DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

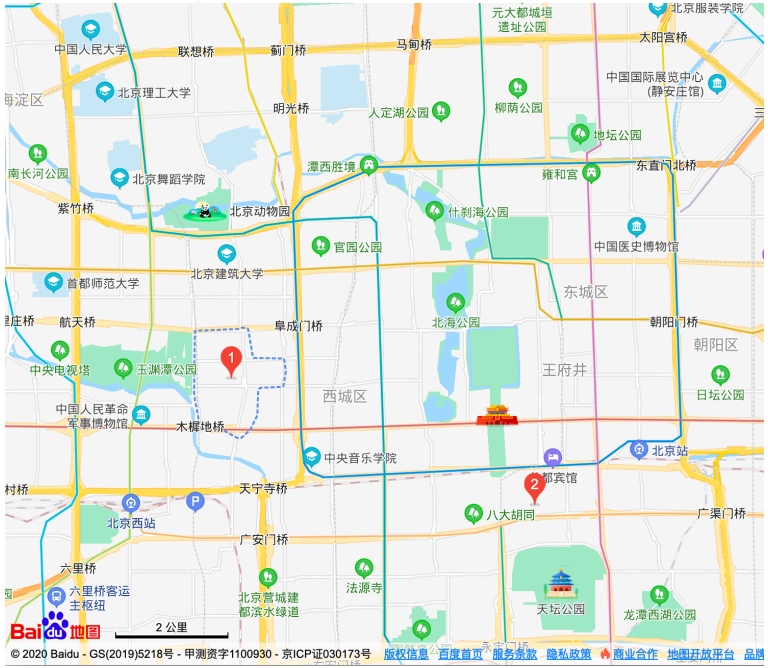
IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

基于京东快递文本识别的自然语言优化之路

京东物流 王梓晨

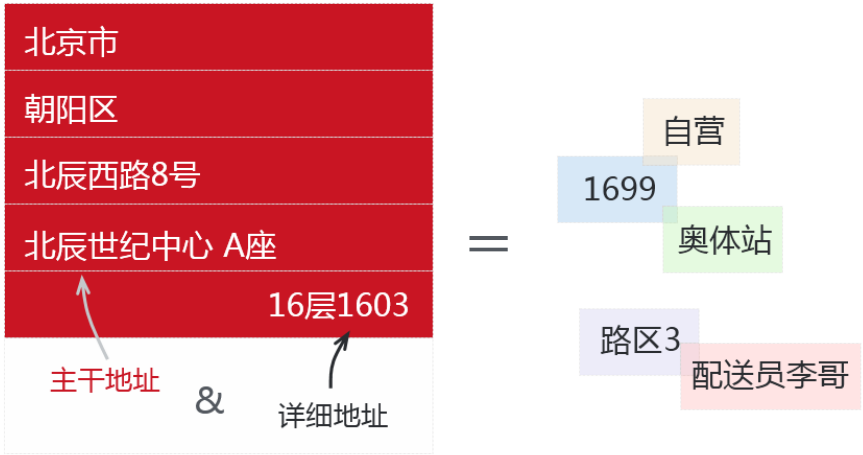
三里河的故事


中国人民大学
联想桥
翁门桥
马甸桥
元大都城垣遗址公园
北京服装学院
太阳宫桥
中国国际展览中心(静安庄馆)
北京理工大学
明光桥
人定湖公园
柳荫公园
地坛公园
南长河公园
北京舞蹈学院
潭西胜境
雍和宫
东直门北桥
紫竹桥
北京动物园
什刹海公园
官园公园
中国医史博物馆
首都师范大学
北京建筑大学
东城区
航天桥
阜成门桥
北海公园
朝阳门桥
朝阳区
中央电视塔
玉渊潭公园
王府井
日坛公园
中国人民革命军事博物馆
木樨地桥
西城区
中央音乐学院
北京站
天宁寺桥
北京西站
八大胡同
广渠门桥
广安门桥
六里桥
六里桥客运主枢纽
北京营城建都滨水绿道
法源寺
天坛公园
龙潭西湖公园
2 公里
© 2020 Baidu - GS(2019)5218号 - 甲测资字1100930 - 京ICP证030173号
版权信息 百度首页 服务条款 隐私政策 商业合作 地图开放平台 品牌


