

# 大规模知识图谱的构建及应用

抖伟雨  
移动浏览产品部  
腾讯

跟着兴叔走不丢

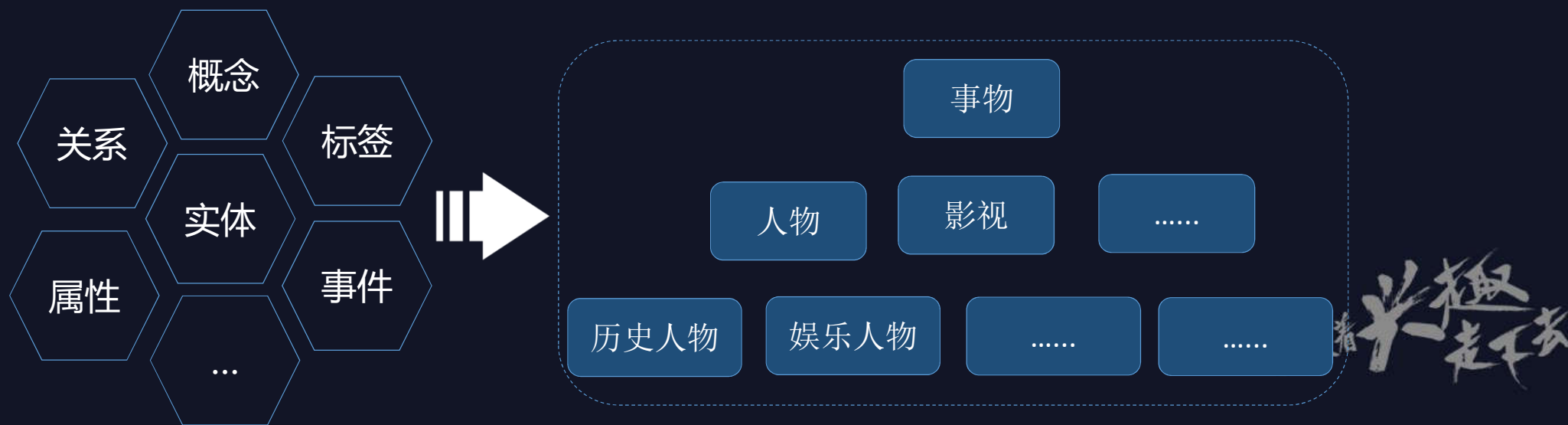
## 目录

- 大规模知识图谱构建
  - 知识图谱简介
  - 知识图谱构建技术
  - 知识图谱目前的使用场景
- 知识图谱在自然语言理解的应用

跟着兴趣走

# 知识图谱 - 汇聚知识，理解客观世界

- Google知识图谱是Google的一个知识库，其使用语义检索从多种来源收集信息，以提高Google搜索的质量。
- 知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系。
- 辛格博士：The world is not made of strings , but is made of things.



