面向企业数据孤岛的联邦排序学习

史鼎元，王晏晟，郑鹏飞，童咏昕

软件开发环境国家重点实验室
大数据科学与脑机智能高精尖创新中心
北京航空航天大学

北京航空航天大學
BEIHANG UNIVERSITY

内容提要

- 研究背景
- 主要挑战
- 解决方案
- 实验验证

内容提要

- 研究背景
- 主要挑战
- 解决方案
- 实验验证

人工智能与数据孤岛

金融风控

数据分散于小型银行

用户行为

数据分散于用户设备

直接合并数据开展训练？

智慧交通

数据分散于出行平台

智能政务

数据分散于政府部门

激增的数据需求使数据孤岛问题凸显

隐私保护立法

愈发严格的隐私保护要求

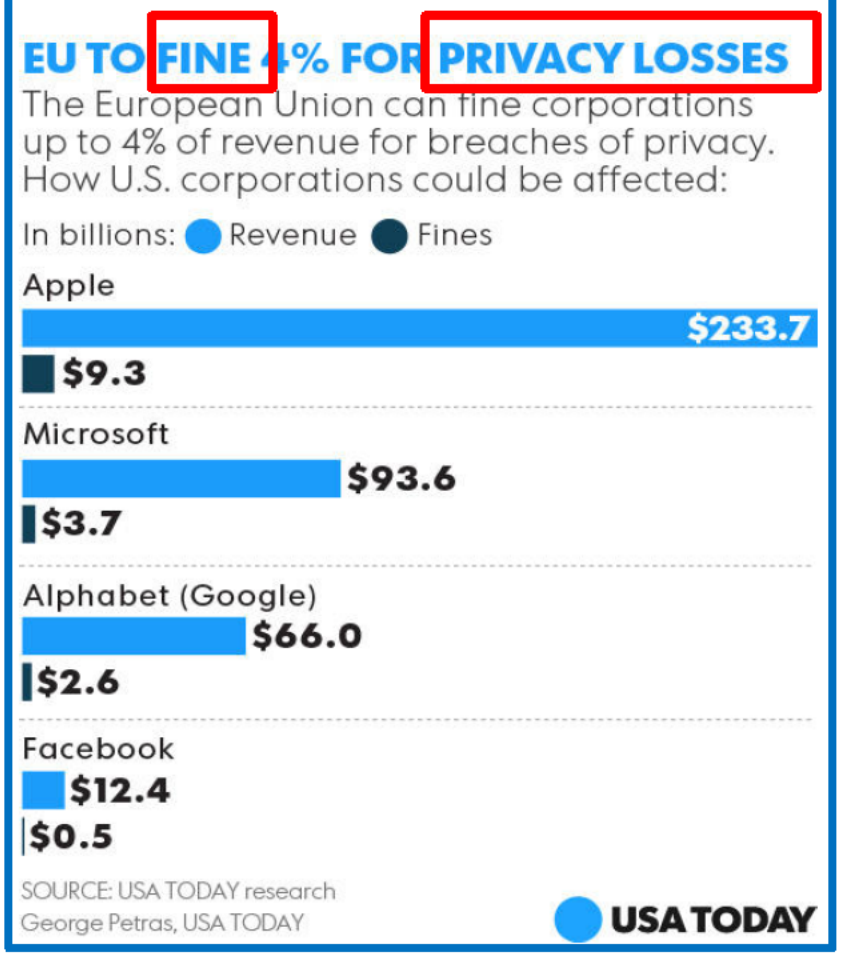
2012 《消费者隐私权利法案》

2017 《网络安全法》

2018 《通用数据保护条例》

限制用户数据随意传输使用

条例出台后多家大企业成为被告

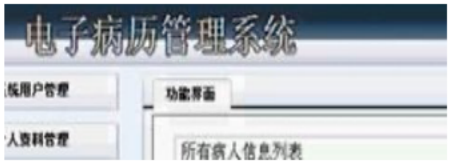


隐私保护立法限制企业间共享用户数据，加剧数据孤岛问题

小规模数据拥有者的检索难题

小规模数据拥有者

中小型医院



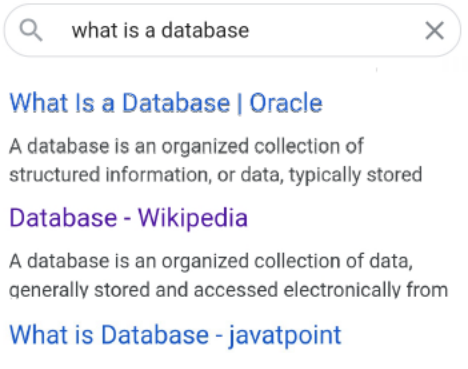
病历检索

中小型机构

情报检索

如何获得
优质检索服务

排序学习模型



评估文档和查询的
相关性并排序

