大规模知识图谱的构建及应用

钭伟雨
移动浏览产品部
腾讯

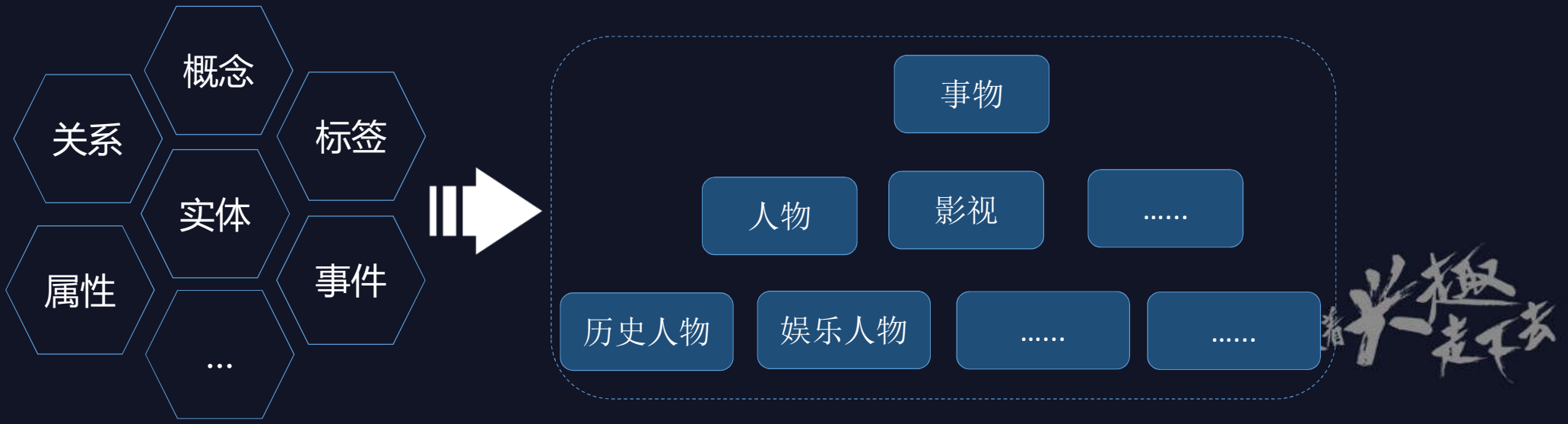
目录

- 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

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

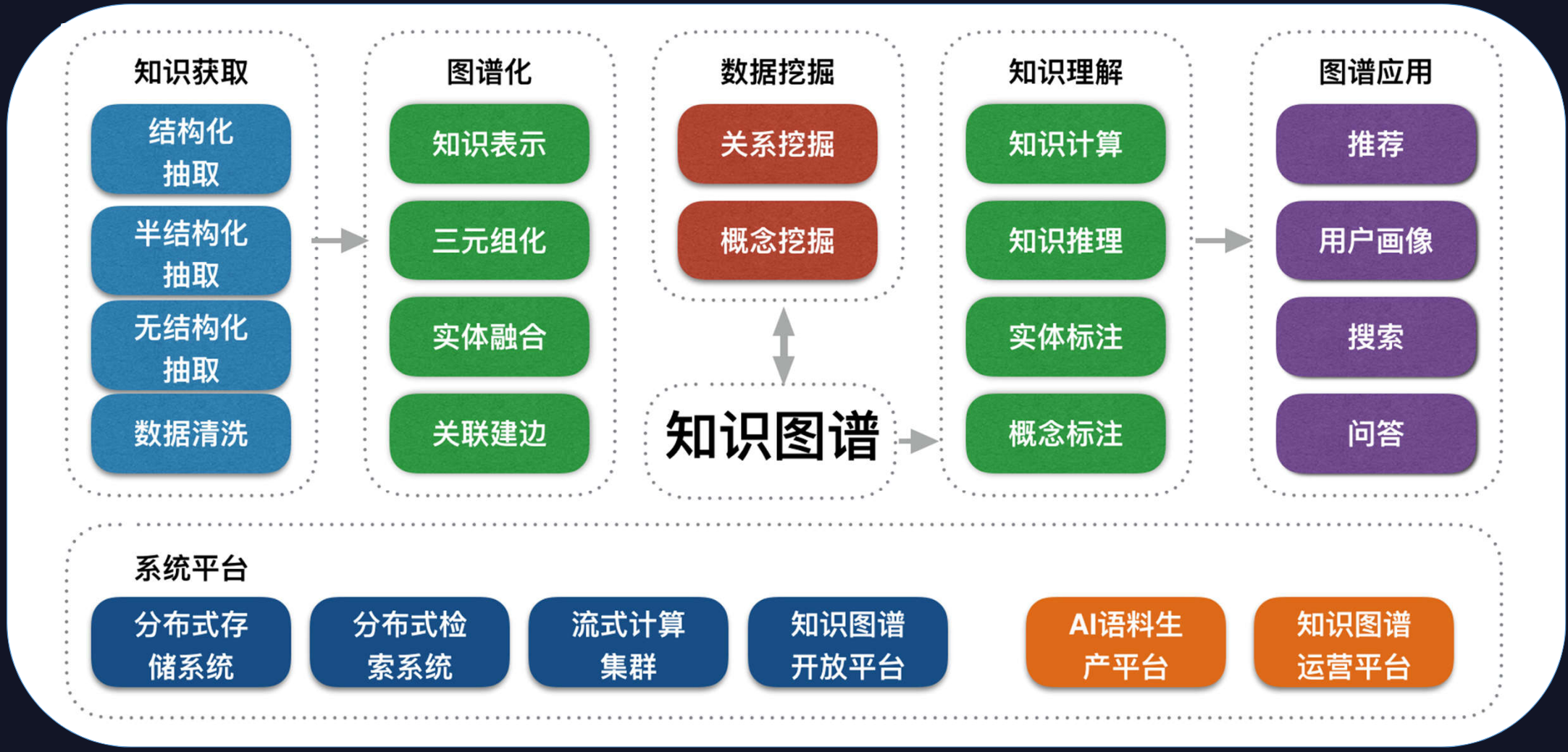
百亿级流水

专业领域

医疗类目
物种类目等

百科类

知识图谱技术架构


知识获取
结构化抽取
半结构化抽取
无结构化抽取
数据清洗
图谱化
知识表示
三元组化
实体融合
关联建边
数据挖掘
关系挖掘
概念挖掘
知识图谱
知识理解
知识计算
知识推理
实体标注
概念标注
图谱应用
推荐
用户画像
搜索
问答
系统平台
分布式存储系统
分布式检索系统
流式计算集群
知识图谱开放平台
AI语料生产平台
知识图谱运营平台