# 手机QQ浏览器 - 我的兴趣世界



市场份额稳居行业第一



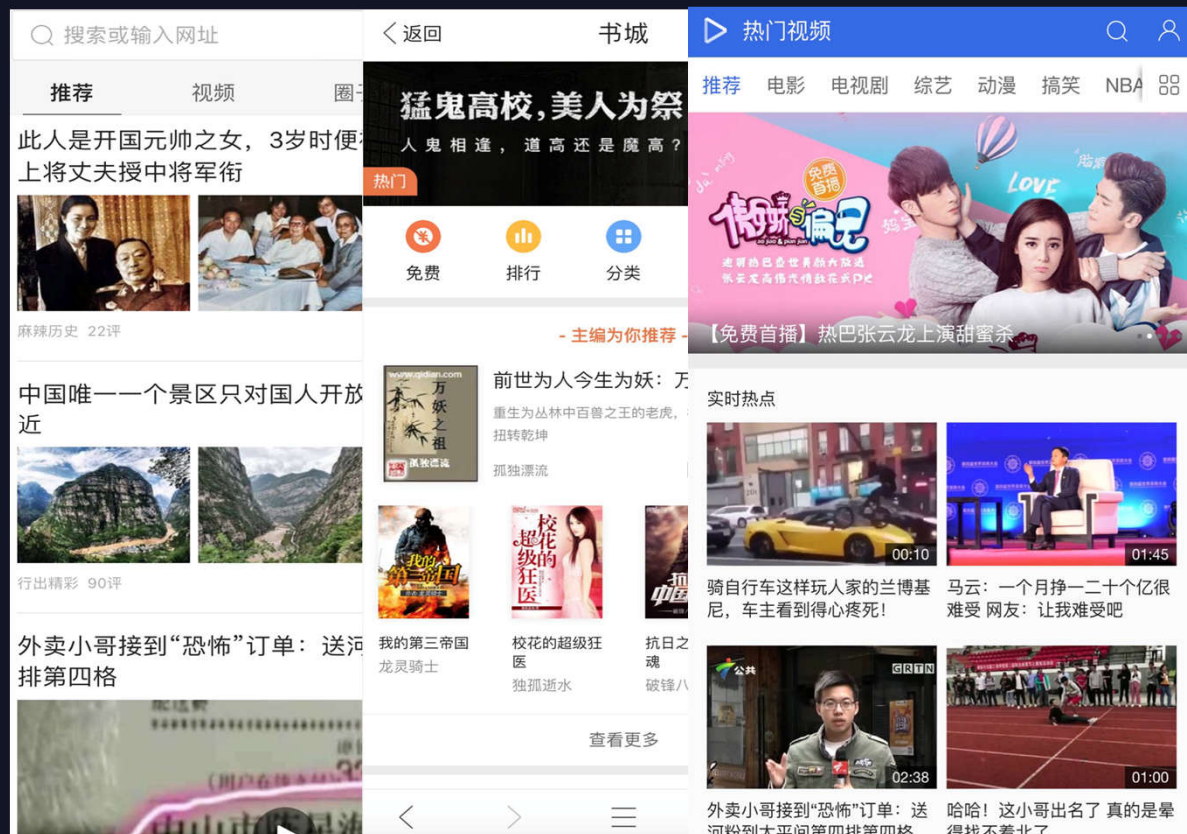
理解内容



理解用户



理解知识



# QQ浏览器知识图谱现状

## 数据规模

亿级 实体  
亿级 关系  
十亿级 三元组

## 覆盖领域

30+大领域：  
人物，影视，体育，音乐，  
文学，医疗等  
300+细分领域：  
历史人物，电影，篮球等

## 各领域垂直类 站点

200+ 站点  
百万级更新 / 天

## Query 资讯数据

百亿级流水

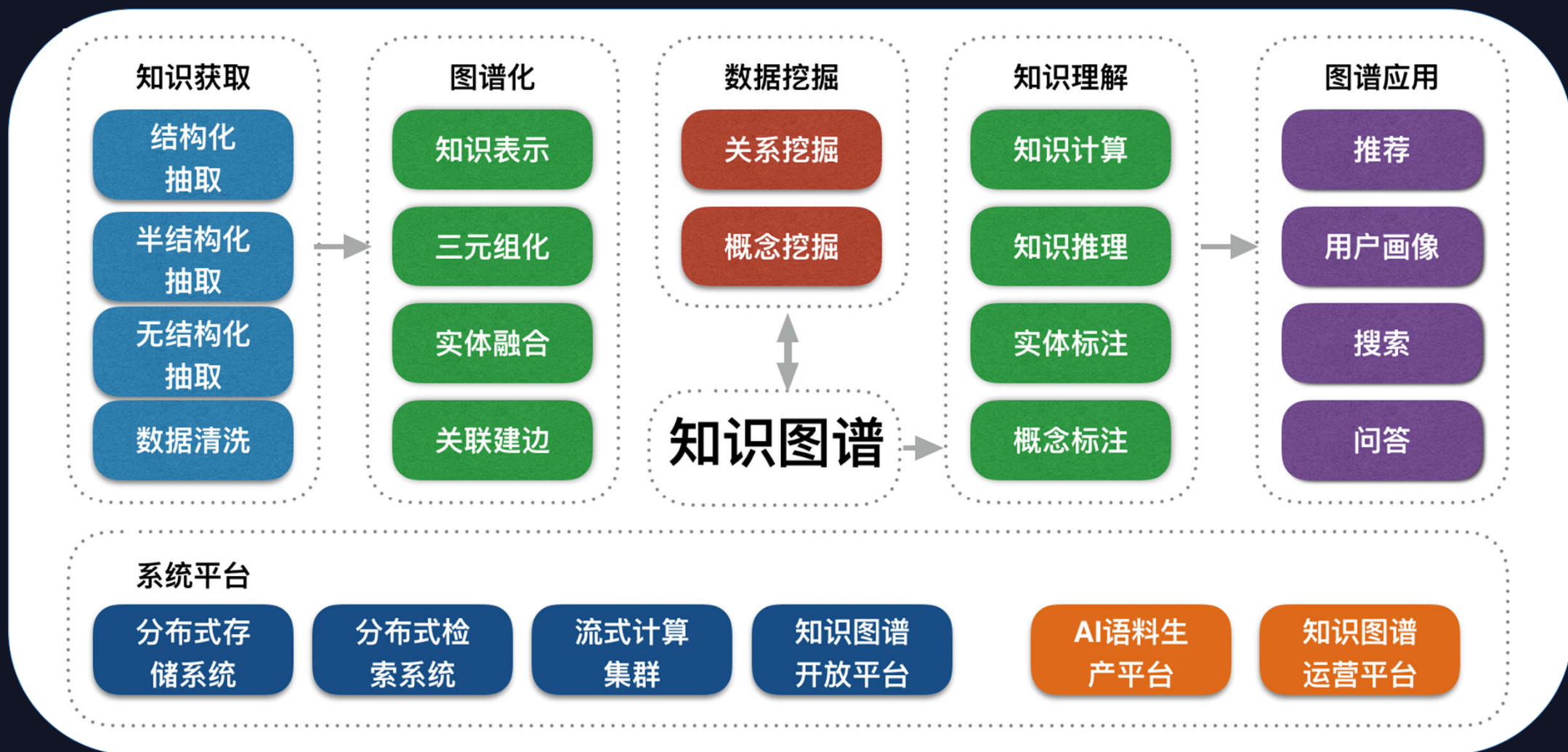
## 专业领域

医疗类目  
物种类目等

## 百科类

跟着兴趣走

# 知识图谱技术架构



# 构建技术 - 关系抽取

- 谔龙王适娴结婚
- 女友 - > 妻子

不用等年底了！谔龙王适娴今日领证结婚

