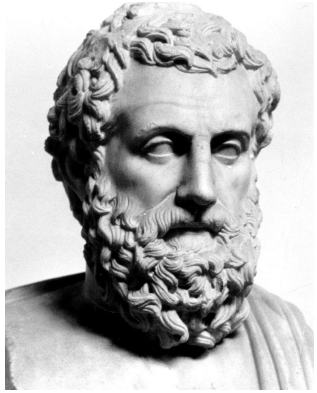
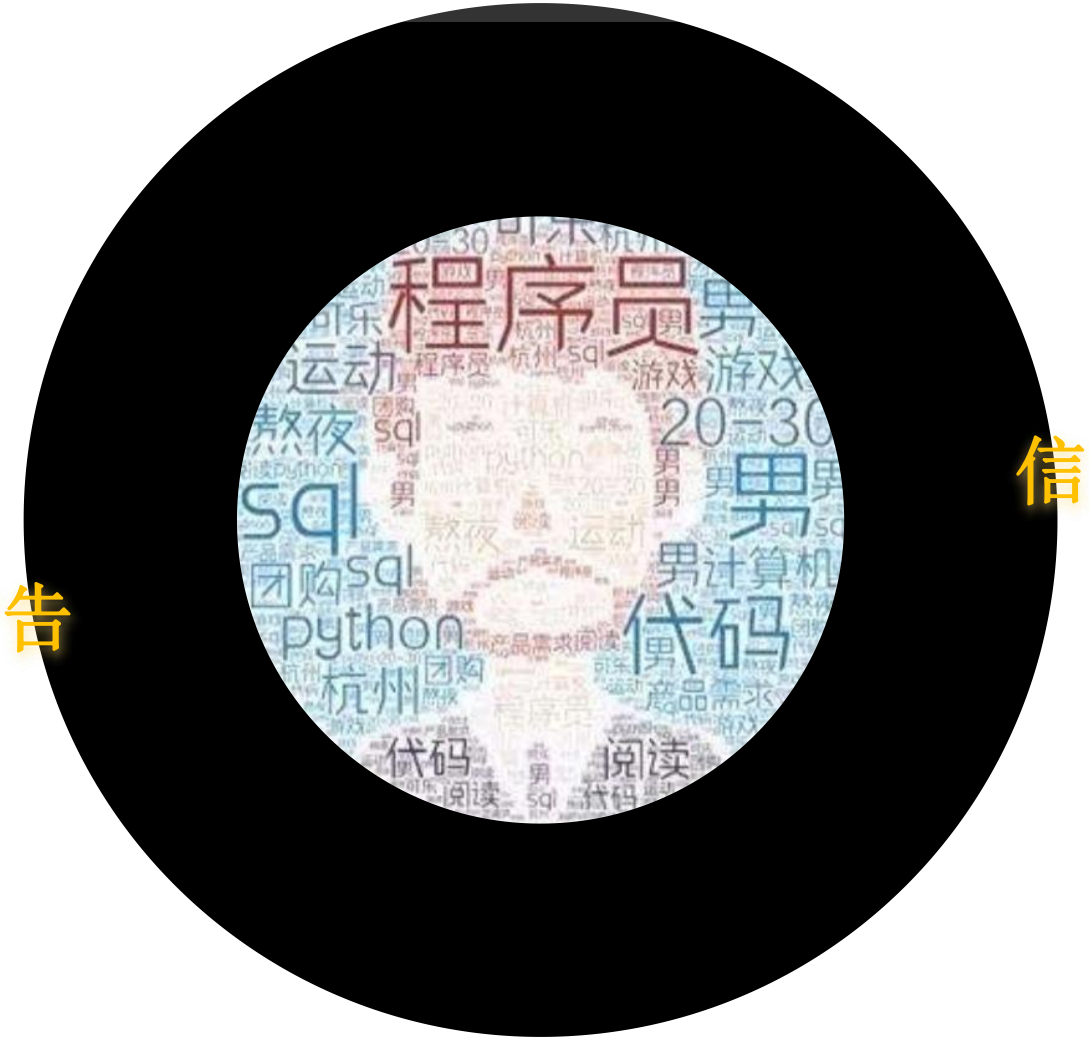
RS&NLP选修
Lesson-01

