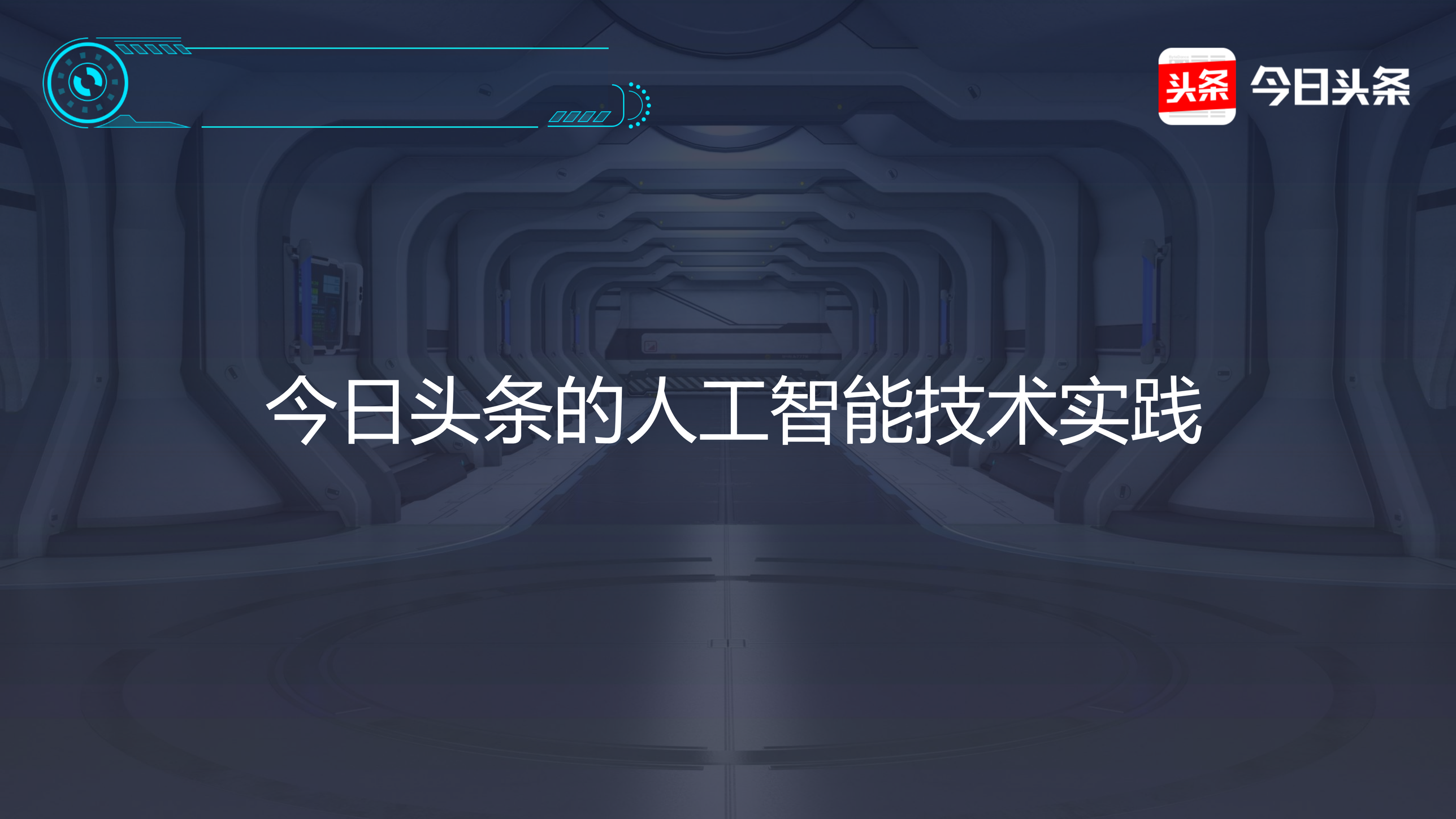

今日头条的人工智能技术实践

媒体形式的历史变迁

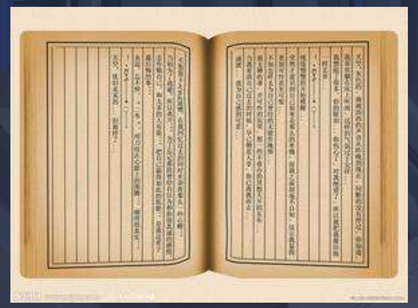
今日头条

传唱史诗

公元前2000年

《吉尔伽美什》

书籍



公元前1000年

《尚书》

报纸

公元200年

《威尼斯公报》

广播&电视

20世纪20年代

匹兹堡KDKA电台

贝尔德的电视机

互联网

20世纪70年代

互联网雏形ARPANET诞生

从人工到智能算法的媒体革命

今日头条