腾讯体育  
2017-11-28 12:51:29



视频：谔龙求婚加长版：为国为家让我们在一起吧，时长约1分59秒



思明区民政局

搜狗百科

11 人物生活

2017年11月27日，谔龙向女友王适娴求婚成功。<sup>[37]</sup>

2017年11月28日，谔龙和王适娴领证结婚。<sup>[38]</sup>

Baidu 百科

2017年羽毛球亚洲锦标赛在武汉落下帷幕，在男单冠军争夺战先失一局的情况下连扳两局以2-1逆转获胜，从而首度夺得亚锦赛男

2017年11月27日，谔龙在厦门向王适娴求婚，预计年底领证。

人物关系 纠错

未婚妻 王适娴



## 构建技术 - 关系抽取的流程

关系定义

样本筛选

模型建立

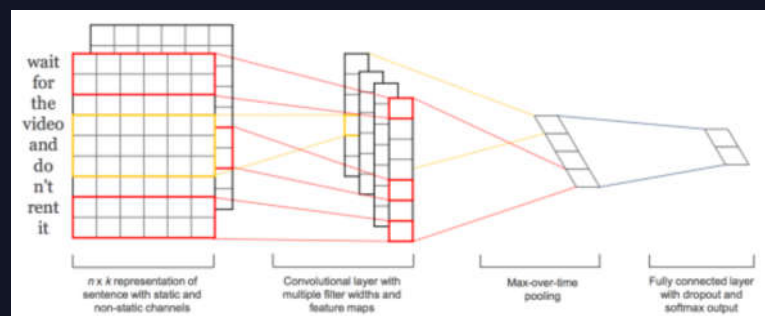
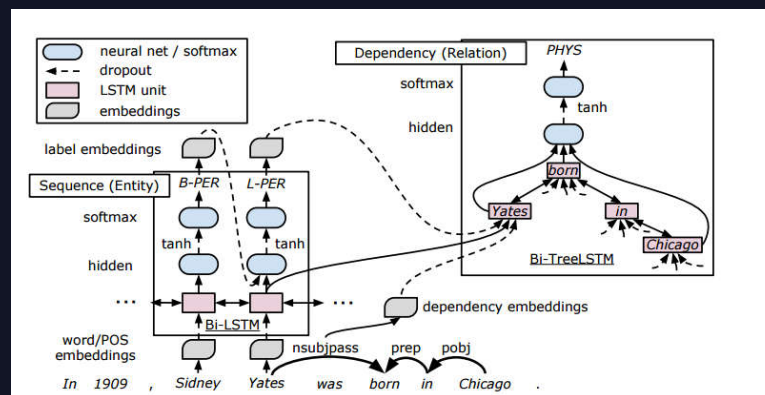
关系预测

