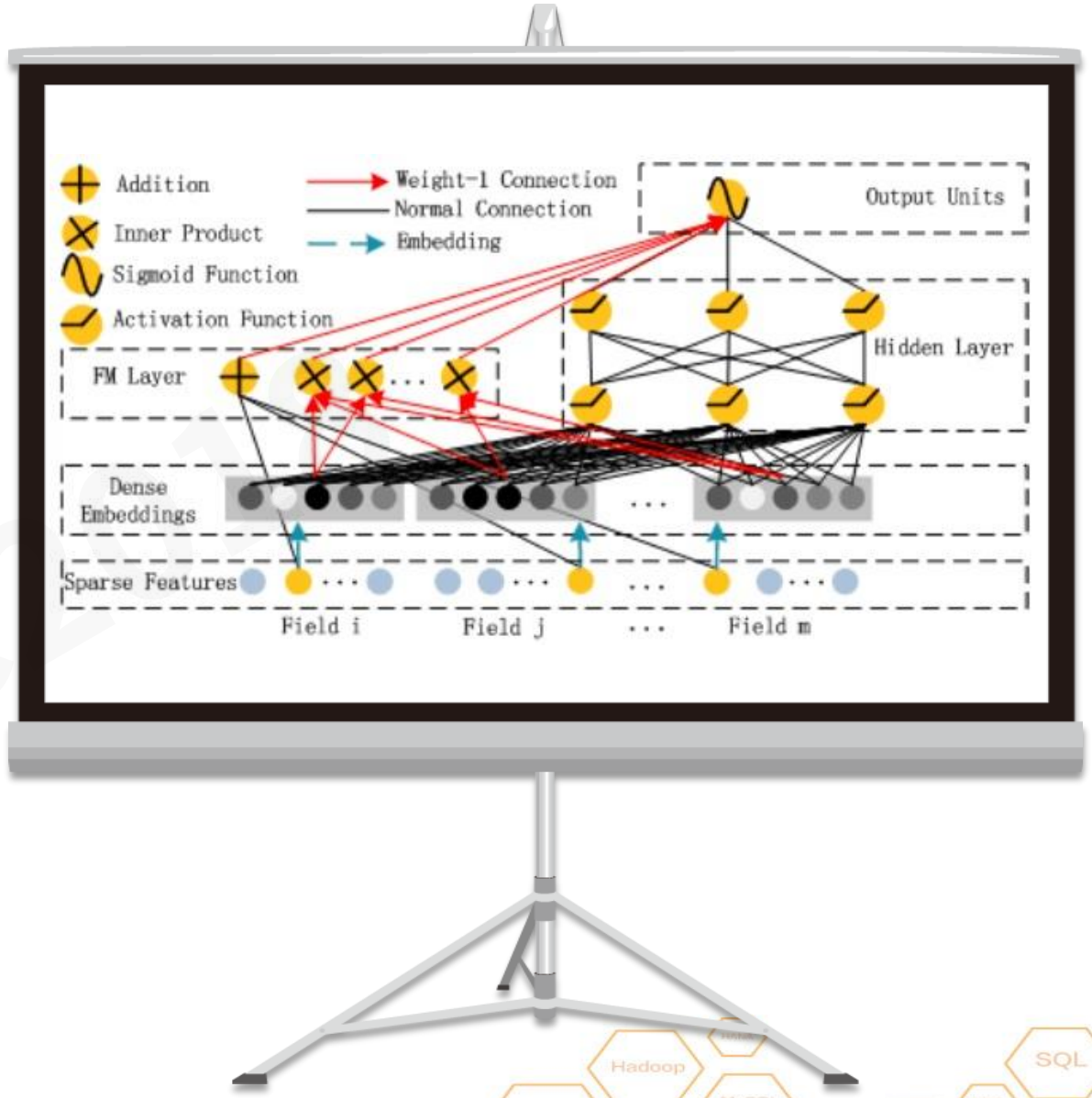
- 通过Deep Models学习高阶特征，增强模型的泛化能力
- 通过Wide Models 建模，增强模型的”记忆能力”



个性化排序

- FM层和Deep层共享Embeddings层的结果
- Embeddings参数训练时被FM和DNN同时更新
- 相比于wide && deep，训练参数没有增加

deepFM



THE 总结 END

- 用户画像、多样性、冷启动、时效性

- 传统机器学习模型->深度模型

- 方法很多，策略和算法同样重要

- 推荐系统的基石

其他重要方面 4

个性化排序 3

召回算法 2

文章NLP 1

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigData
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168