从文本分类到分类系统

Text Classification and Classification System

报告人：叉烧

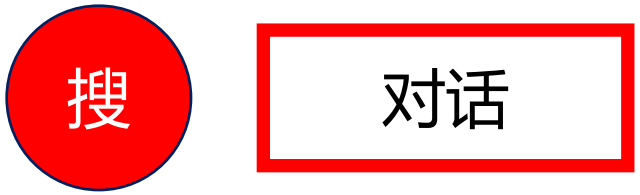
日期：2021年9月11日

目录
CONTENTS

PART ONE

背景

落地场景



细分意图

下游处理

Query：七里香——music/album
Query：周杰伦——Singer
Query：爱情来的太快就像龙卷风——lyrics

Query：怎么去北京——机票/火车票
Query：怎么去故宫——导航/公交/出租
Query：怎么去天堂——闲聊/？

落地背景

工业界技术设计的几个注意点：

- 可用性。方案模型能用，流程能跑通。
- 可控性。能有较为灵活的干预机制。
- 可靠性。能够支撑一定的在线压力。（耗时和并发）
- 快速迭代。不求最好，可用后通过迭代逐步优化。
- 算法效果。较高的算法效果。

PART TWO

分类模型

CLS任务

规则和词典	人工特征&ML	Embedding+DL	Bert+ fine tuning
正则	tfidf+ml	fasttext	BERT-CLS
词典匹配		textcnn	RoBERTa
		HAN	albert

- 工业应用上，规则词典仍然因为可控、稳定性占据主
- Emb深度学习方案给出稳妥的模型基线。
- 预训练语言模型效果具有统治级的地位，但由于复杂度和模型体积原因受到一定限制。
- 以搜代分可以做长尾召回的处理。

Fasttext: https://arxiv.org/abs/1607.01759 , TextCNN: https://arxiv.org/abs/1408.5882
HAN: https://www.aclweb.org/anthology/N16-1174.pdf
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
RoBERTa : A Robustly Optimized BERT Pretraining Approach
ALBERT: A Lite BERT For Self-Supervised Learning Of Language Representations
https://zhuanlan.zhihu.com/p/110648517 , https://zhuanlan.zhihu.com/p/183852900
我的分类文章：
https://mp.weixin.qq.com/s/WZZ8Qtxxd_TpSzgMhfAcSA， https://mp.weixin.qq.com/s/KrblC_JcjmPOZji4E1yaHg ，
https://mp.weixin.qq.com/s/DMkC0olB5KF_MsPlPr36nQ