跟着兴趣走



# 构建技术 - 关系抽取模型

- 深度学习方法（2015年）
  - 基于LSTM的关系提取
  - 问题：需要大量样本
- 优化方案
  - SPTree + LSTM
  - 降低训练样本
  - 问题：不够准确
- 基于CNN的关系提取
  - CNN 提取高阶特征：例如句法结构
  - ATT model 代替SPTree 提取关键语义特征。
- SPTree + LSTM 关系提取
  - Without Embedding (64%)
  - With Embedding 70%
- CNN 关系提取
  - Without Embedding 72%
  - With Embedding 84%
- CNN + 规则 with embedding: 96%



跟着兴一老师

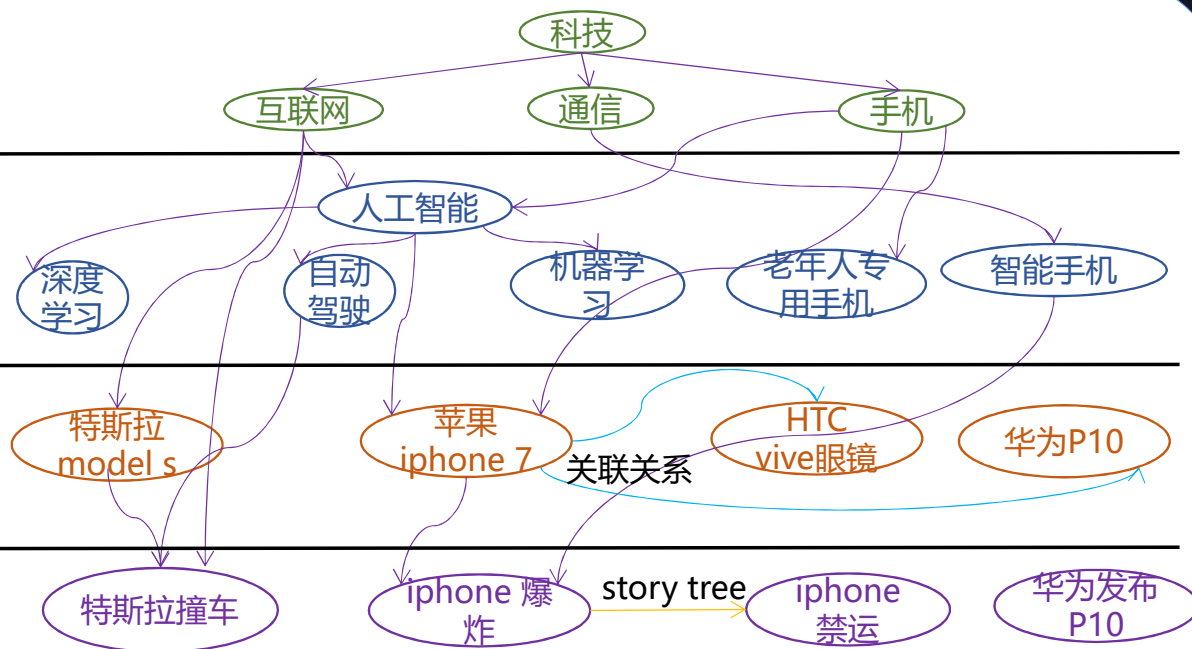
# 构建技术 - 概念图谱

主题分类层

概念层

实体层

事件层



主题分类：汽车;美系汽车  
概念：美系豪华车;  
驾驶感出众的车  
实体：凯迪拉克XT5  
事件：凯迪拉克XT5发售



# 构建技术 - 概念挖掘与上下位计算

冷启动



QA数据 → 泛需求识别 → (C,E)Pair挖掘

垂直站点 → 网页解析 → (C,E)Pair挖掘

自动样本标注

同义归一

(C,E)Pair挖掘



例行化

query

query分类

concept识别

entity识别

同义归一

$$prob(concept|query) = \sum_{t \in titles} p(concept|title)p(title|click)$$

title分类

点击权重

基于<q,t>对齐语料识别

美国|美系SUV汽车大全|最新车型报价及图片|排行榜|爱卡汽车

狂野动感|六款进口美系SUV推荐|凤凰汽车|凤凰网

美系suv推荐

等选车内容,更多美国|美系SUV选车资讯内容尽在爱卡汽车  
newcar.xcar.com.cn/car... - 百度快照

美国|美系SUV汽车排行榜|美国|美系SUV汽车排名推荐|爱卡汽车

跟着兴趣走

## 使用场景 - 直接满足用户的需求



跟着兴叔走不丢

## 使用场景 – 更智能的助手服务

### 播放音乐

“我想听菊花台”