HB京0037

地址有规范吗？

收件姓名是来自这个星球吗？



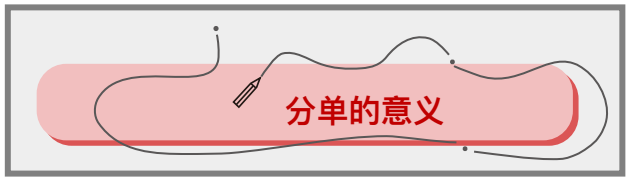
四级地址是什么？
北太平庄桥属于海淀还是朝阳？

业务意义

根据订单收货地址、商品属性、卖家类型等计算末端配送站信息

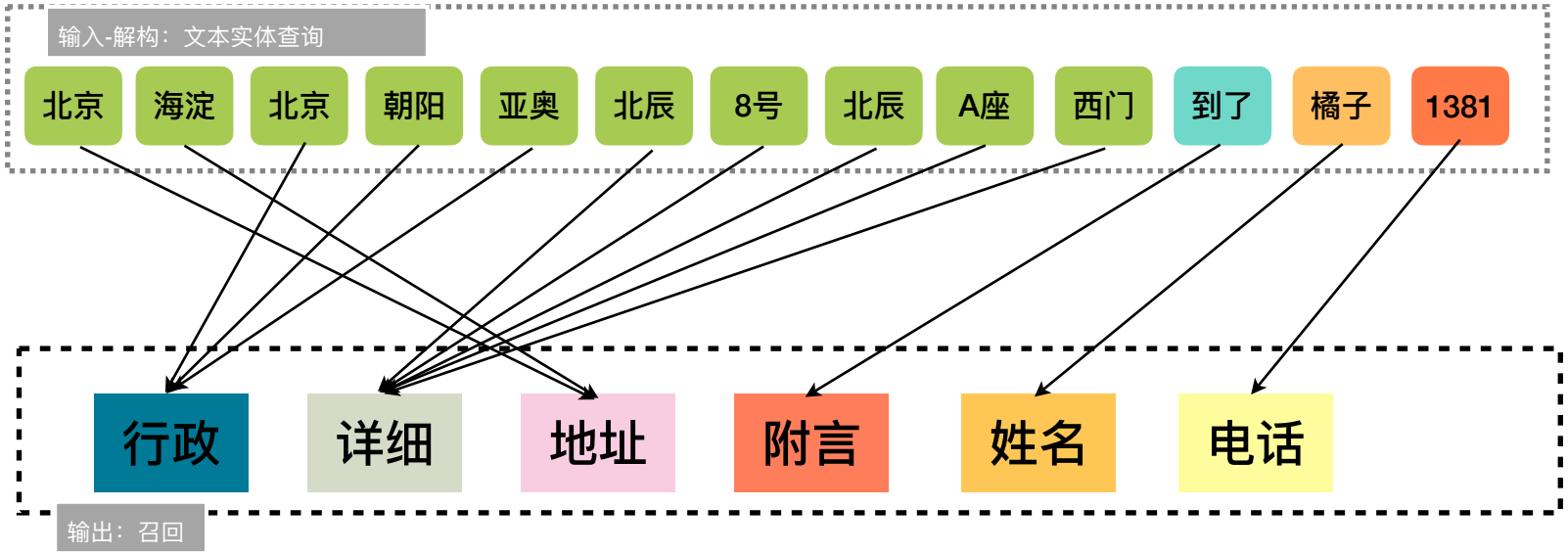
从ofc获得订单信息， 下传订单信息到运单系统， 控制下传速度

为仓库、分拣提供打印时的包裹标签信息：站点，路区，京鲜达等信息



1. 作为订单流程的一部分， 下传订单信息到运单系统
2. 实现分拣中心快速作业
3. 指导站点提前安排配送资源
4. 让KA商家提前知道包裹配送的路由信息

目标：结构化


输入-解构：文本实体查询
北京
海淀
北京
朝阳
亚奥
北辰
8号
北辰
A座
西门
到了
橘子
1381
行政
详细
地址
附言
姓名
电话
输出：召回

怎么做分词?