人工

分发

管理

创作

智能算法

2016年是一个历史拐点

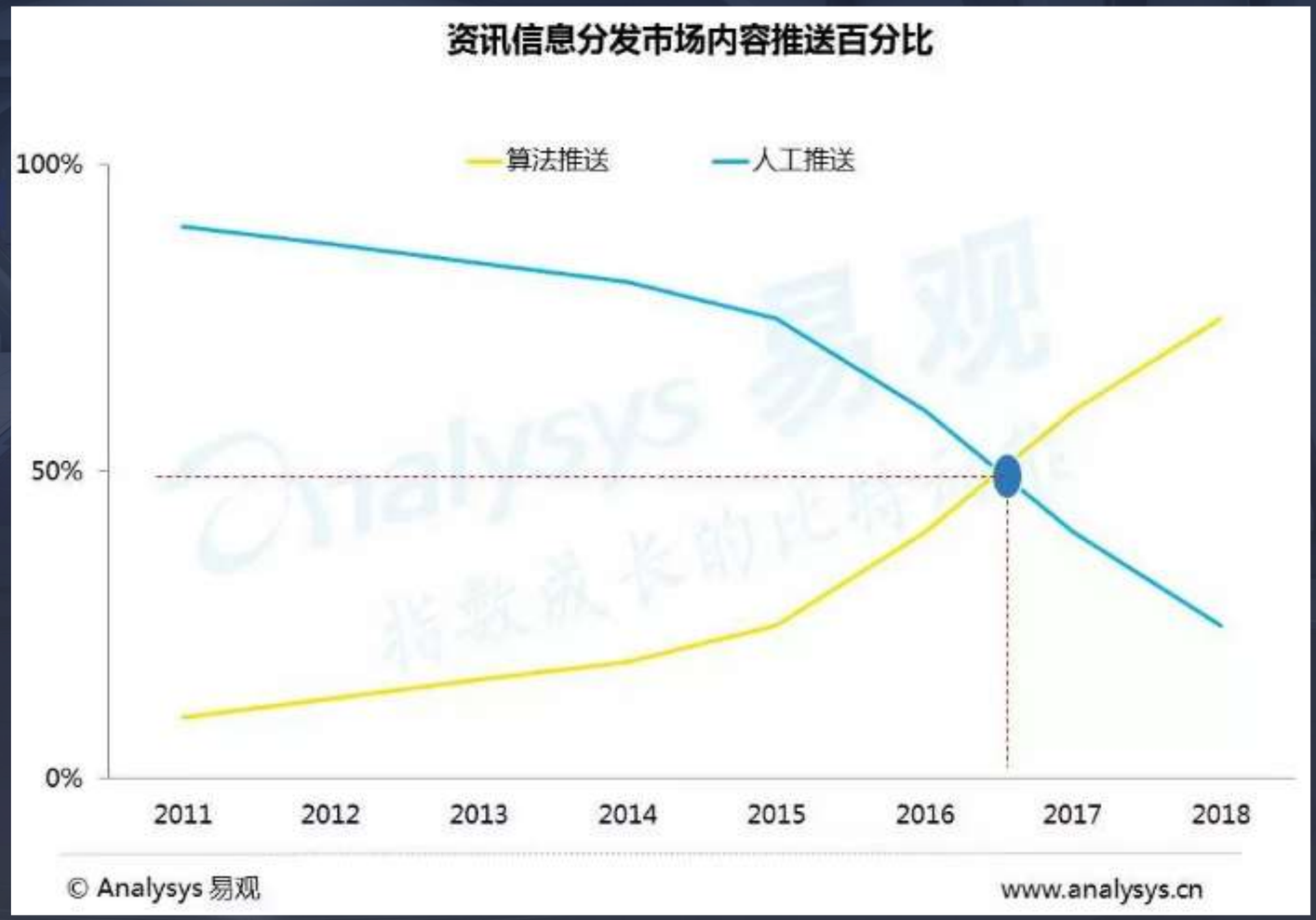
今日头条

“近日，第三方监测机构易观发布了一个具有“里程碑式意义”的数据：2016年，在资讯信息分发市场上，算法推送的内容将超过50%。

这将成为一个分水岭。它意味着，我们以后接触到的信息，将主要由“智能机器人”为我们准备，而以往看似不可或缺的“人工编辑”角色，则不可避免地被边缘化。

自2012年今日头条开启算法分发的尝试，4年之后，算法时代正式宣告来临。”