“我想听张碧晨的歌”

“来首中国好声音的歌”



跟着兴叔走平路



## 使用场景 – 更丰富的内容形态



# 使用场景 – Feeds推荐



主题分类：汽车; 美系汽车  
概念：美系豪华车；  
驾驶感出众的车  
实体：凯迪拉克XT5  
事件：凯迪拉克XT5发售

跟着兴趣走



# The Application of Knowledge Graph On Natural Language Understanding

**Calvin Lai** 赖坤锋

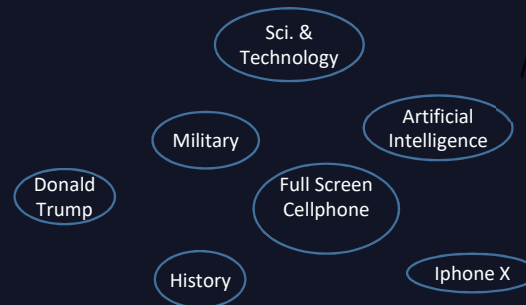
Mobile Browser Product Department  
Tencent Inc.



## Starting With Recommender



Memory Based

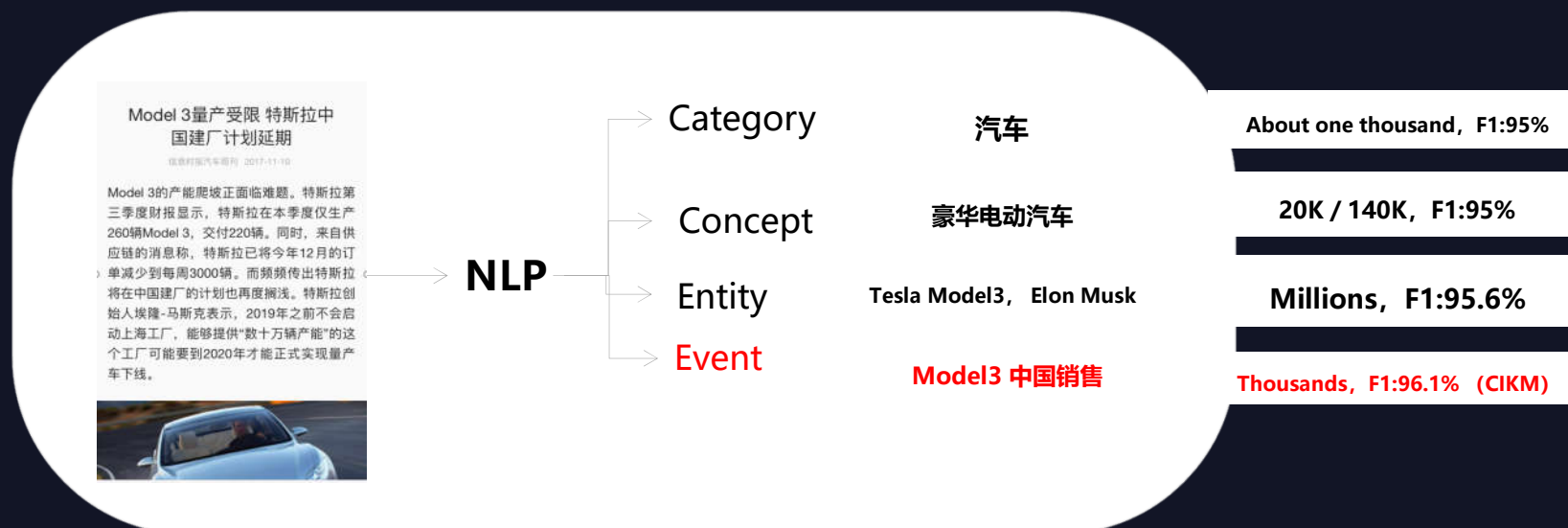


Content Based



跟着米叔走

## Understand the contents



Category: SVM & RCNN

Concept: Three methods towards concept inference

Entity: LSTM+ LP (Google 2015)