训练

Google

Bai du 百度

Bing

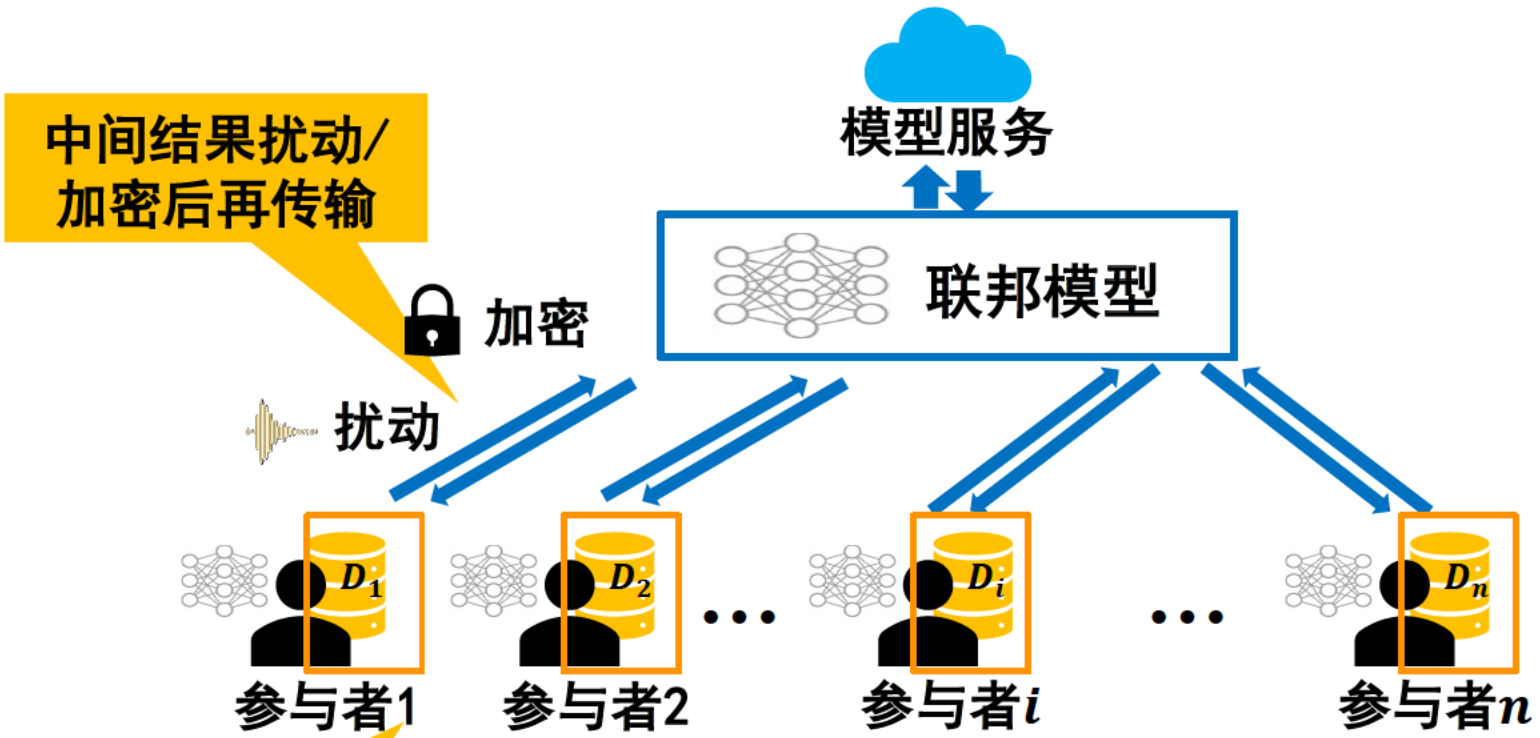
海量查询语句

海量网页文档

隐私限制阻碍数据共享

隐私限制阻碍数据共享与模型训练，无法获得优质检索服务

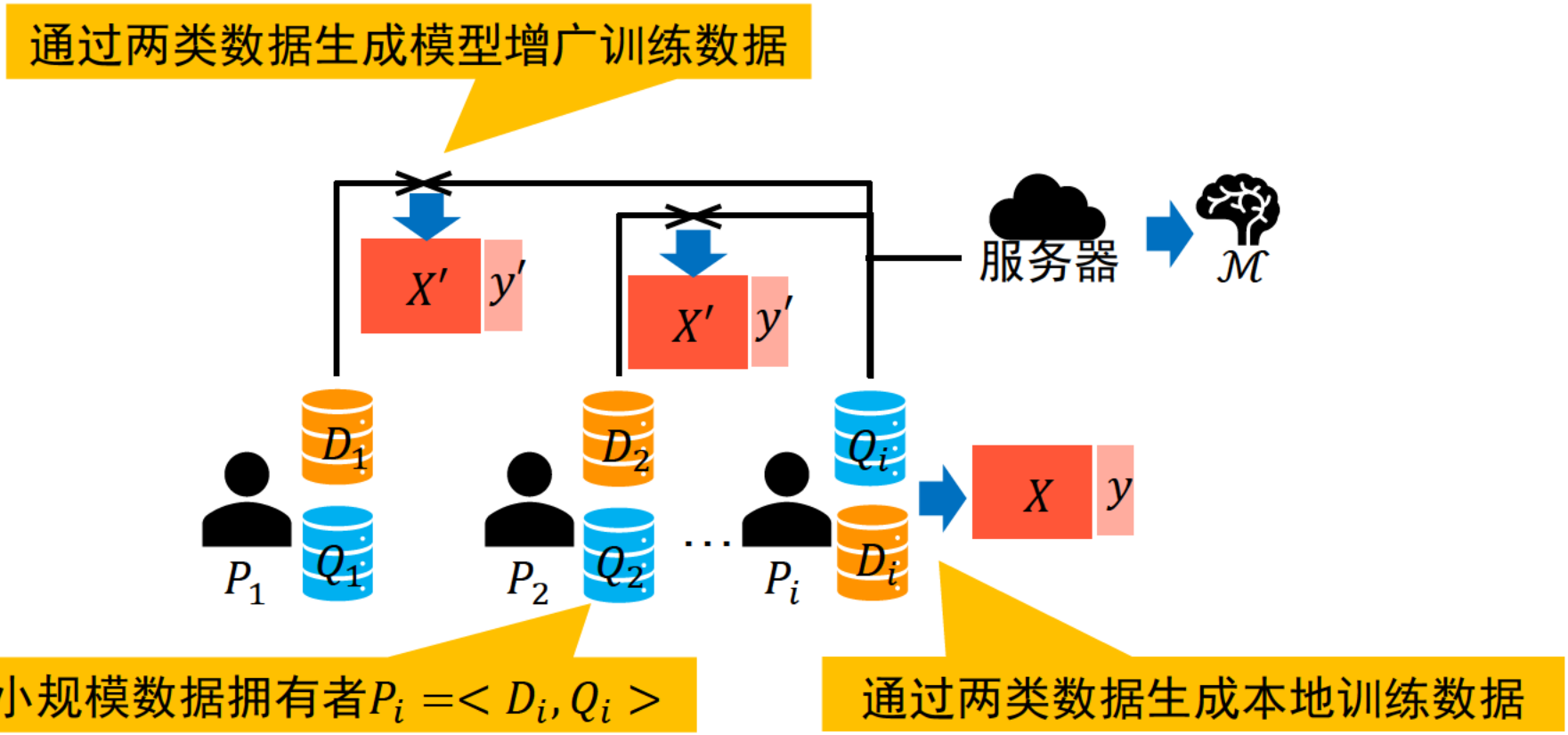
联邦学习：数据孤岛破局方法



联邦学习让小规模数据拥有者能安全训练高效排序学习模型

面向企业数据孤岛的联邦排序学习

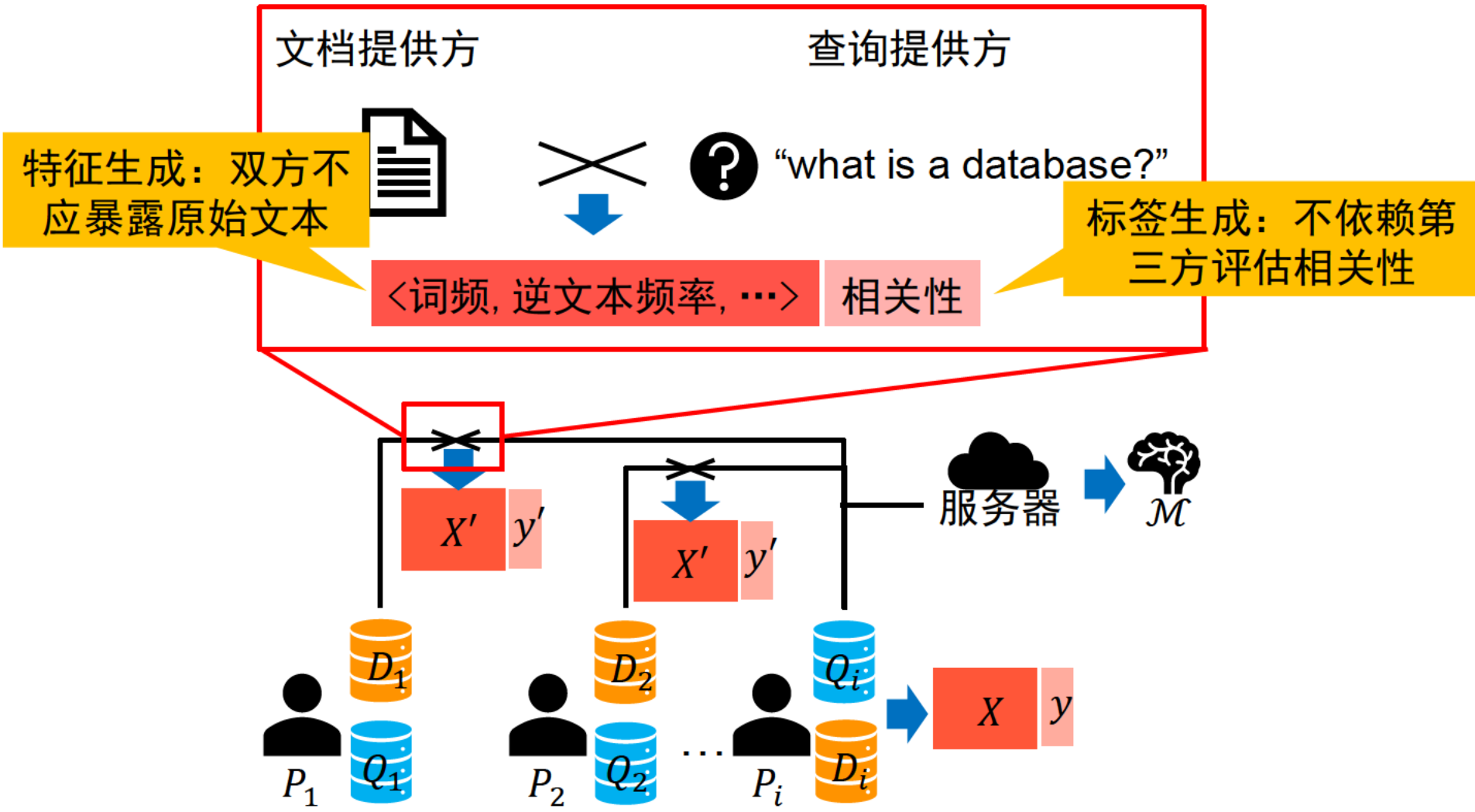
- 设计要求
  - 各方隐私能够保护
- 设计目标
  - 协作训练高效排序学习模型