xx老师& 开课吧人工智能学院课程组

2019.9

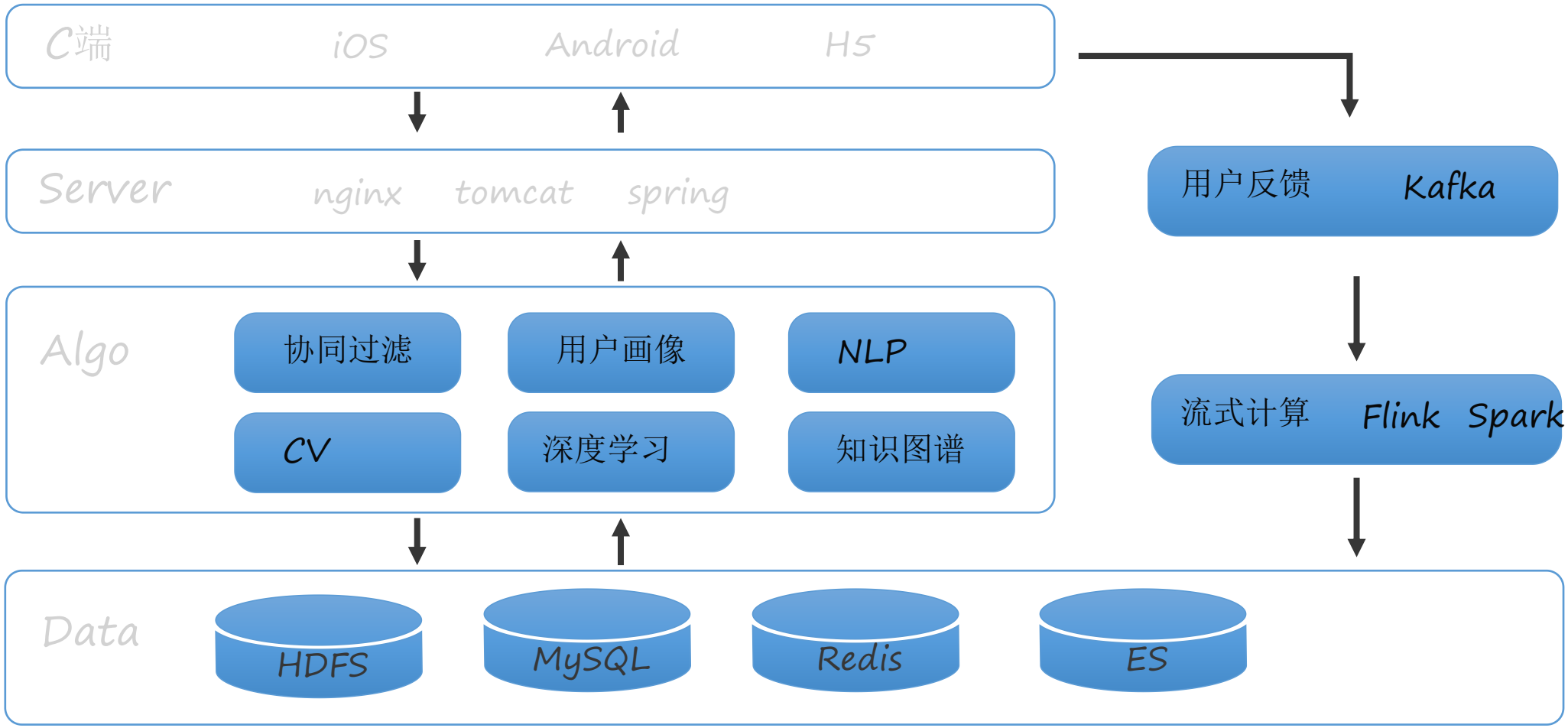


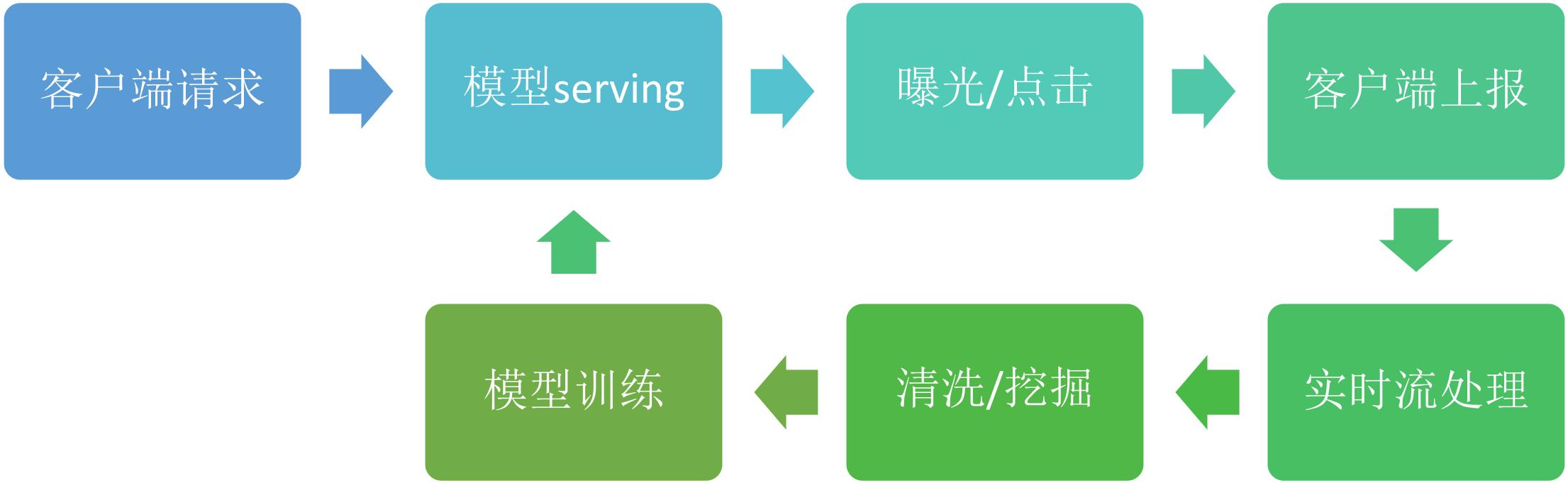


计算广告

信息流推荐

1/4 推荐系统大有不同


C端
iOS
Android
H5
Server
nginx
tomcat
spring
Algo
协同过滤
用户画像
NLP
CV
深度学习
知识图谱
Data
HDFS
MySQL
Redis
ES
用户反馈
Kafka
流式计算
Flink
Spark


客户端请求
模型serving
曝光/点击
客户端上报
实时流处理
清洗/挖掘
模型训练

“用户画像是真实用户的虚拟代表，是建立在一系列真实数据之上的目标用户模型。” by Alan Cooper

文本特征case

查找文章: 4688699423 提交

4688699423 莎娃连续17次不敌小威 07-10 13:18 rate:18 展开>>

文章Profile

一级分类	展开>>
news_sports	2.5957

二级分类	展开>>
news_sports/tennis	0.7201

关键词2							展开>>
西班牙	0.9915	小威	0.9858	穆古拉扎	0.9845	女单决赛	0.9641
俄罗斯	0.9475	莎拉波娃	0.9282	莎娃	0.9208	小威廉姆斯	0.9199
委内瑞拉	0.8738	锦标赛	0.7582	温网	0.6409	大满贯	0.5660
半决赛	0.4663						