Event: CIKM 2017

跟着兴叔走不丢

## Creating the event

- Line 1: from the crawler
  - High accuracy but low recall
- Line 2: from Query-Doc pair
  - High accuracy but low recall
- Line 3: from document clustering
  - High recall, but accuracy is low

排名	关键词	搜索指数
1	南苏丹一村庄遭袭	11308
2	朝鲜试射弹道导弹	20250
3	杀妻藏尸案开庭	48369
4	张国荣追悼会	19959
5	美俄军机危险相遇	14067
6	会车时发生车祸	167079
7	刘国栋上任村长	127549
8	遼龙求偶成功	9060
9	中国增城举行出炉	263119
10	黄子韬高调改名	3726

排名	关键词	搜索指数
1	Jasper 李翊	611157
2	何穗与程璧影合影	505479
3	甜食找乐子	390442
4	范冰冰收工和工作人员聚餐	270711
5	林俊杰伟大的渺小	251365
6	有一款手机 让你戒掉手机	204003
7	2岁男童已经转秀	201394
8	阿拉雷 实力	180729
9	羽绒服的正确穿法	166362
10	车上装了1.4万双阿迪达斯	155691

### 朝鲜试射导弹

#### 朝鲜试射导弹的最新相关信息

日本称已截获朝鲜导弹发射信号 近日或将再射导弹



来源: 观察网 据日本媒体当地时间11月28日凌晨报道, 日本政府官员声称, 根据最近朝鲜发出的无线电信号判断, 朝鲜很可能于几天内再次进行远程导弹试射...

新浪 1天前

朝鲜发射新型洲际导弹 金正恩亲笔签发... 网易

2小时前

朝鲜发射新型洲际导弹 金正恩亲笔下令... 网易

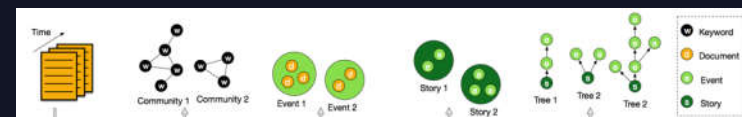
3小时前

亚盘汇市 朝鲜发射导弹避险货币反应有限... 理财18

1小时前

金价微涨 朝鲜发射导弹抵消美元与美股走... 和讯网

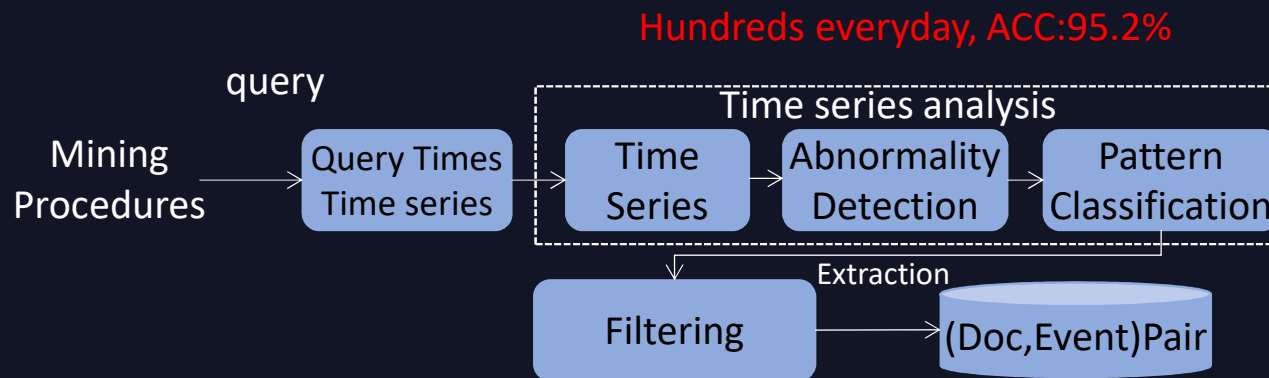
8小时前



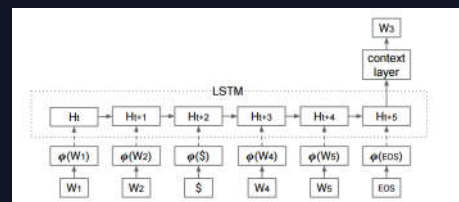
Bang Liu. Growing Story Forest Online from Massive Breaking News. CIKM. 2017.