构建技术 - 关系抽取

- 谌龙王适娴结婚
- 女友 - > 妻子

构建技术 - 关系抽取的流程

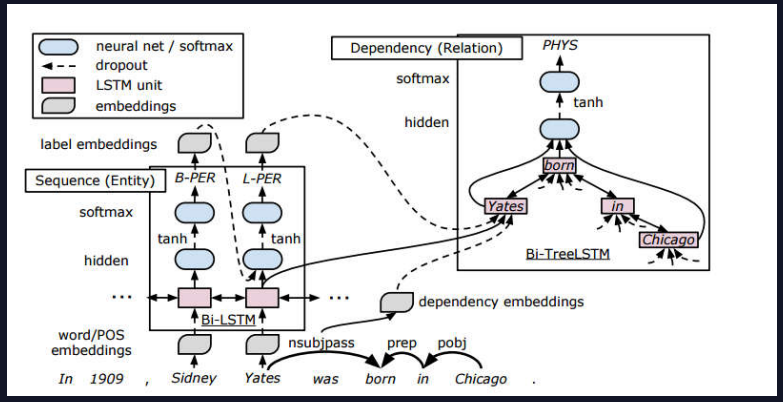
关系定义 → 样本筛选 → 模型建立 → 关系预测

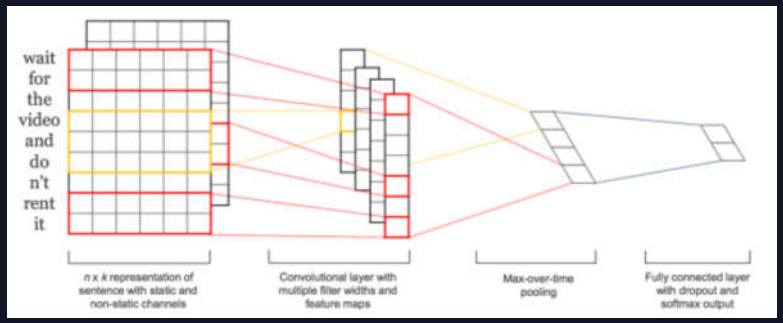
构建技术 - 关系抽取模型

- 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

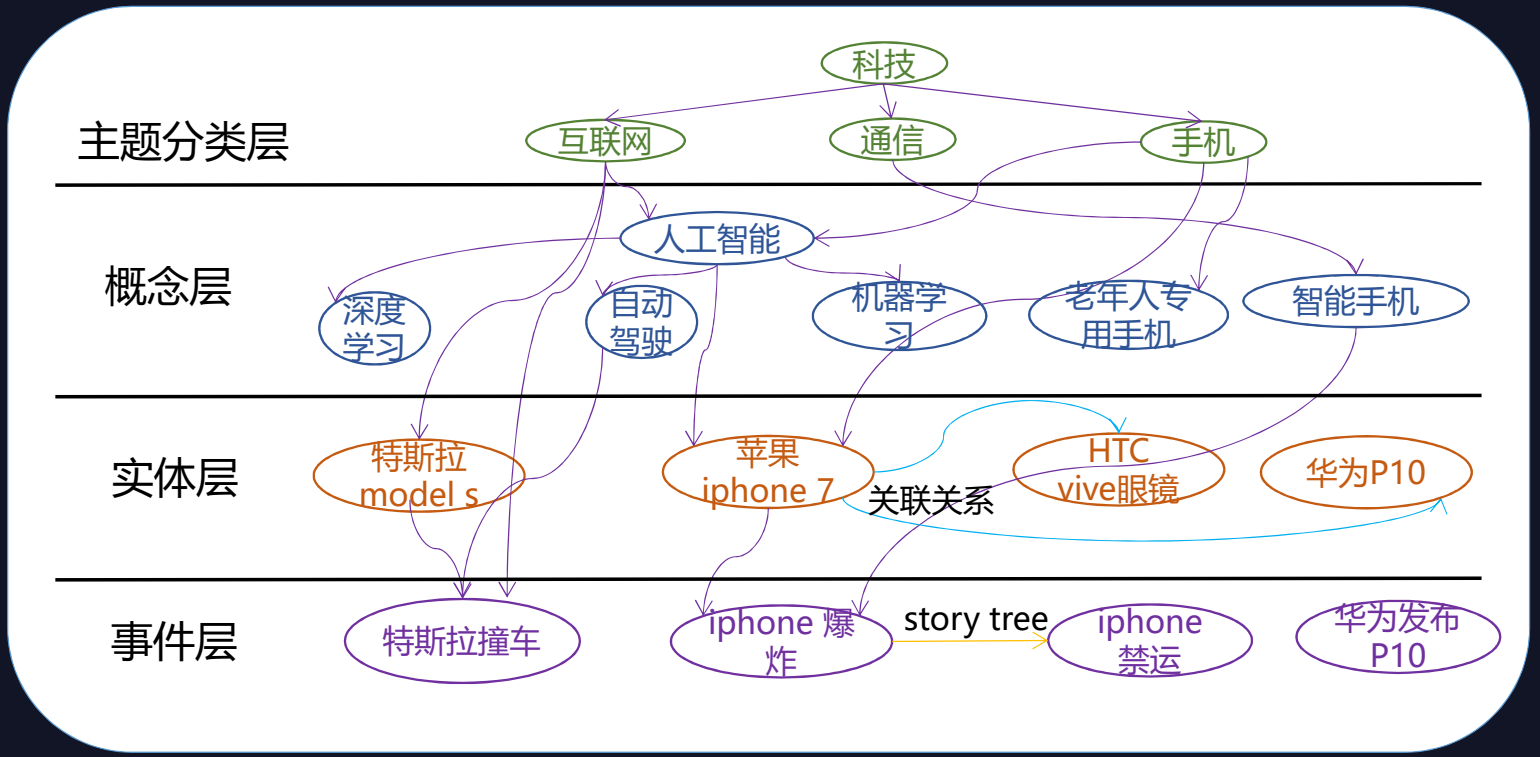
- 基于CNN的关系提取
  - CNN 提取高阶特征：例如句法结构
  - ATT model 代替SPTree 提取关键语义特征。