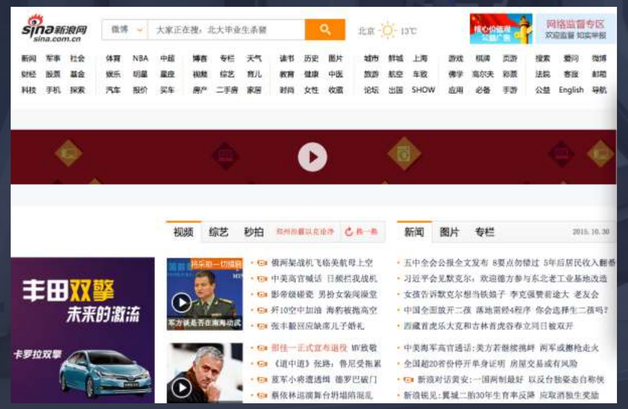
----- *《钛媒体》*



互联网时代内容分发形式的变革

今日头条

门户



社交媒体&社交网络

推荐引擎

智能算法给内容分发带来的红利

今日头条

平台类型	优势	缺陷
门户	人工精选内容，质量有保障	分发效率低，日均分发数百条内容，用户长尾兴趣无法被满足
社交媒体	个性化，内容分发效率高	信噪比低，充斥大量不感兴趣的内容
社交网络	个性化，内容分发效率高，互动性好	信噪比低，充斥大量不感兴趣的内容
算法推荐	个性化，内容分发效率高	需要更多，更好的数据