高亮关键词							展开>>
西班牙	0.9976	莎拉波娃	0.9886	俄罗斯	0.9856	小威廉姆斯	0.9831
委内瑞拉	0.9823	小威	0.9498	穆古拉扎	0.9463	温网	0.9323
半决赛	0.7198	女单决赛	0.7114	大满贯	0.6948	波兰	0.6094

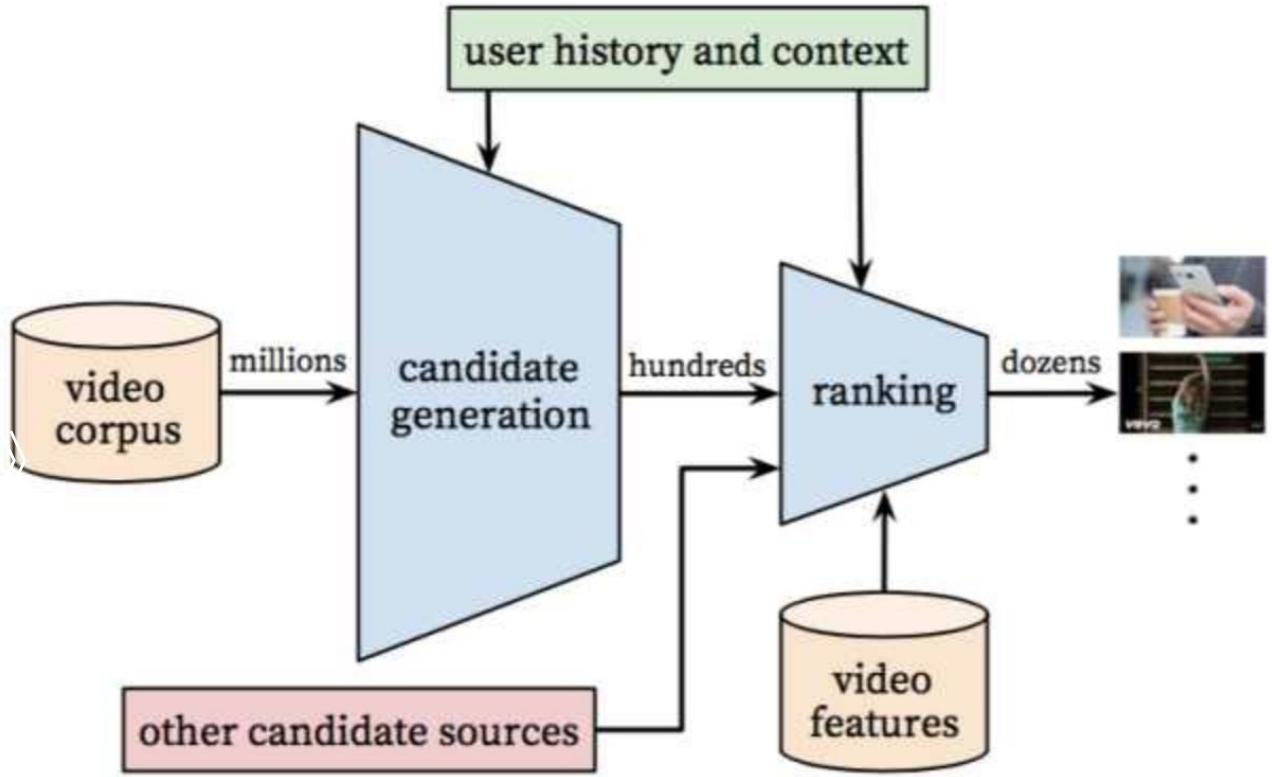
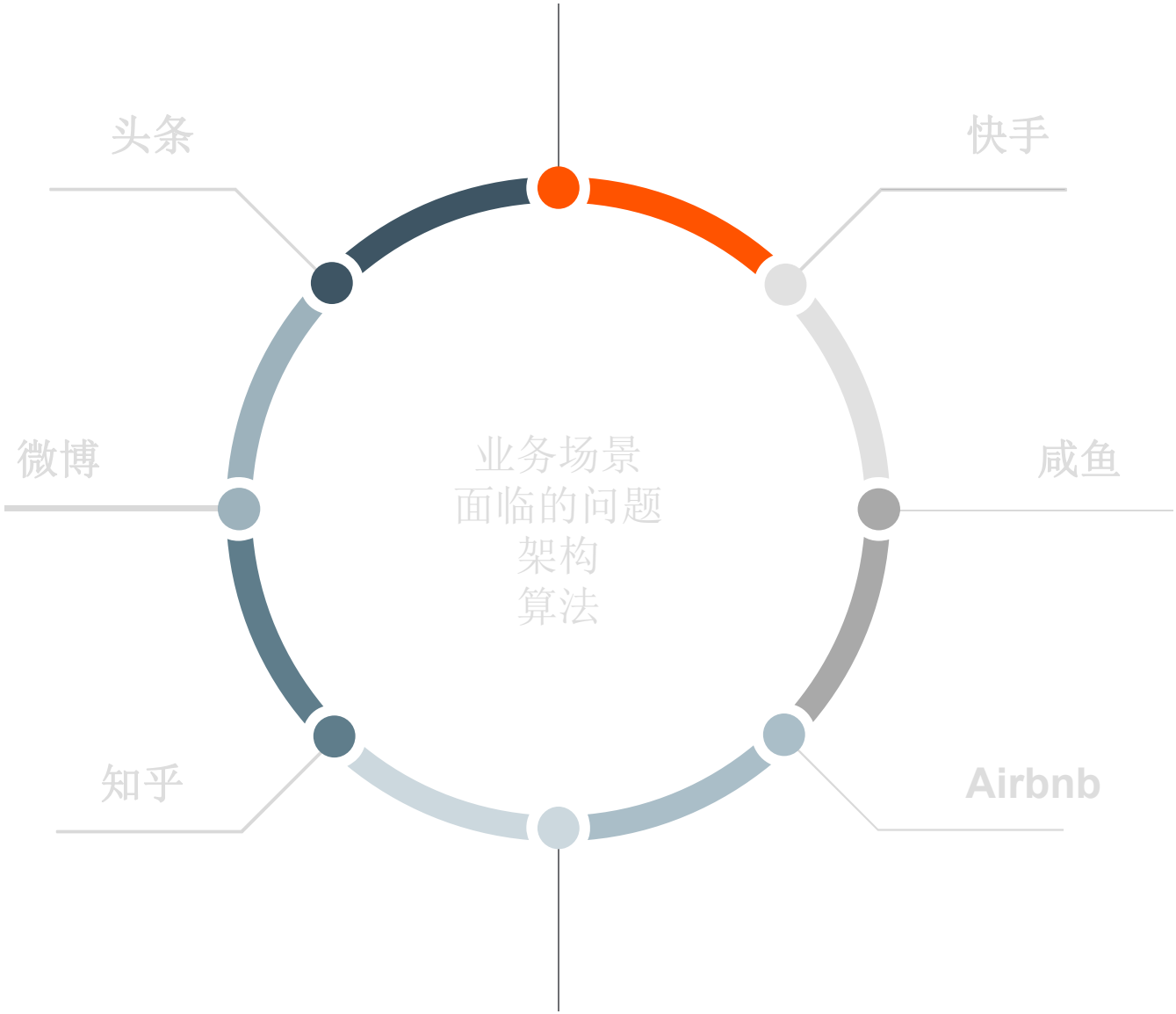