语料库
地址库

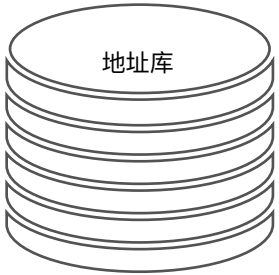
到货请联系我前电话
先电话联系
送前电话联系以便留人签收电话联系

姓名库

吴**
汪**

电话库

1392300****
1527015***



北京大兴亦庄朝林广场A座
小熊收18661861888

分词模型

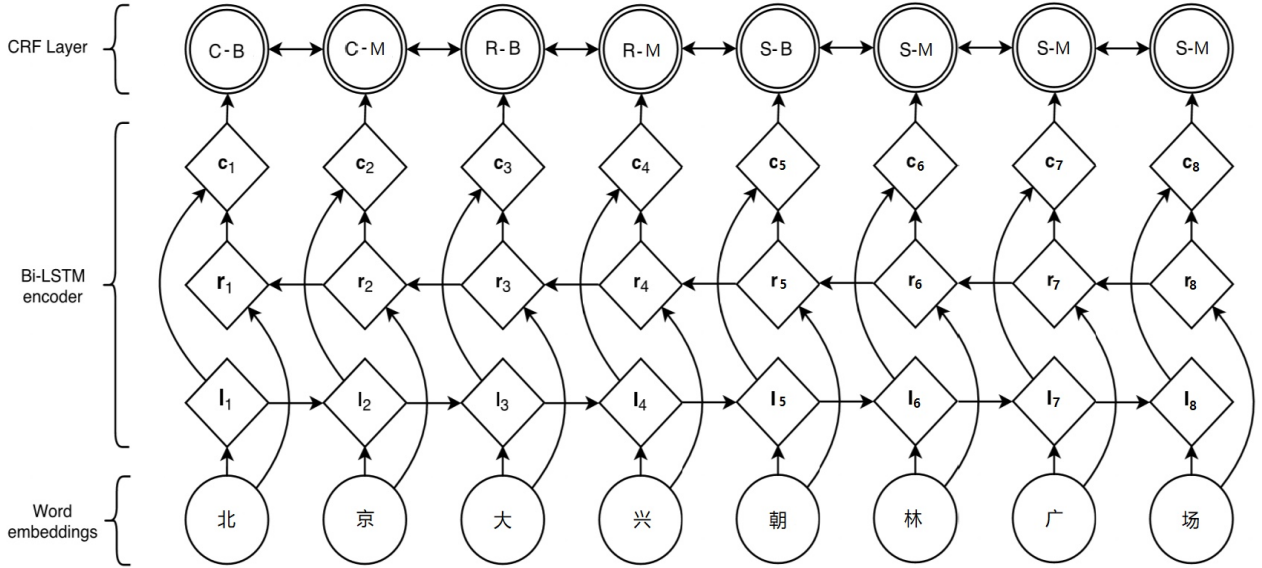
label名称	词语
城市	北京
区县	大兴
poi	朝林大厦
人名	张三
电话	**18661861888**
附言	**到了电话**

地址分词

词典 VS 神经网络

嘉铭园-二区	c13332	下一级
武警一师家属楼	c13514	下一级
飘亮阳光广场	c13728	下一级
北京市朝阳区外国语小学二部	c13729	下一级
名人 大厦	c13731	下一级
慧忠北里（扩建）	c14070	下一级