字节跳动是移动互联网成长最快的公司之一

开创最早　技术最领先　用户规模最大

一直被模仿，从未被超越：大量的“山寨”头条追赶，其中不乏互联网巨头，均落后于头条。已有美、日、印尼等国的互联网公司宣称自己是“xx（本国）”的今日头条。

2012. 3
字节跳动成立

2012. 8
今日头条APP上线

2012. 12
日活跃用户超过100万

2013. 5
B轮融资

2013. 7
日活用户超过1000万

2014. 6
C轮融资

2014. 12
超越所有国外同行

2015. 11
日活超过3000万
推出短视频平台

2016. 8
用户规模超过5. 5亿
日活跃用户超过6000万

今日头条

传统内容分发平台纷纷拥抱算法推荐

今日头条

除了以今日头条为代表的新兴智能推荐平台，传统新闻APP，浏览器，搜索应用，社交平台（Facebook，微博）纷纷上线资讯智能推荐功能

《Facebook披露信息流排序方式：发布新功能》 2013 新浪科技
《Twitter调整消息流排序 不再严格按时间顺序》2016 新浪科技

Twitter近年股价变化

Facebook近年股价变化

基于智能算法的内容推荐

今日头条

推荐系统的核心算法可以根据用户标签，内容标签和情景信息，计算用户对内容感兴趣的概率$P(y|x_u, x_i, x_c)$

模型优化的目标是什么？

今日头条

负反馈

用户点击

页面停留时间

关注

点赞

评论

收藏

分享

多目标如何融合？

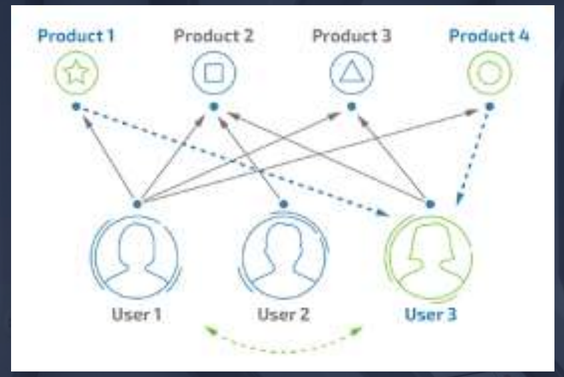
今日头条

$$P(y|x_u, x_i, x_c) = \prod \left(P_i(y|x_u, x_i, x_c) \right)^{\delta i} / Z$$

- 以提升所有正向行为的概率为目标
- 用指数权重衡量各个目标的重要度
- 不断调整权重尽量使得所有目标都有提升
- 有没有一个终极目标可以代替多目标？

典型推荐算法

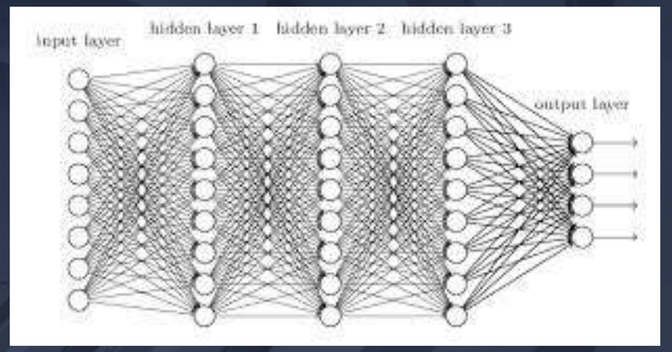
今日头条



协同过滤

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic Regression

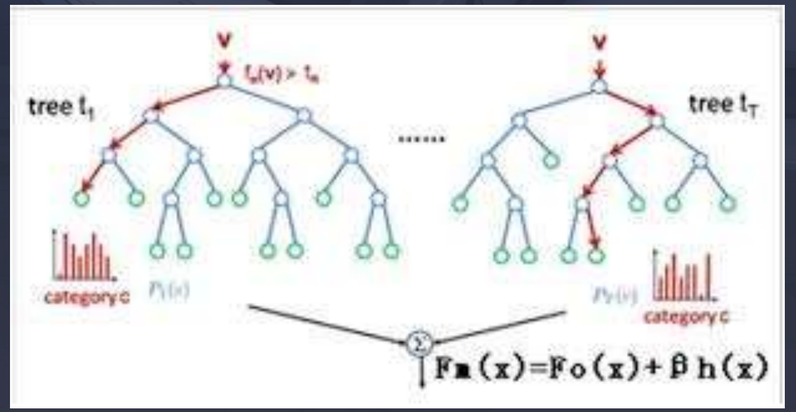


DNN

$$\check{y}(x) = w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} x_j x_{j'} \langle v_j v_{j'} \rangle$$

$$\check{y}(x) = w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} x_j x_{j'} \sum_{f=1}^{k} v_{f,j} v_{f,j'}$$