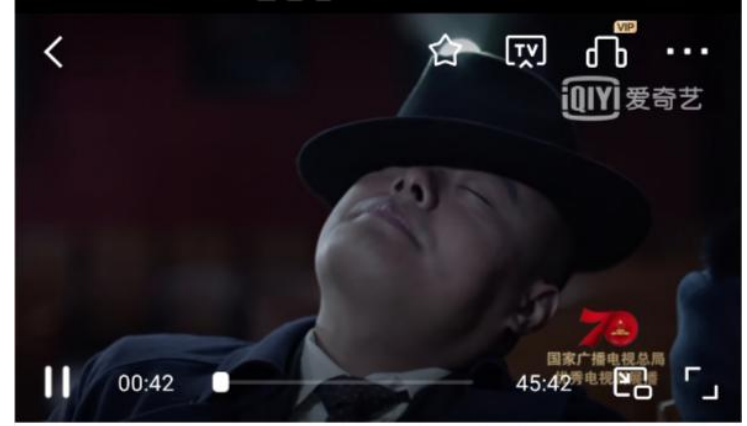
Figure 2: Recommendation system architecture demonstrating the "funnel" where candidate videos are retrieved and ranked before presenting only a few to the user.


头条
快手
微博
咸鱼
业务场景
面临的问题
架构
算法
知乎
Airbnb

“不必拘泥于架构，而是要把业务抽象成模型，进而探求解决方案”