  - 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%





构建技术 - 概念图谱



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

同义归一

(C,E)Pair挖掘

自动样本标注

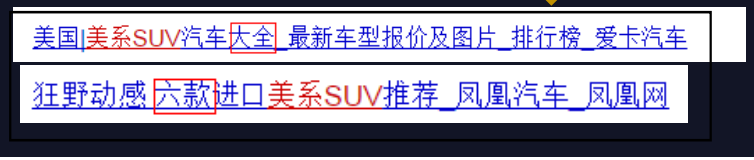
例行化

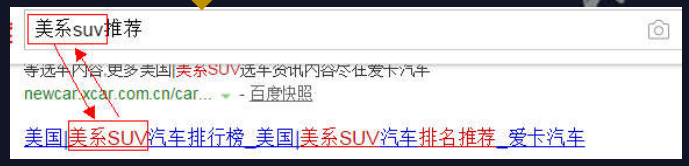
query → query分类 → concept识别 → entity识别 → 同义归一

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

title分类　　点击权重

基于<q,t>对齐语料识别





跟着兴趣走下去

使用场景 - 直接满足用户的需求

跟着兴趣走下去

使用场景 – 更智能的助手服务

跟着兴趣走下去

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

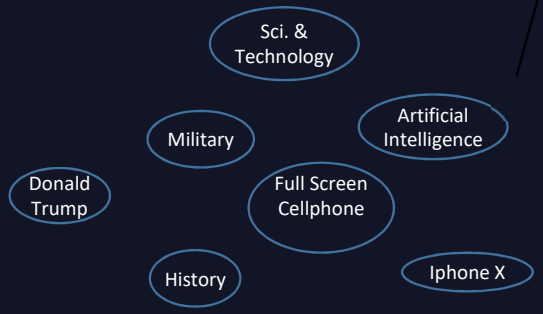
跟着兴趣走下去

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

跟着兴趣走下去

Understand the contents

NLP

Category	汽车	About one thousand，F1:95%
Concept	豪华电动汽车	20K / 140K，F1:95%
Entity	Tesla Model3， Elon Musk	Millions，F1:95.6%
Event	Model3 中国销售	Thousands，F1:96.1% （CIKM）

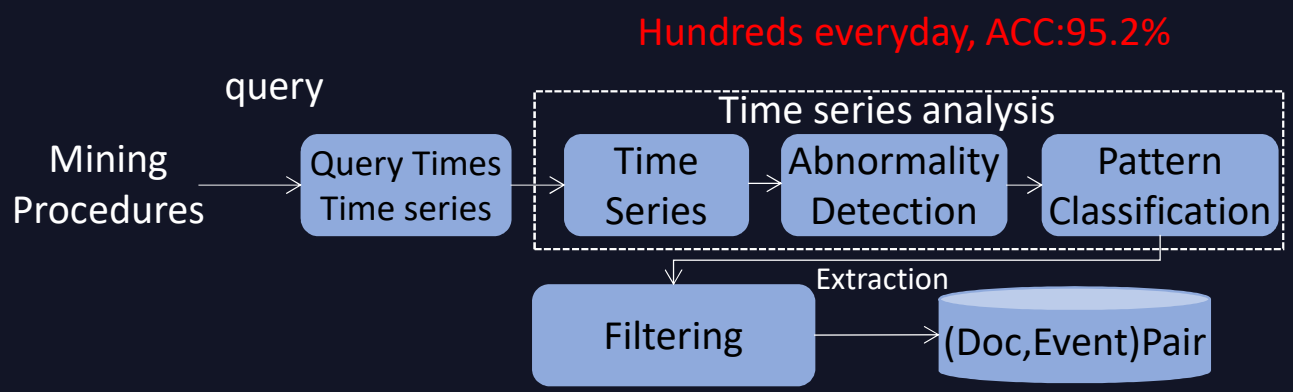
Category： SVM & RCNN
Concept： Three methods towards concept inference
Entity： LSTM+ LP（Google 2015）
Event： CIKM 2017

Creating the event

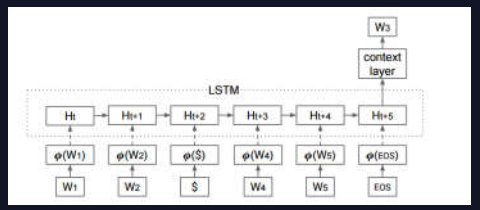
- Line 1: from the crawler
  - High accuracy but low recall
- Line 2: from Query-Doc pair
  - High accuracy but low recall
- Line 3: from document clustering
  - High recall, but accuracy is low