跟着兴一老干

## About Line 2: Query-Doc event generation



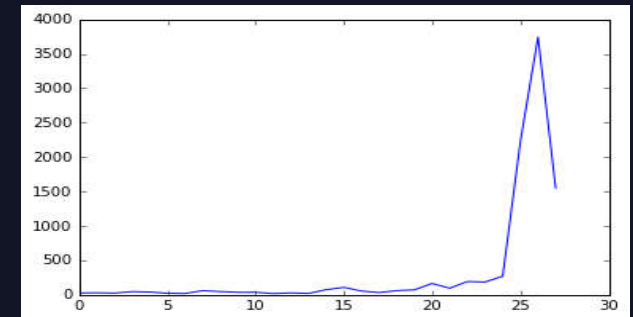
$$Prob(news|query) = \sum_{t \in titles} P(news|t) P(t|query)$$



Click rate

## Time Series Patterns

- Seasonal: e.g. 163 mailbox
- Hot topic: e.g. Kejie's war with Alphago
- Stable: e.g. Novels
- Cold topic: e.g. long tailed topic

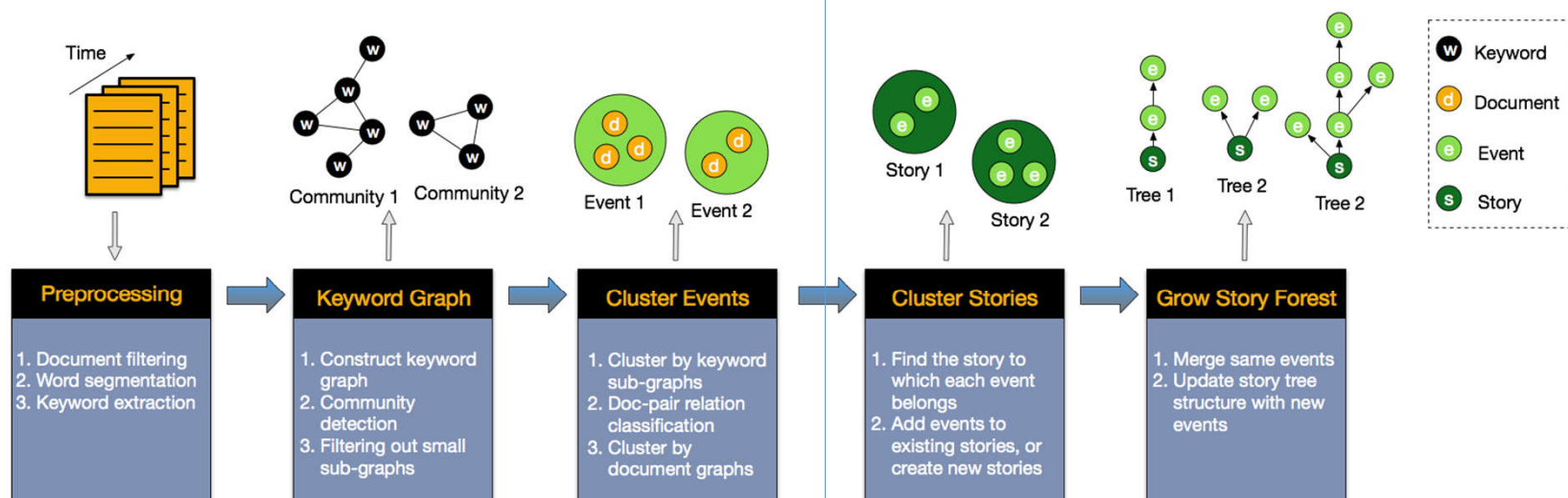


跟着兴一老干

## About Line 3: Story Systems

Cold start stage

Incremental update stage



跟着兴趣走

## The evaluation of story systems

- **Flat Cluster:** cluster by stories, no structure.
- **Story Timeline:** organizes events linearly by time.
- **Story Graph:** calculates a connection strength for each pair of events and connect the pair if the score exceeds a threshold.
- **Event Threading:** appends each event to its most similar earlier event. Similarity measured by TF-IDF.