By Gerrard

2/4 从实际业务出发



视频　　评论 893

猜你喜欢

铁血红安

侠义豪情 战将传奇

热度1674

铁血尖刀

英雄尖刀队

热度1793

春去春又回

民国版基督山伯爵

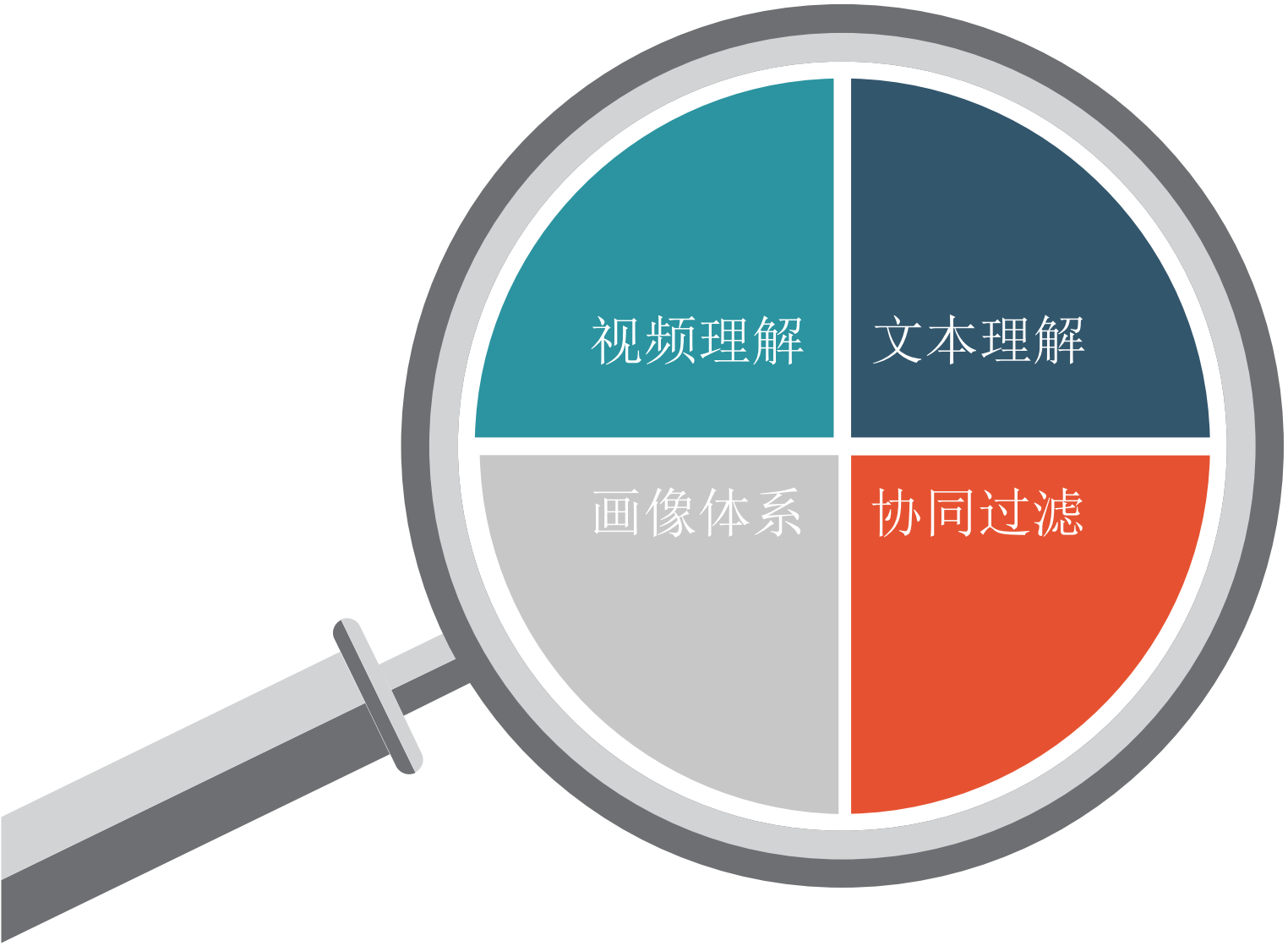
热度1720

为了新中国前进

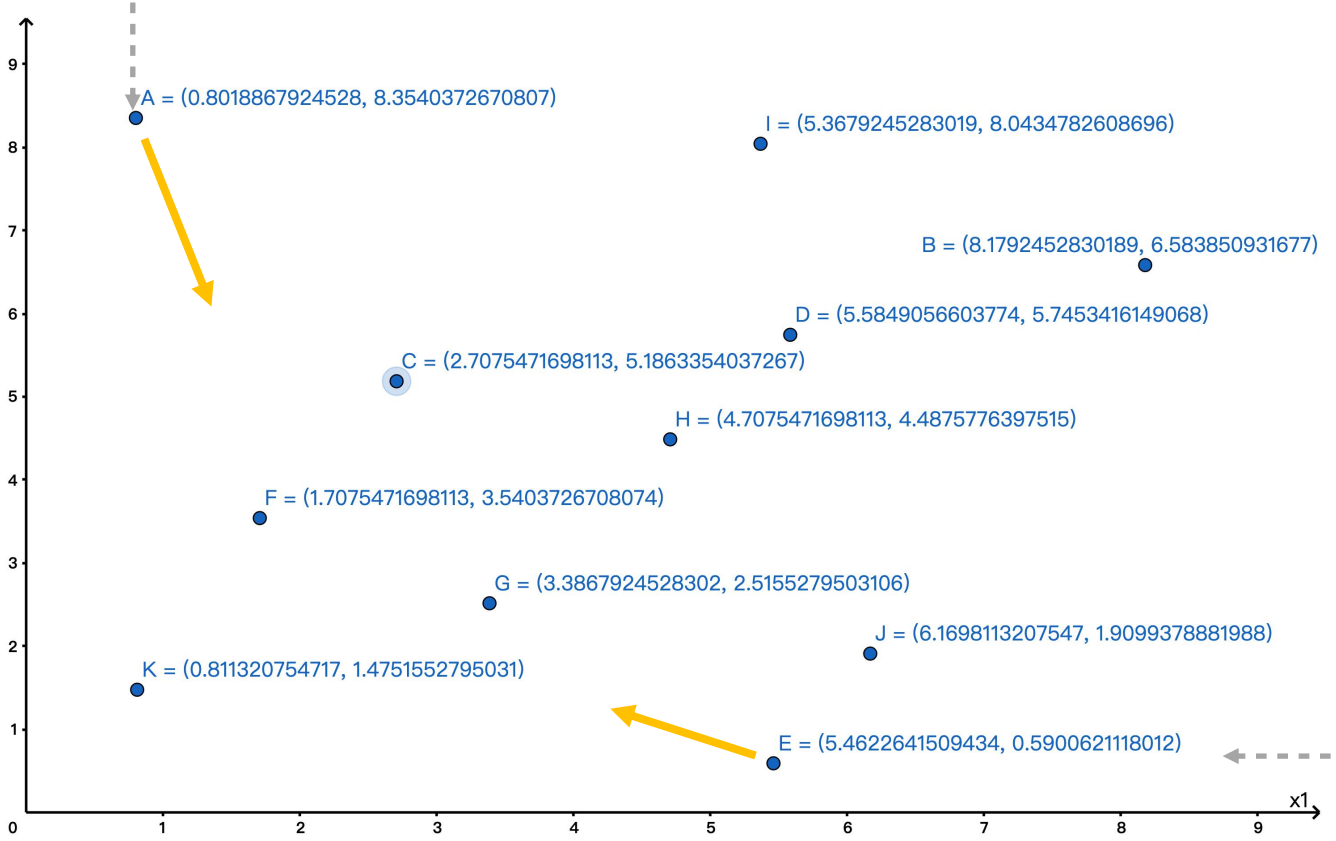
打造"中国兄弟连"

热度1610

鬼谷门


视频理解
文本理解
画像体系
协同过滤

时间戳	Userid	Itemid	Title	……
1571987459	6789	7263898123	谍战深海	……
1571823459	28376	9184798217	铁血红安	……
……	……	……	……	……



「量子波动速读」这种明显的骗局，家长仍...