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

About Line 2: Query-Doc event generation



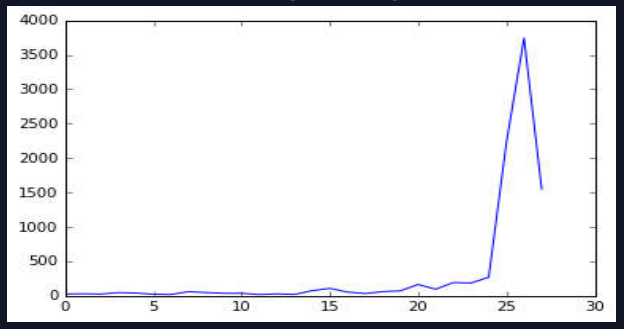
$$Prob(\ news|\ query) = \sum_{t \in titiles} P(news|t)\ P(t|query)$$



Click rate

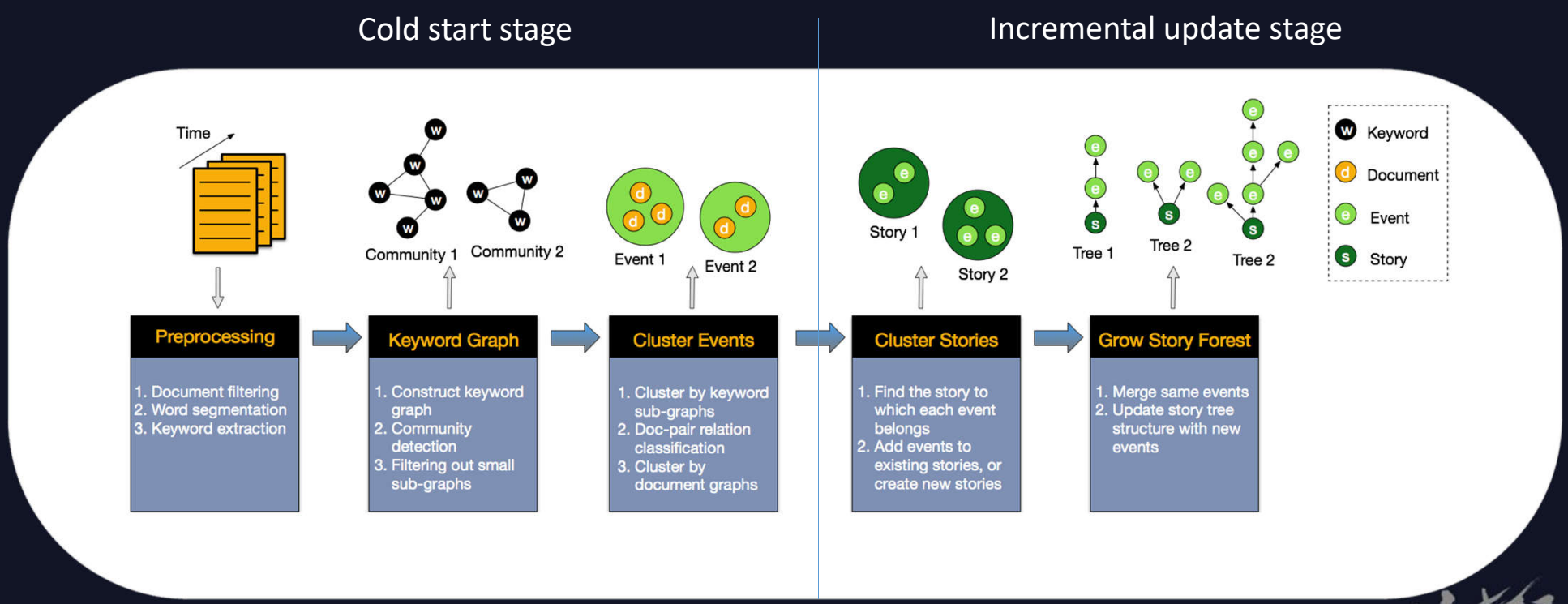
Time Series Patterns

- 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



跟着兴趣走下去

About Line 3: Story Systems



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.