Factorization Machine



GBDT

典型推荐特征

今日头条

相关性特征

关键词匹配
分类匹配
主题匹配
来源匹配

上下文特征

上一刷内容
最近N次推荐内容
最近N天推荐内容

环境特征

地理位置
时间

热度特征

全局热度
分类热度
主题热度
关键词热度

协同特征

点击相似用户
兴趣分类相似用户
兴趣主题相似用户
兴趣词相似用户

Bias特征

用户先验点击率
用户性别
用户年龄

基于智能算法的文本内容分析

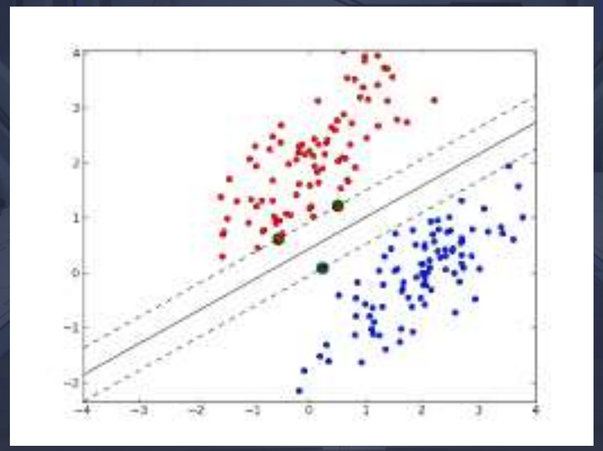
今日头条

文本内容自动分析是新闻推荐系统的基石，主要应用包括精细分类，主题分析和实体词提取


拜仁失利球迷怀念瓜帅 安帅命门恰是瓜帅最大优势

骇客足球 2016-11-20 16:17

客场0-1不敌多特，拜仁将榜首的位置暂时交给了升班马莱比锡RB，虽然赛季还很漫长，现在谈论争冠还为时过早，但是球迷们已经逐渐意识到，这支拜仁和以往不一样，他们没有了以往那样在德甲的统治力，甚至有球迷已经开始怀念瓜迪奥拉时代的拜仁。