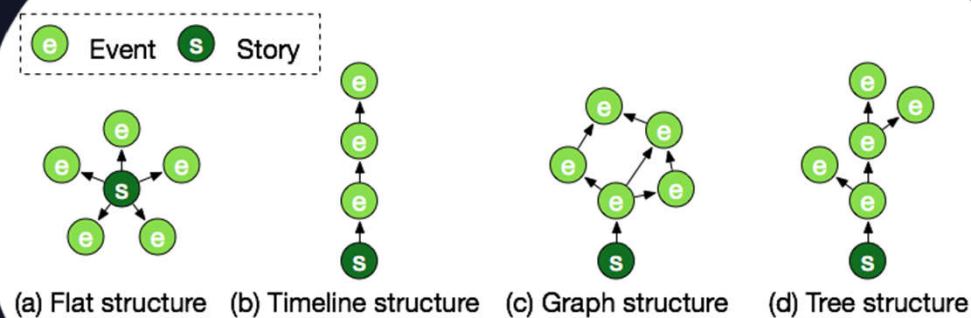


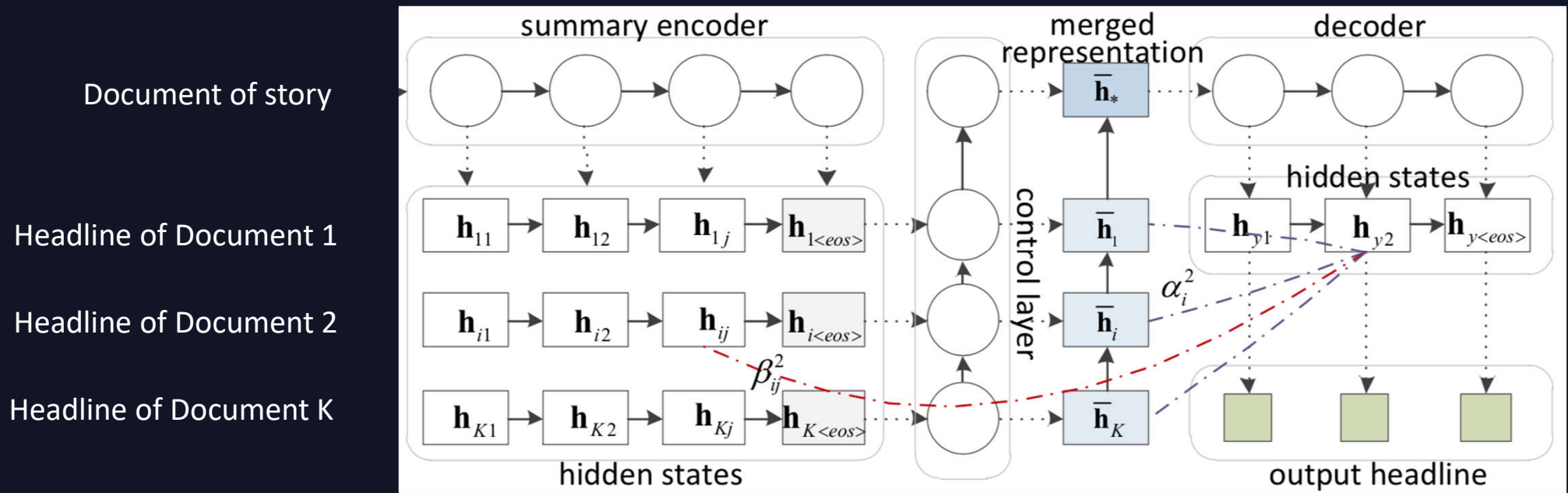
Table 3: Comparing different story structure generation algorithms.

	Tree	Flat	Thread	Timeline	Graph
Correct edges	82.8%	73.7%	66.8%	58.3%	32.9%
Consistent paths	77.4%	—	50.1%	29.9%	—
Best structure	187	88	84	52	19

跟着兴趣走



## Story Captioning – Hierarchical Attention



Rouge-1 : 28.5

$$C_t = f(\{h_{ij}\}) = \sum_{i=1}^K \sum_j \alpha_i^t \beta_{ij}^t h_{ij}$$

$$\alpha_i^t = \frac{\exp(g(\bar{h}_i, h_{yt}))}{\sum_{s=1}^K \exp(g(\bar{h}_s, h_{yt}))}$$

$$\beta_{ij}^t = \frac{\exp(g(h_{ij}, h_{yt}))}{\sum_w \exp(g(h_{iw}, h_{yt}))}$$

跟着兴叔走不丢

## Summarization

- 大规模知识图谱构建
  - 知识图谱简介
  - 知识图谱构建技术
  - 知识图谱目前的使用场景
- 知识图谱在自然语言理解的应用

跟着兴叔走

Thank you

跟着兴叔走平路