刘欢：遇事不决，量子力学 解释不通，穿越时空 脑洞不够，平行宇宙 画面老土，追求复古 不懂配色，萨博朋克...

224 赞同 · 22 评论

如何看待量子波动速读的出现？

甲烷：存在，原理如下 受训者（简称a）与培训者的母亲（简称m）进入量子纠缠态，进行洛伦兹变换，压缩m...

136 赞同 · 22 评论

北京量子波动速读是什么妖魔鬼怪？

窗边的小豆豆：为了利益不择手段

0 赞同 · 0 评论

wwq88：只见过量子粒子群优化算法，[强]的很，但这个～是个什么玩意[捂脸]咋不上天呢

0 赞同 · 0 评论

白嘉禾：[图片] 绍兴量子波动速读是什么妖魔鬼怪？ [图片] 湖南量子波动速读是什么妖魔...

0 赞同 · 0 评论

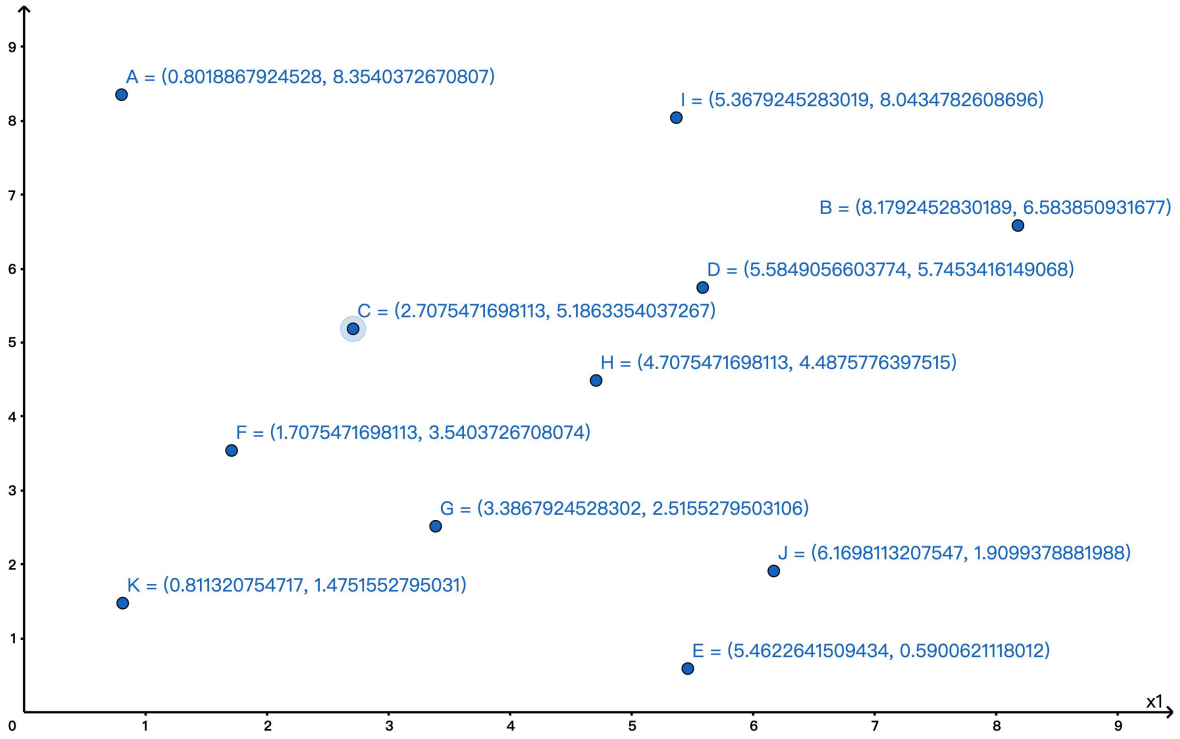
PARADISEDA：已经不是交智商税这么简单了，是简直侮辱人的智商。这种东西还有人上当，是真没救了。