内容提要

- 研究背景
- 主要挑战
- 解决方案
- 实验验证

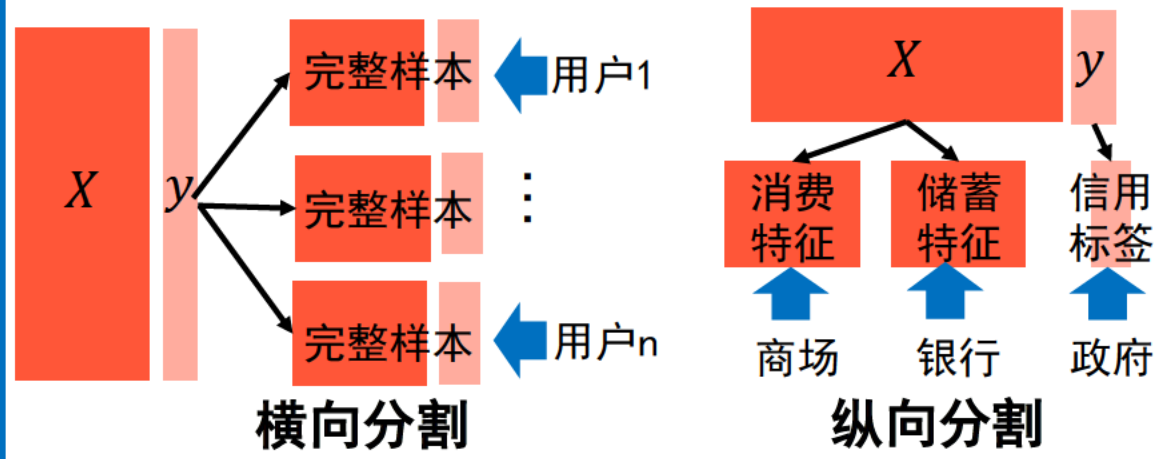
挑战：数据增广难度大



隐私保护要求加大了特征和标签的生成难度

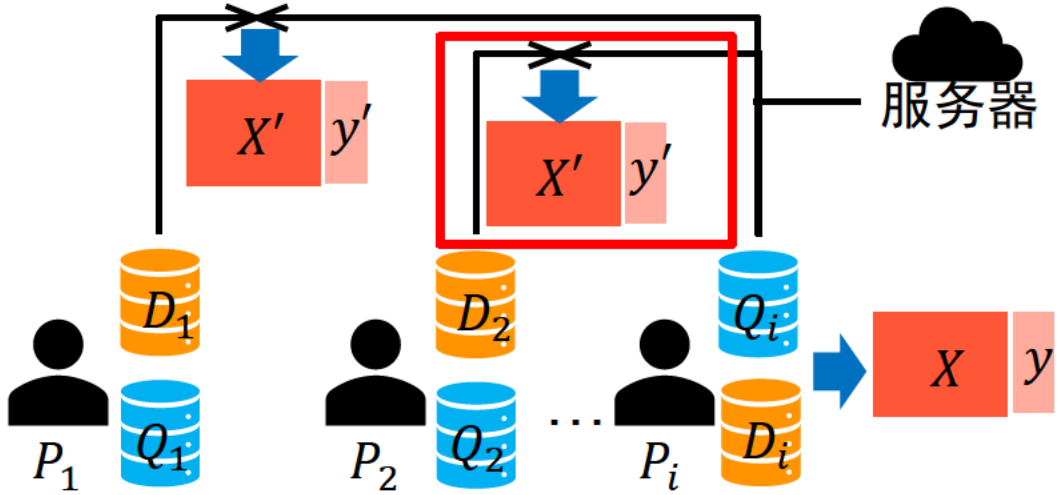
挑战：数据增广难度大

常规数据分割形式



特征/标签仅需一方数据生成

联邦排序学习分割



特征/标签需要两方数据生成

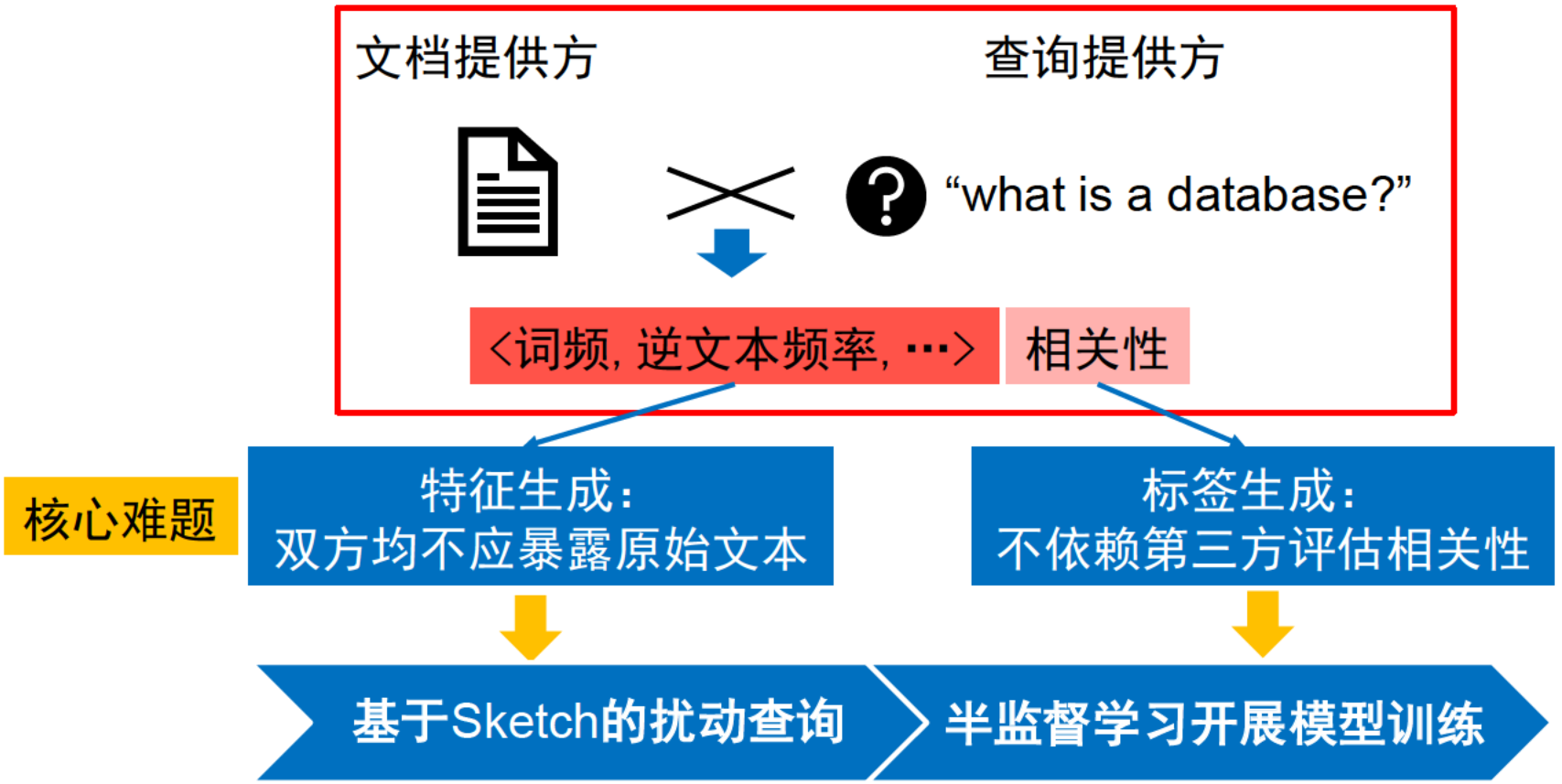
联邦排序学习引出了新的数据分割方式：交叉分割

内容提要

- 研究背景
- 主要挑战