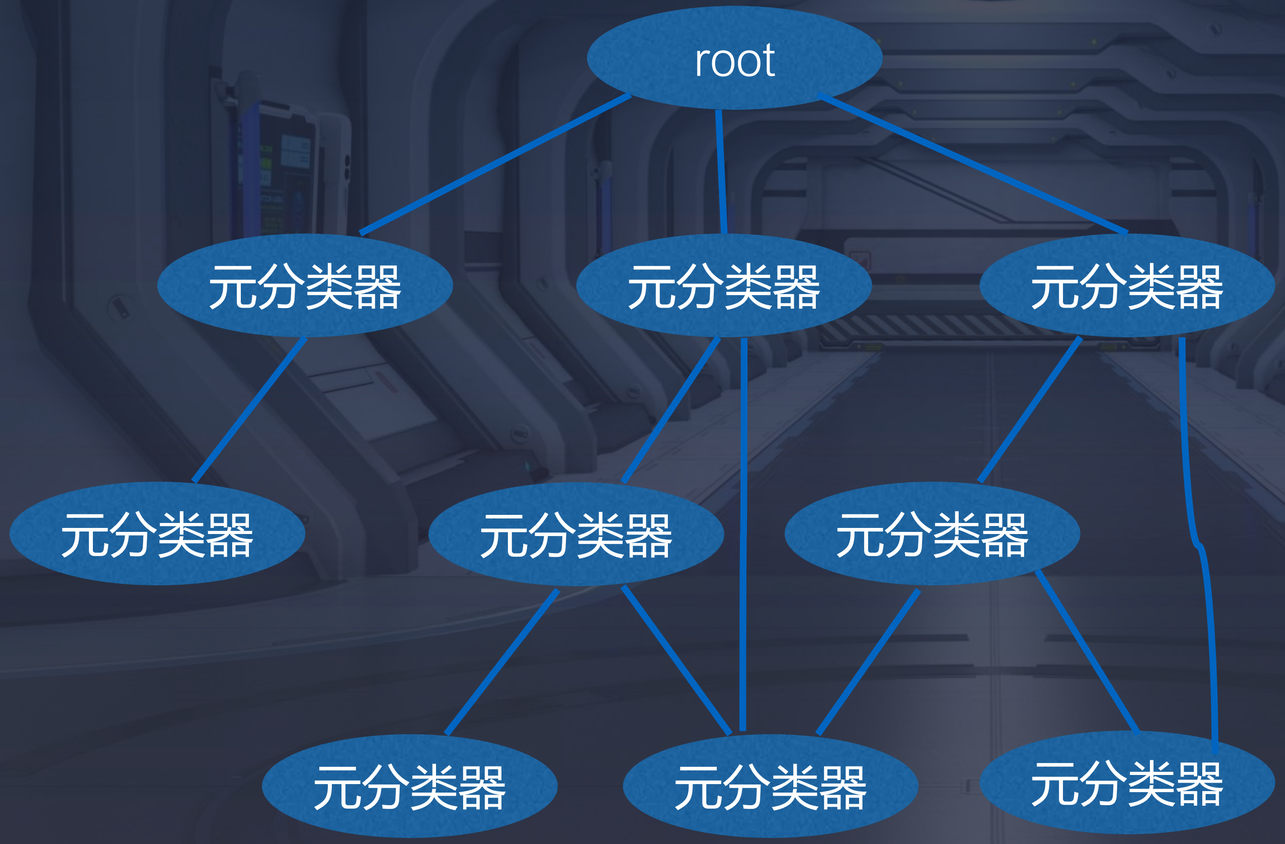


体育/足球/德甲

拜仁，安切洛蒂，多特蒙德

典型的层次化文本分类算法

今日头条



元分类器类型：

- SVM
- SVM + CNN
- SVM + CNN + RNN

实体词识别算法

今日头条

英超-利物浦0-0曼联，德赫亚频频开挂

原创 肆客足球 2016-10-18 07:54

北京时间10月18日凌晨03:00，2016-17赛季英超联赛第八轮焦点战打响，红军利物浦坐镇安菲尔德球场迎战红魔曼联，上演两队第197次双红会。上半场，红军采用高压反抢限制曼联进攻，在高空球方面，曼联则占据优势。半场双方互无建树。易边再战，双方攻势渐起，德赫亚两次神扑将利物浦极具威胁的进攻化解。全场战罢，双方0-0握手言和。积分榜上，利物浦落后榜首的曼城2分排在第4，曼联积14分排在第7位。

分词&词性标注

英超 N 利物浦 N 0-0 曼联 N ，德赫亚 N 。

抽取候选

英超联赛
利物浦足球俱乐部*
利物浦市*
曼联俱乐部
德赫亚
。。。

去歧

英超联赛
利物浦足球俱乐部
曼联俱乐部
德赫亚
。。。

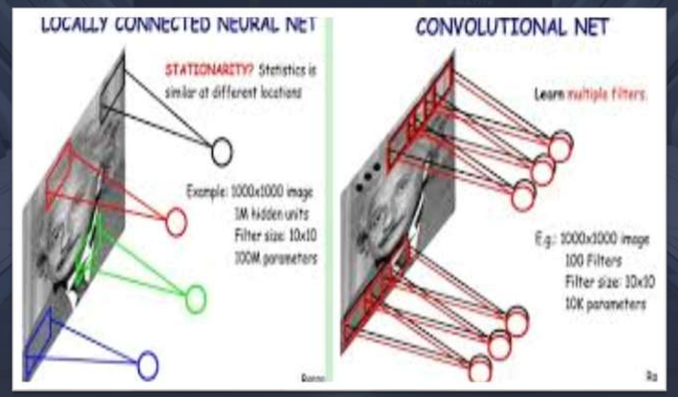
计算相关性

新版实体词	展开>>
大卫·德赫亚	0.9973
利物浦足球俱乐部	0.9899
曼彻斯特联足球俱乐部	0.9835
英格兰足球超级联赛	0.9565
兹拉坦·伊布拉希莫维奇	0.6718
卢克·肖	0.6559
韦恩·鲁尼	0.6387
埃姆雷·詹	0.6320
保罗·博格巴	0.6196
迈克尔·卡里克	0.5185