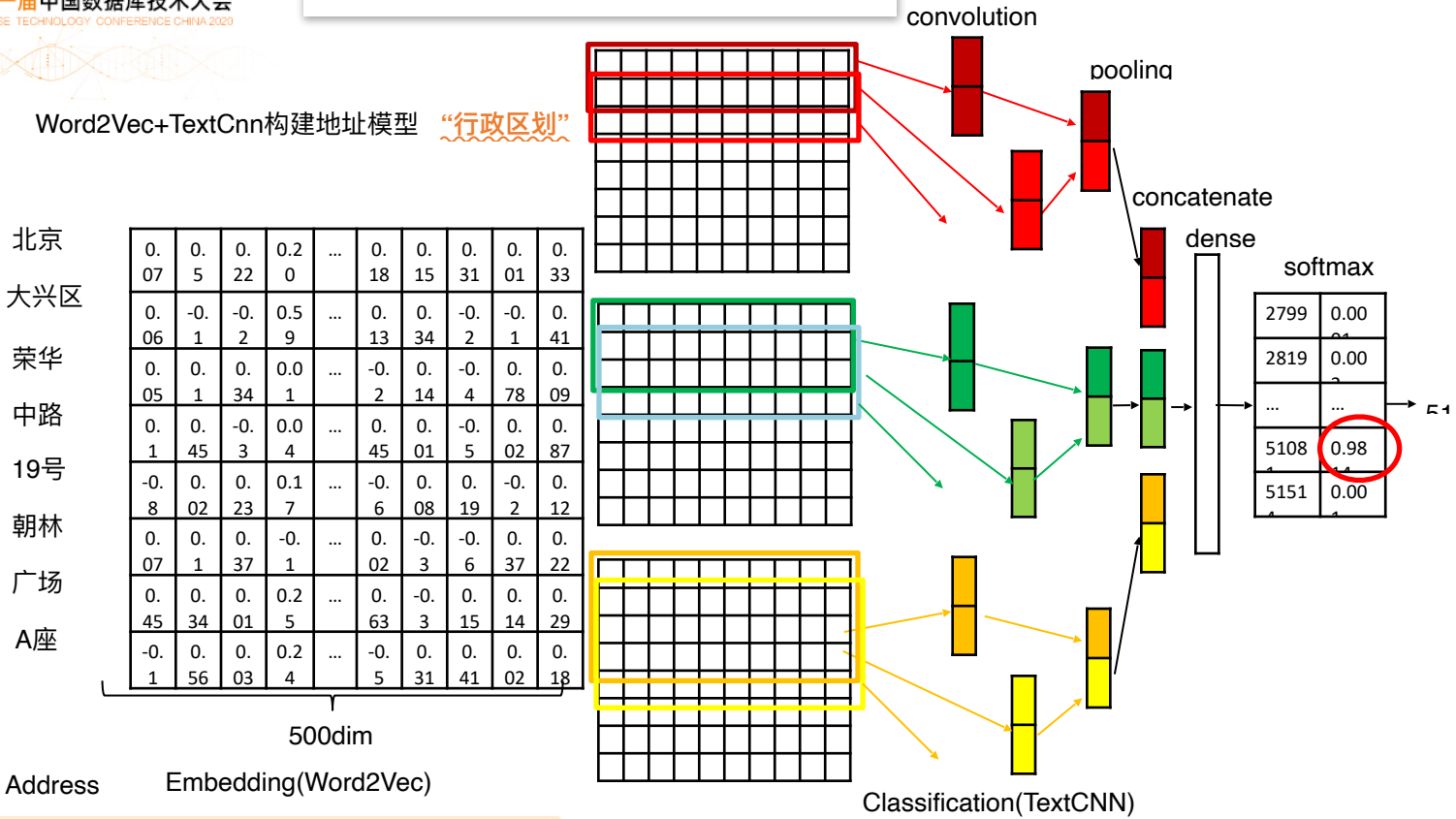
文本解析网络结构



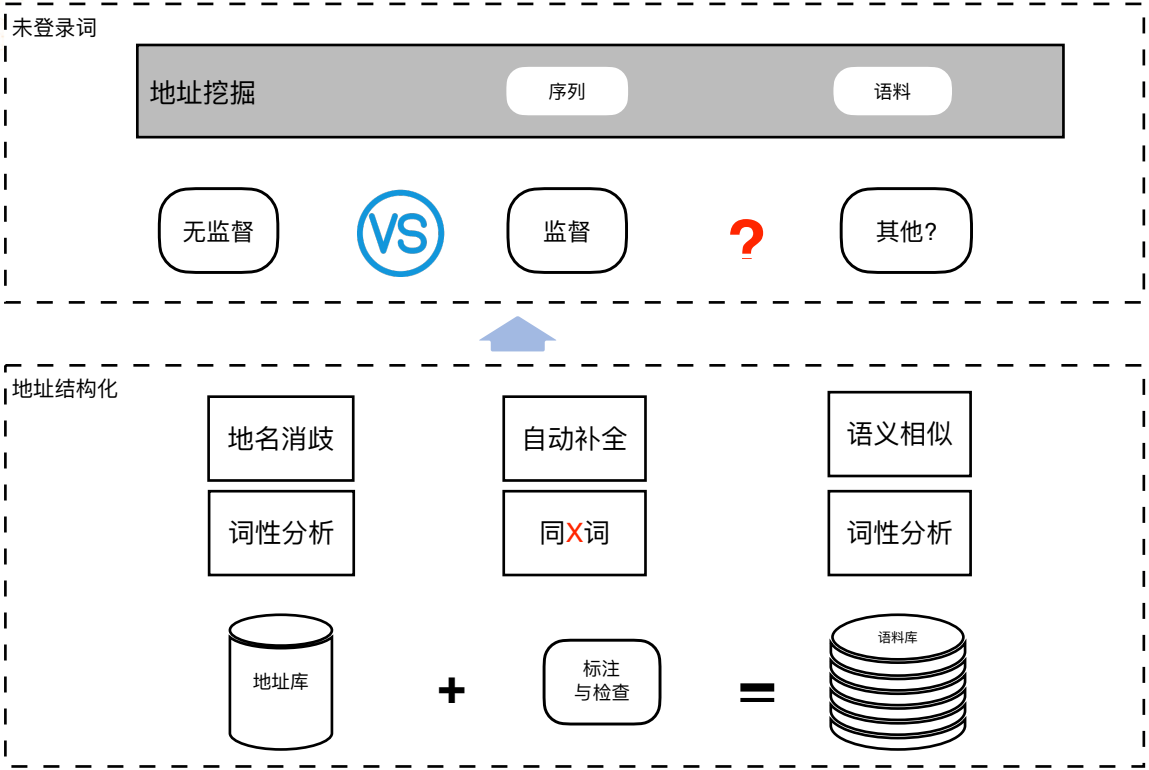
+词库 … 遗留问题?