- 解决方案
- 实验验证

解决方案框架

- 数据增广难度大



数据库领域技术解决联邦排序学习的核心难题

解决方案流程

基于Sketch的扰动查询 → 半监督学习开展模型训练

- 例：词频合作生成
  - $TF_{i,j}(t_k, d) = \frac{TC_{i,j}(t_k,d)}{l_d}$

出现次数：隐私信息

文档长度：公开信息

解决方案流程

基于Sketch的扰动查询 | 半监督学习开展模型训练

文档提供方 | 查询提供方

文档单词

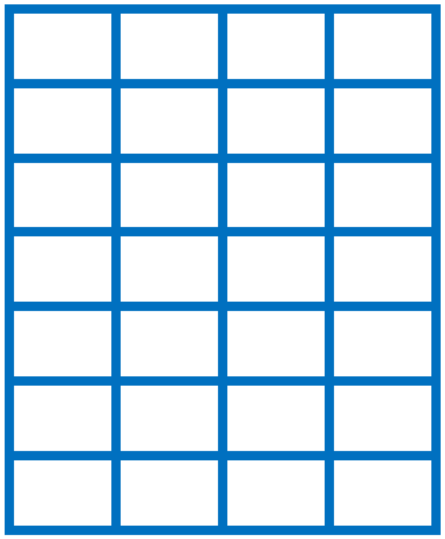
文档

t_1	t_2	...