基于智能算法的图像识别

今日头条

对全自动化的智能推荐引擎而言，准确快速的图像识别对于分析内容特征，广告色情识别至关重要



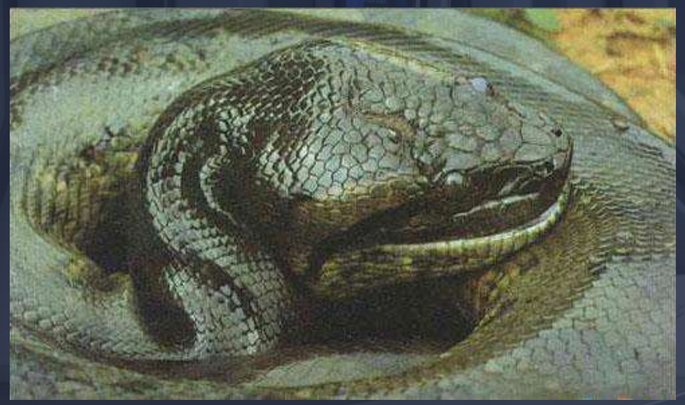
奥巴马

美国

国际

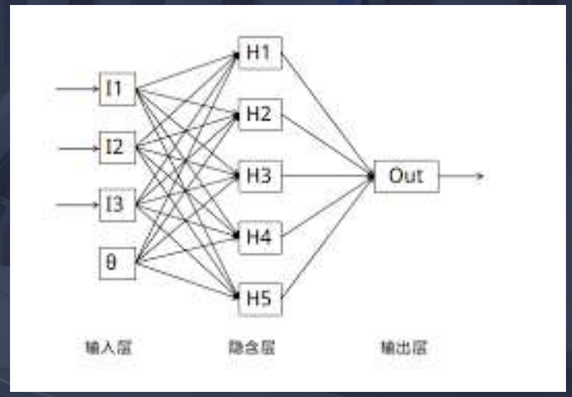
一种识别易引起不适图片的算法

今日头条



标注样本图片

基于ImageNet训练好的CNN抽取向量



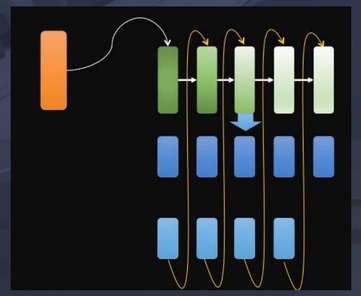
训练NN分类器

基于智能算法的写稿机器人

今日头条

人工智能已经可以在财经报道，体育赛事报道等领域自动创作内容，可读性完全可以媲美人工编辑

图片库



直播语料素材

算法辅助视频封面选择

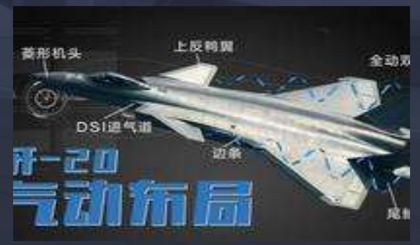
今日头条

封面选择对视频的点击率有重要影响，智能算法可以自动给出封面建议，减少视频上传者的选择成本

预估点击率 0.1

预估点击率 0.3

预估点击率 0.2



预估点击率 0.2

算法自动生成视频集锦

今日头条

智能算法可以从体育比赛，MV等长视频中自动抽取精彩片段，甚至生成gif，可以节省用户时间和流量


Q&A