CLS任务

fasttext

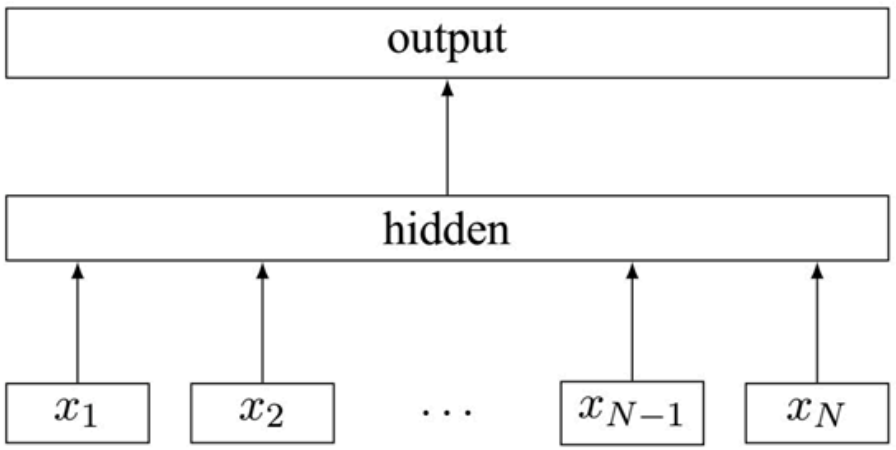
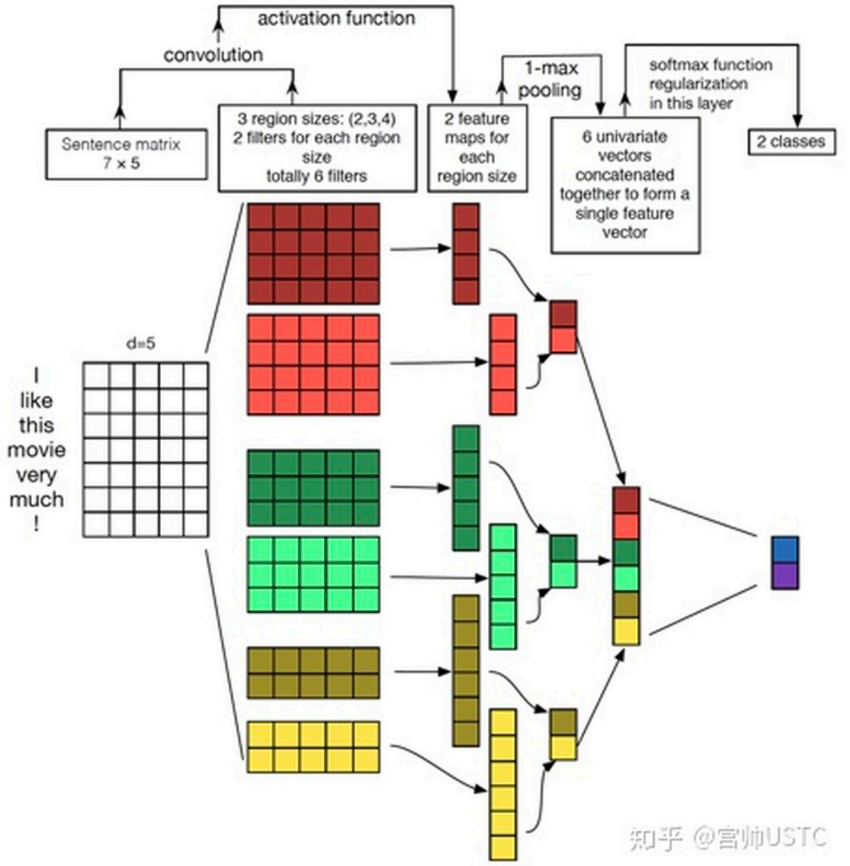


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

- 浅层模型，浅层信息
- 高性能，GPU用都嫌浪费。

https://zhuanlan.zhihu.com/p/32965521

textcnn

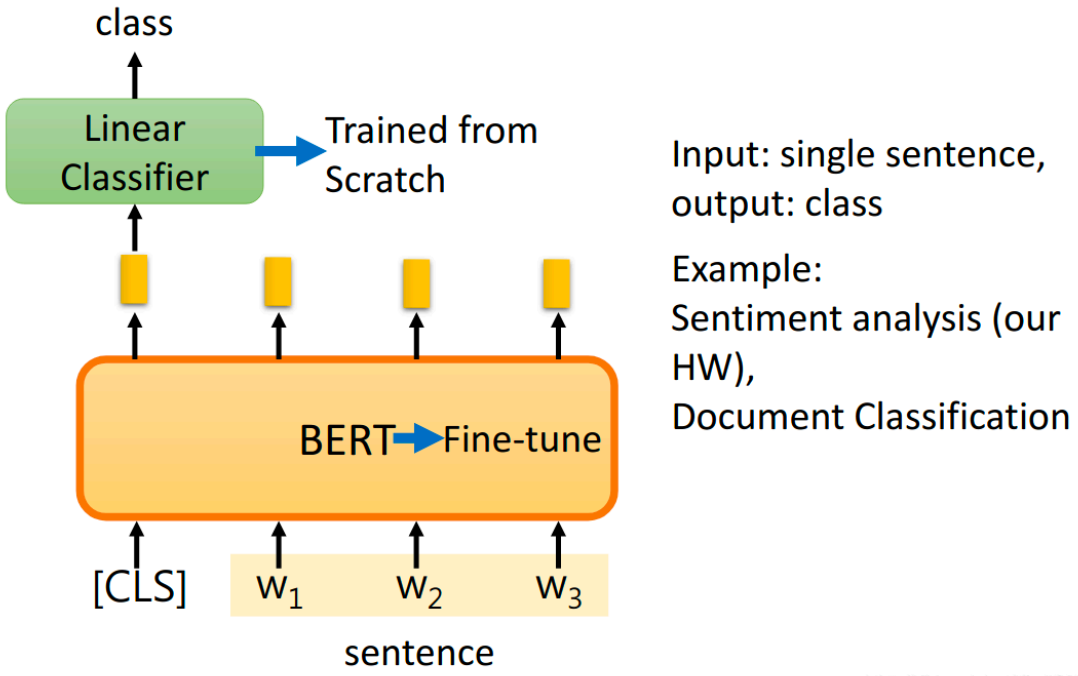


- 下限比较高。
- 轻松达到ms级。
- 落地场景其实不会距离bert太远。（因为短板不是模型）

https://zhuanlan.zhihu.com/p/40276005

CLS任务

Bert-cls



- 上限高。
- 体积大、性能差。（几十甚至上百ms）

- 加速手段：
  - 多机多卡。（$$$$$$）
  - ONNX/TVM加速。
  - 离线cache。
- 能提升在线的耗时和并发性，但是对资源的消耗非常大，人、卡、时间等。