查询

t_2

i	$h_i(t_2)$
1	2
2	4
3	3
4	2
5	3
6	2
7	1



Sketch每一行对应一个哈希函数

解决方案流程

基于Sketch的扰动查询　半监督学习开展模型训练

文档提供方　查询提供方

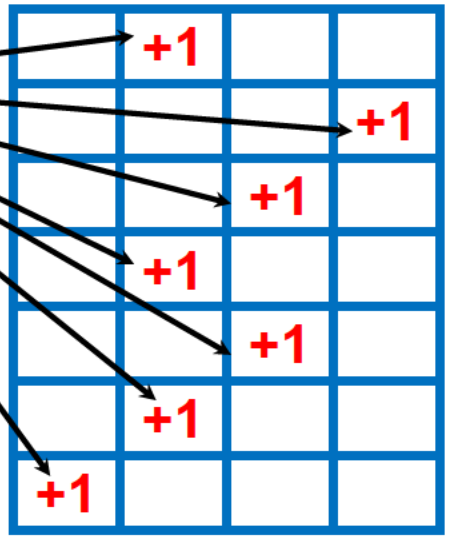
文档

t_1	t_2	...

查询

t_2

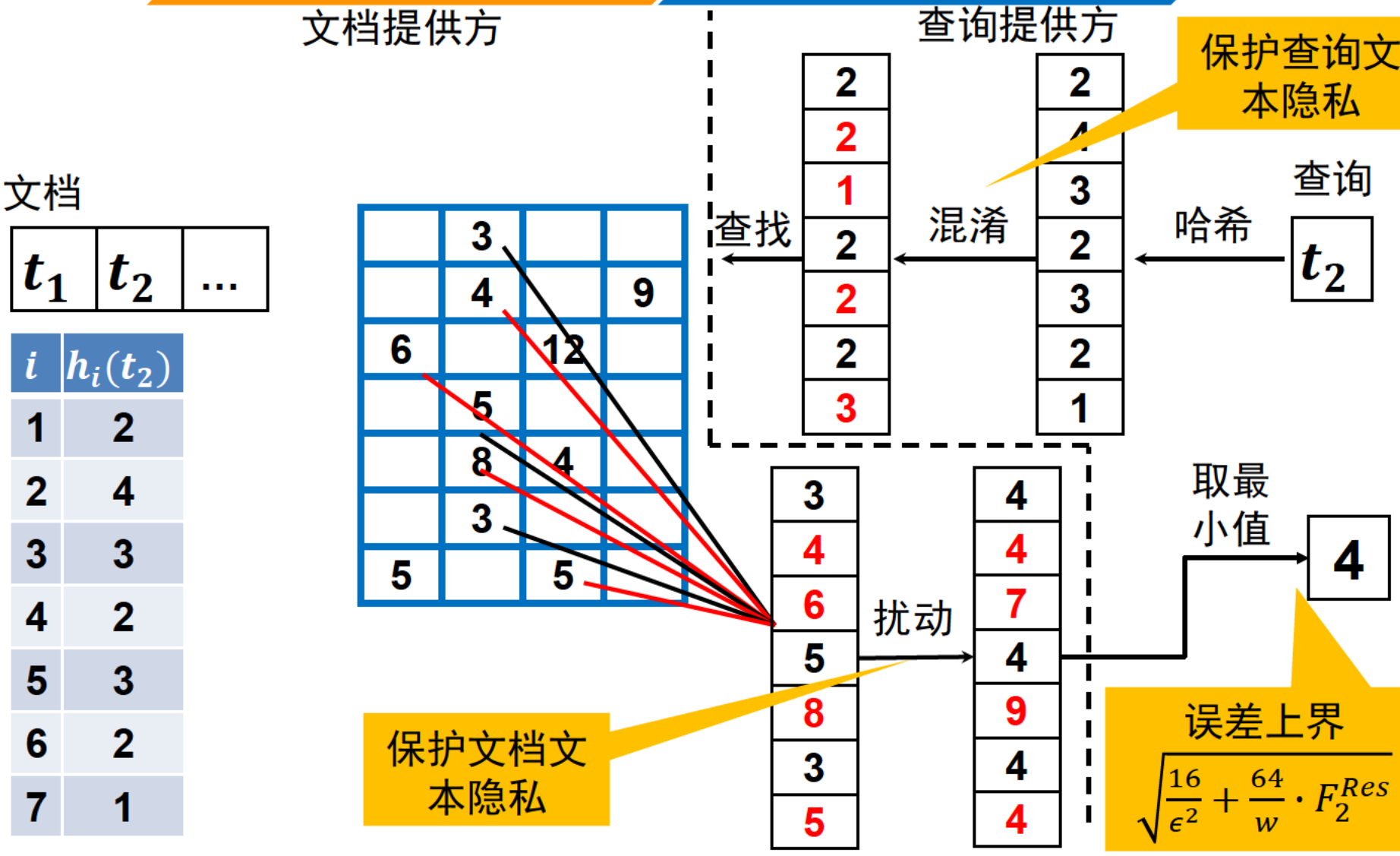
i	$h_i(t_2)$
1	2
2	4
3	3
4	2
5	3
6	2
7	1