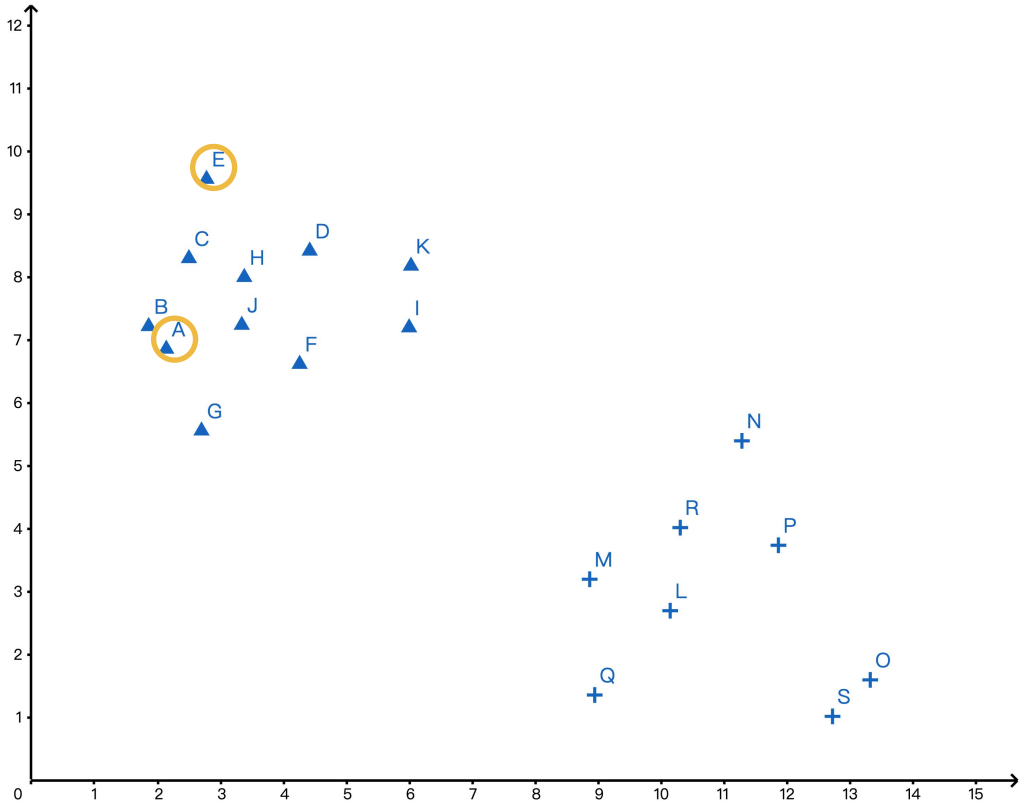
9 赞同 · 0 评论

通过模型计算
越相近*item*距离越近

取最邻近的内容
作为候选集

建立倒排索引
快速找到相似内容





$$p(w_j \mid w_i) = \sigma(u_i^T v_j) \prod_{k=1}^{N} \sigma(-u_i^T v_k)$$

where $\sigma(x) = 1/1+\exp(-x)$, N is a parameter that determines the number of negative examples to be drawn per a positive example. A negative word w_i is sampled from the unigram distribution raised to the 3/4rd power. This distribution was found to significantly outperform the unigram distribution, empirically [8].

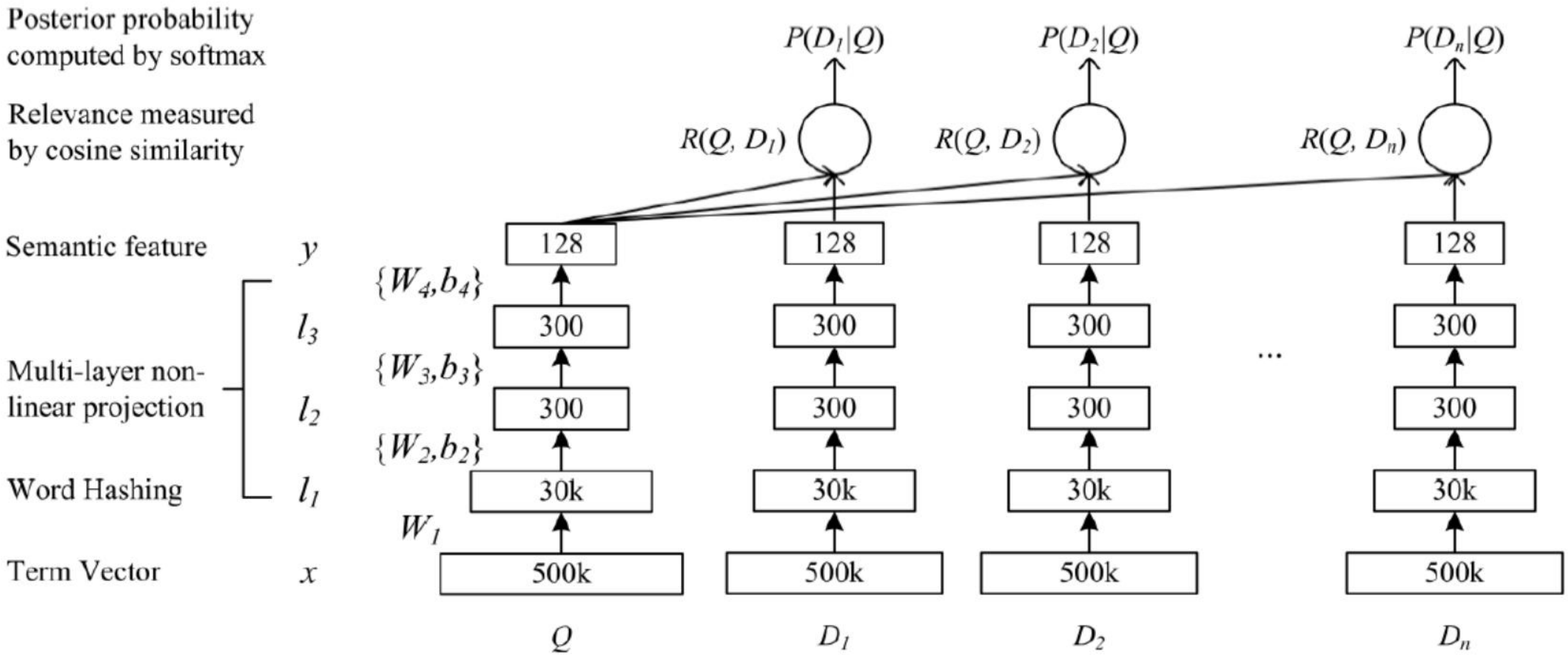
工具	描述	作用	替代
Spark	分布式数据处理框架	将行为日志处理成特定格式	Hive/Presto/MR/python
Gensim	Python算法包	利用Word2vec算法求相似性	Spark/其他开源实现
Faiss	FB家开源的相似性搜索库	为item的向量建立索引	Annoy