语义相似度

Word2Vec+TextCnn构建地址模型 “行政区划”



结构化基础数据建设架构



未登录词-编辑距离

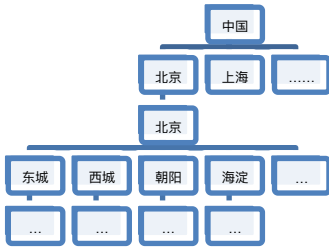
北京市朝阳区中关村街道中关村南一条7号

行政区划
北京市
朝阳区
中关村街道
中关村

地址树路径
北京市：*[0,0]*
朝阳区*: [0,0,2],[5,12]*
中关村街道*: [0,0,3,10]*
中关村*: [0,0,3,10]*

对候选地址进行惩罚值计算
若词语不落在候选路径上，对路径累加惩罚值，惩罚值设为该词不为行政区划词的概率的对数值。

地址树



输出
除噪结果

行政区划
北京
北京市
海淀区
中关村街道

惩罚值最小的路径
[0,0,3,10]

反馈
提供概率值

存储

别名
北京市：北京
中关村街道：中关村街道
中关村：中关村街道

输出
别名库

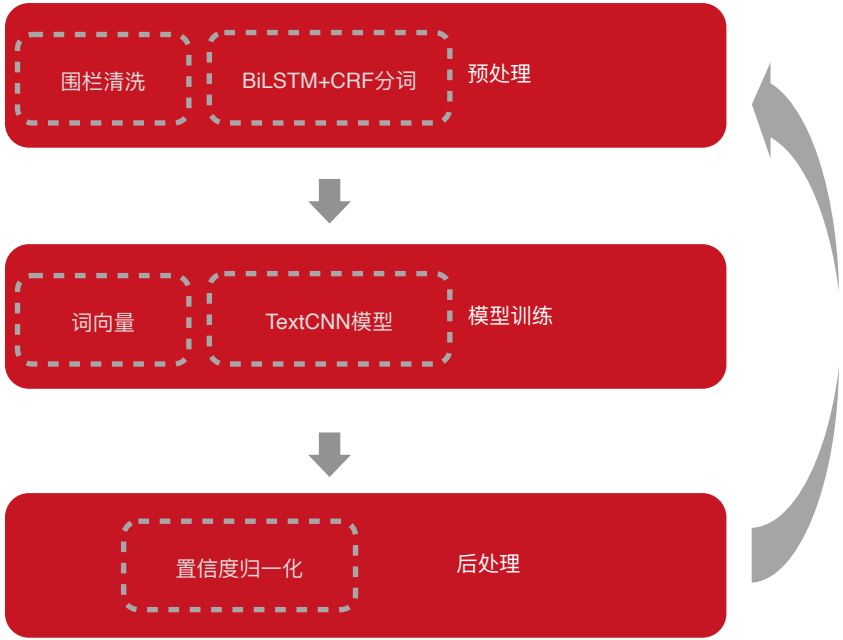
行政区划	别名库
亦庄镇	{'亦庄': 723586, '亦庄镇': 29369...}
双流区	{'双流区': 1189560, '双流县': ...}
成都	{'成都市': 10792101, '成都': 523446}
.....	.....