对每一行，选一格+1

解决方案流程

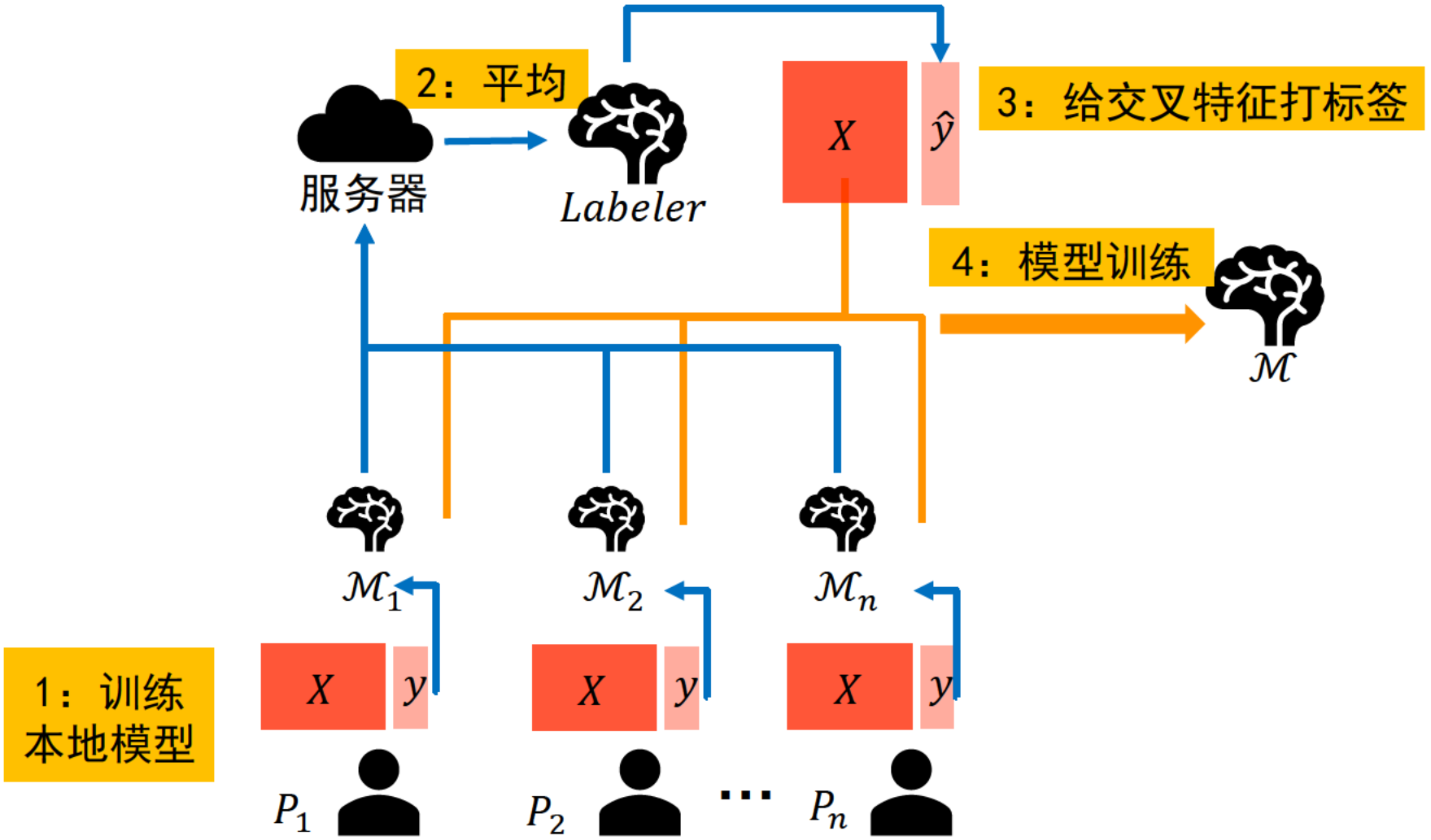
基于Sketch的扰动查询　半监督学习开展模型训练



i	$h_i(t_2)$
1	2
2	4
3	3
4	2
5	3
6	2
7	1

解决方案流程

基于Sketch的扰动查询 半监督学习开展模型训练



内容提要

- 研究背景
- 主要挑战
- 解决方案
- 实验验证

数据集与数据规模

- 数据集
  - MS MARCO文本数据集

- 数据规模
  - 构造4个企业作为联邦参与方
  - 各企业数据规模：
    - 200条查询语句
    - 40000篇文档
    - 带标签特征数据2.4万条
    - 无标签特征数据5.97万条
  - 测试集规模：2.8万条