3/4 NLP&RS的千丝万缕

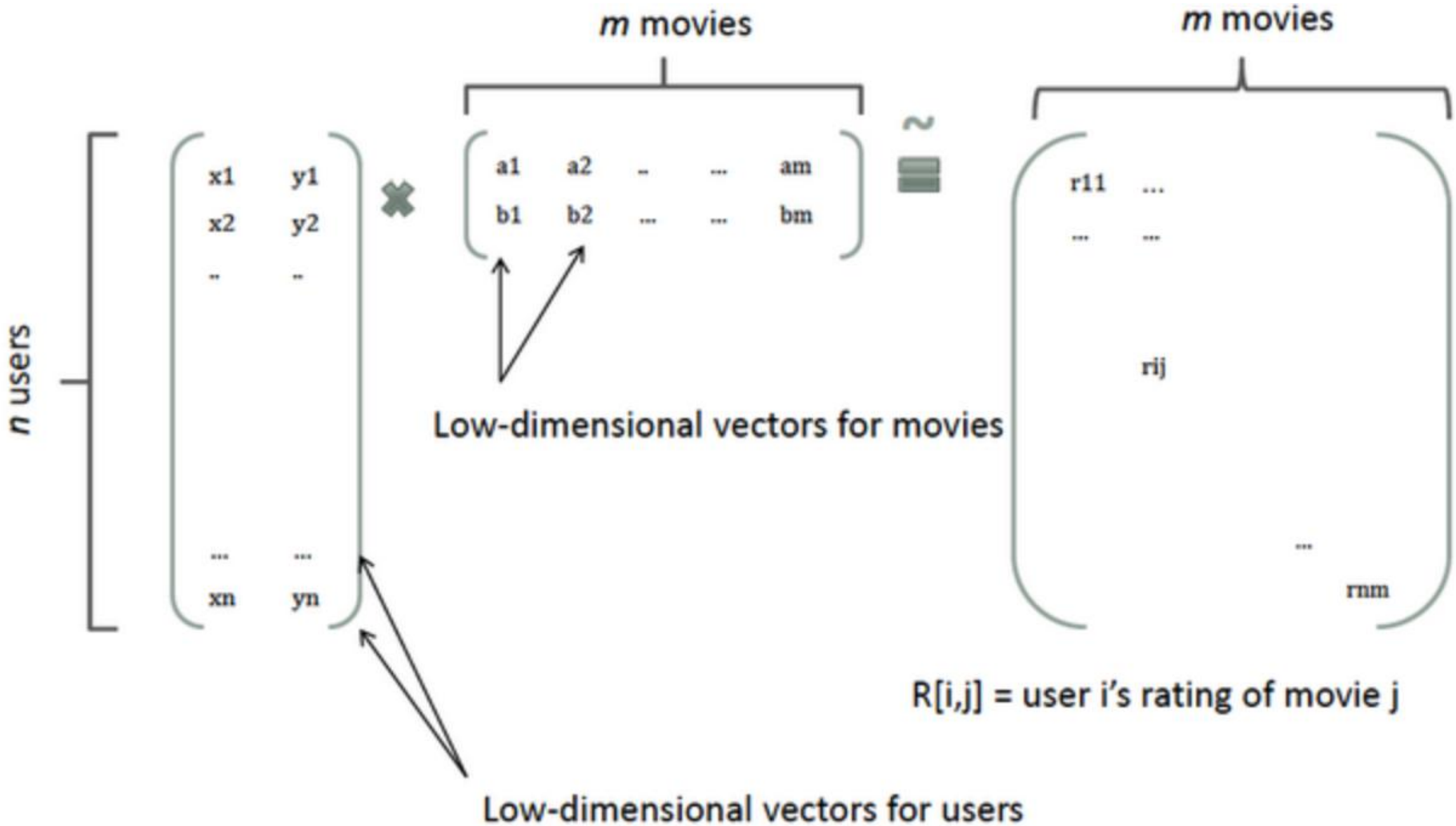
对比项	RS	NLP
基本假设	邻近的浏览，兴趣点相似	context相似的word，有相似的语义
正例	滑窗内的2条浏览记录	滑窗内的2个单词（w, c）
负例	在用户画像内随机采样	在语料库中随机采样
下采样	CTR过高的热门内容	停用词等无实际意义的单词

“NLP和RS是两个具体的应用方向，其背后的思想有很多共通之处”

By Gerrard


Posterior probability computed by softmax
Relevance measured by cosine similarity
Semantic feature
y
Multi-layer non-linear projection
Word Hashing
Term Vector
x
l_3
l_2
l_1
W_1
$\{W_2,b_2\}$
$\{W_3,b_3\}$
$\{W_4,b_4\}$
$P(D_1|Q)$
$P(D_2|Q)$
$P(D_n|Q)$
$R(Q, D_1)$
$R(Q, D_2)$
$R(Q, D_n)$
128
300
30k
500k
...
Q
D_1
D_2
D_n

4/4 回到实际场景


m movies
m movies
n users
x1 y1
x2 y2
a1 a2 ... am
b1 b2 ... bm
r11 ...
rij
rmm
Low-dimensional vectors for movies
R[i,j] = user i's rating of movie j
Low-dimensional vectors for users

$$M_{ij}^{\text{SGNS}} = W_i \cdot C_j = \vec{w}_i \cdot \vec{c}_j = PMI(w_i, c_j) - \log k$$