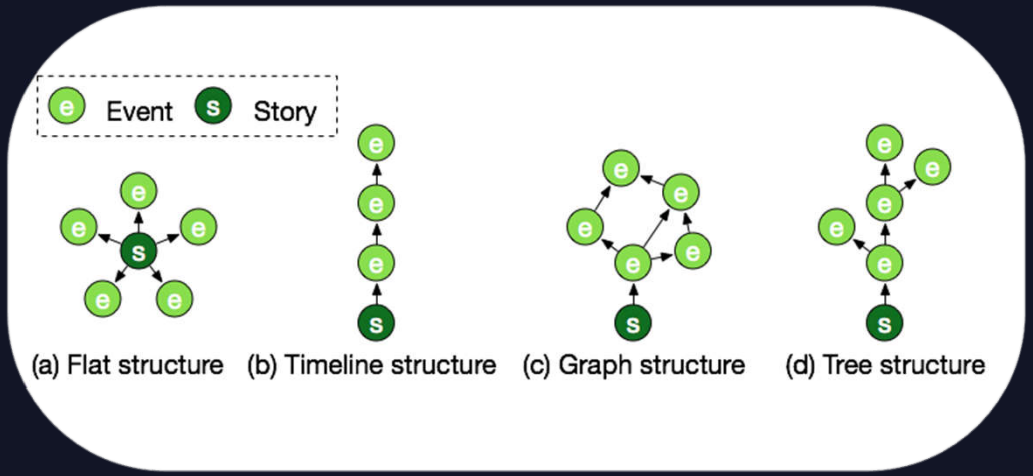


Table 3: Comparing different story structure generation algorithms.

	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

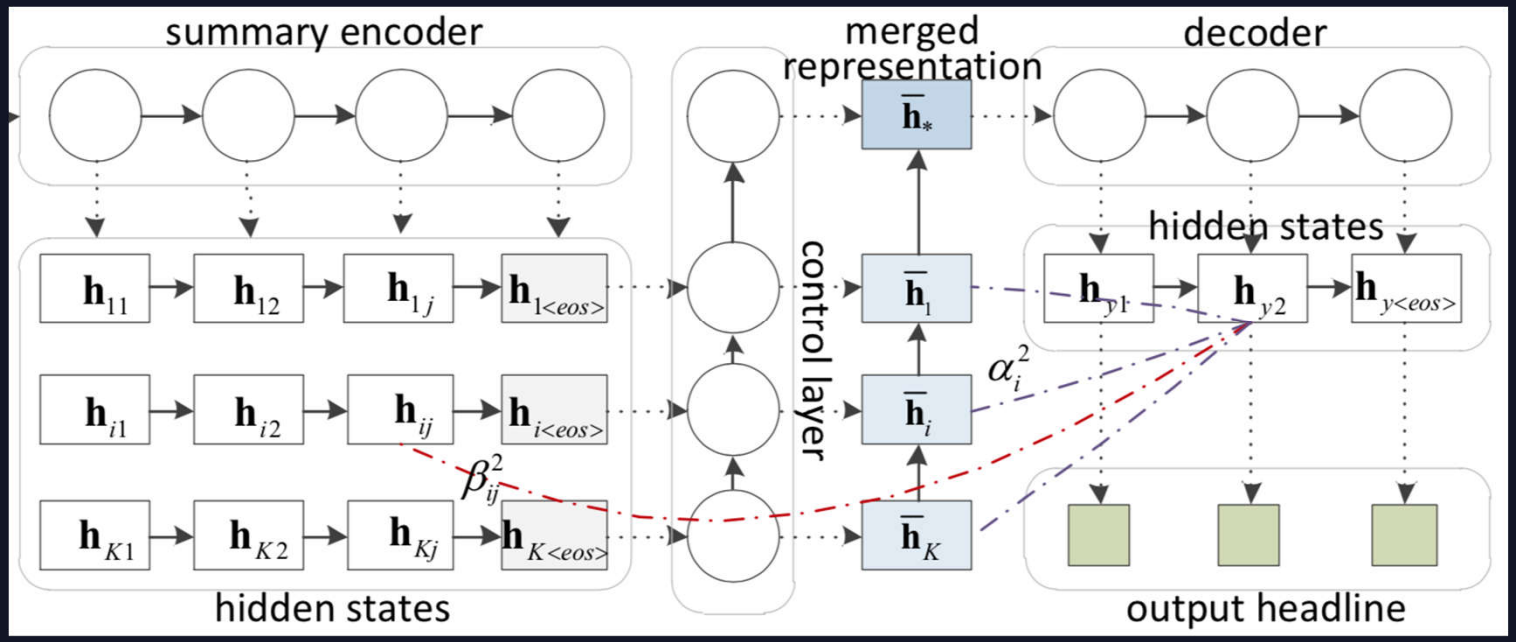
Story Captioning – Hierarchical Attention

Document of story

Headline of Document 1

Headline of Document 2

Headline of Document K



Rouge-1： 28.5

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

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

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

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

Q&A

Thank you