考核指标与对比方法

指标名称	含义
期望倒数排名(ERR)	相关度高的文档排序位置是否靠前
平均准确率(MAP)	排序学习模型给出的顺序与按相关度降序排序的差别
归一化折损累计增益(nDCG)	综合考虑相关性和排序位置的评估指标
前十名归一化折损累计增益(nDCG@10)	排在前十名文档的归一化折损累计增益

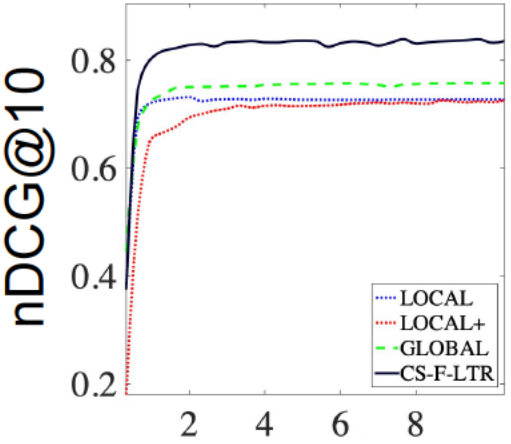
方法名称	训练使用的数据
Local	本地数据
Local+	本地数据+基于自身查询语句的增广数据
Global	所有本地数据
CS-F-LTR	所有本地数据+所有增广数据

有效性检验

		ERR	nDCG@10	nDCG	MAP
Local	企业A	0.5672	0.7274	0.8049	0.4809
	企业B	0.5357	0.6676	0.7779	0.4573
	企业C	0.5637	0.7094	0.7952	0.4788
	企业D	0.5887	0.7466	0.8173	0.5055
Local+	企业A	0.5628	0.7257	0.8026	0.4776
	企业B	0.5448	0.7064	0.7932	0.4779
	企业C	0.5766	0.7314	0.8091	0.4936
	企业D	0.5988	0.7475	0.8247	0.5039
Global		0.5877	0.7577	0.8247	0.5040
CS-F-LTR		**0.6679**	**0.8358**	**0.8760**	**0.5525**

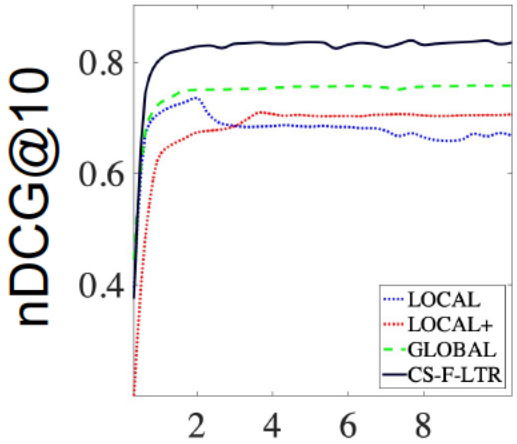
所提联邦学习方法训练的模型性能优于其他方法

迭代过程检验



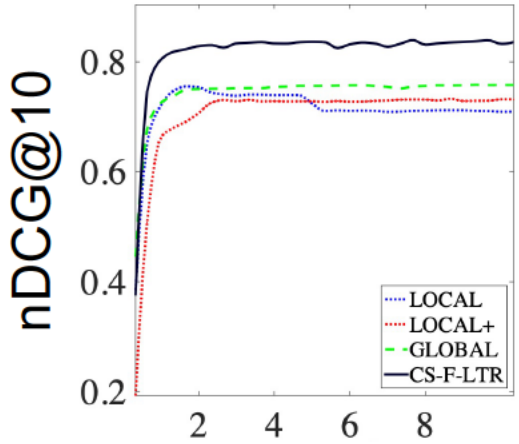
所提方法

企业A　迭代轮次



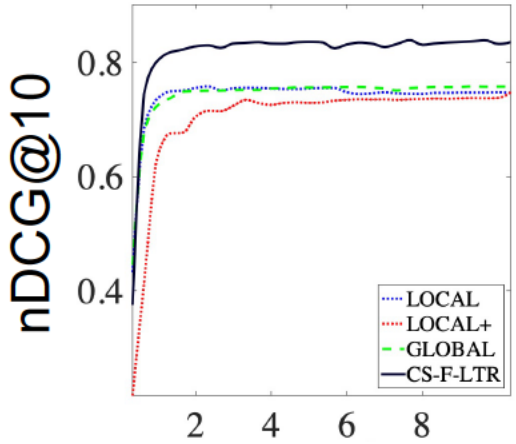
所提方法

企业B　迭代轮次



所提方法

企业C　迭代轮次



所提方法

企业D　迭代轮次

从收敛过验证所提方法的最优性

贡献总结

- 提出联邦场景下的排序学习问题，明确其不同于常规横/纵向分割的交叉分割方式
- 提出基于Sketch的扰动查询方法和半监督学习方法，增广样本数据训练排序学习模型
- 通过在真实文本数据的实验证明算法的有效性

Q & A

Thank You