相似度匹配

输出相似度匹配

	建模	前处理	检索过程	后处理
	倒排表	重复率干扰	❶ term匹配 — 优点 / 缺点	打分机制
	切词|不切词	召回限制	tf/idf — 用户书写习惯不一致 / 大词？小词？ / 查找倒排索引表 topN	

相似度匹配



相似度匹配-打分

$$Score(Q,d) = \sum_{i}^{N} W_i * R(q_i,d)$$

W 每个词在全文档的权重值
$R(q_i,d)$ 每个查询词在当前文档中的得分

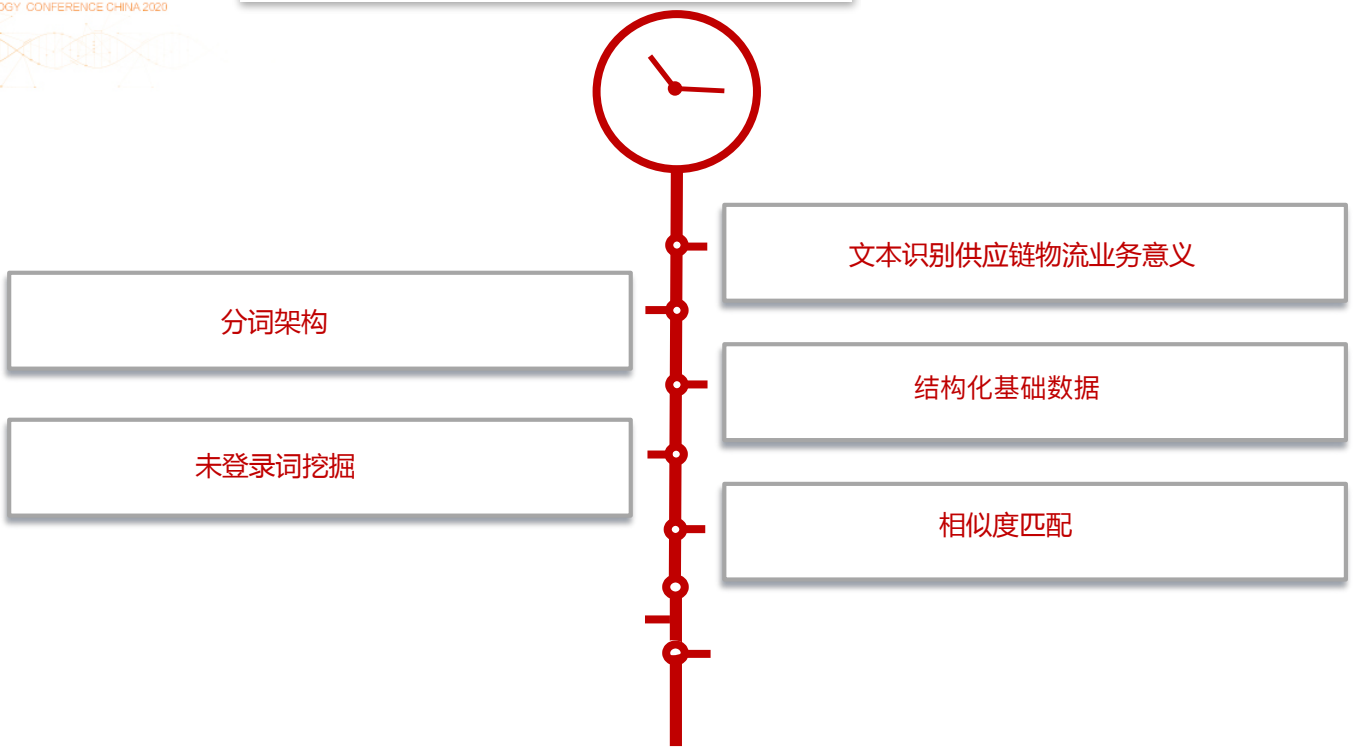
jaccard index(A∩B/A∪B)

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

线性回归并以及 sigmoid 归一化来构建评估搜索返回结果**置信度**的模型

$h_\theta(x) = \theta_0 + \theta_1 x\_tfidfscore + \theta_2 x\_siteoccupation + \theta_3 x\_jaccard$

集腋成裘 · 聚沙成塔



THANKS