https://blog.csdn.net/qq_36618444/article/details/106479882

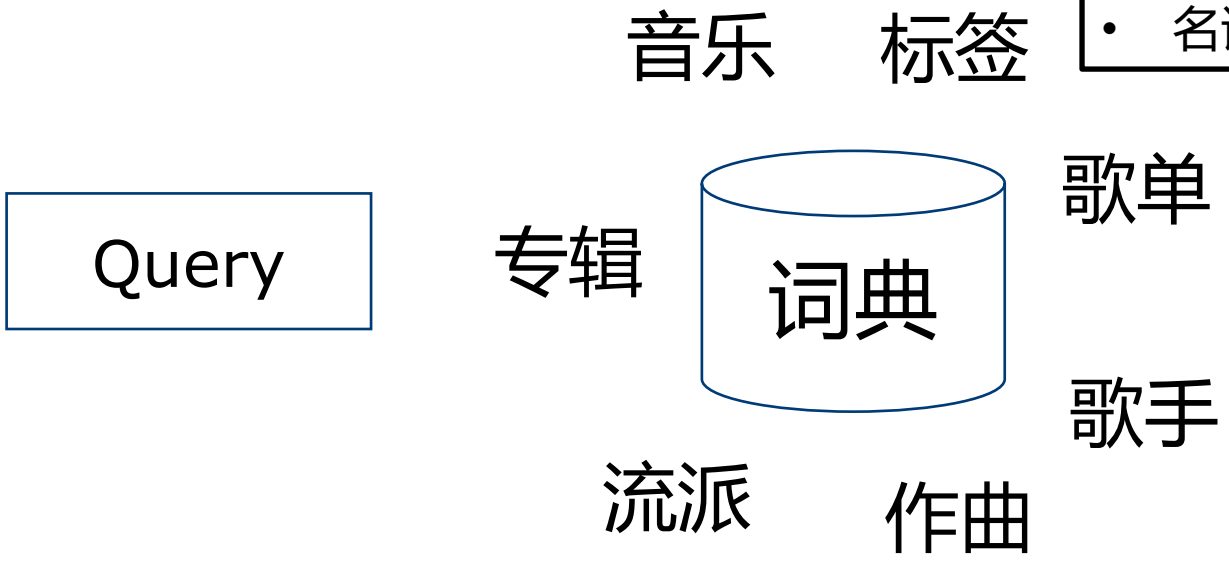
PART THREE

分类方法

分类方法——规则&词典

Query：七里香——music/album
Query：周杰伦——Singer

- 高精准，一般词典可靠的情况下，90%准确基本没问题。
- 高性能，算法设计好，则是一个和库大小无关的复杂度，只与query长度有关。
- 灵活可控。
- 词典需要挖掘。（难道标注样本就不需要？）
- 需要对业务、数据有较高理解。
- 名词性强的，知识型的分类。



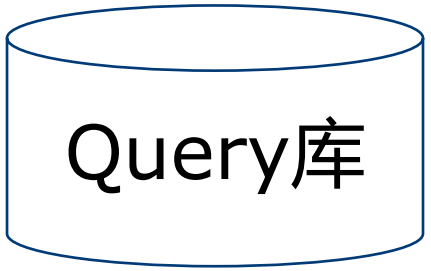
七里香 \001 歌名 \002 专辑
周杰伦 \001 歌手 \002 作曲
说唱 \001 流派
电音 \001 流派

算法题：
给定Query，找出Query中所有库里面包含的词汇。
输入：**我想听**周杰伦的七里香。
输出：周杰伦，七里香。

分类方法——以搜代分

- 灵活可控。
- 长尾零散的意图可用。
- 模糊搜索带来的泛化能力。
- 库带来的资源增加，且需要具备一定工程能力。
- 知识型分类。

Query



爱情来的太快就像龙卷风 \001 lyrics
手机蓝牙连不上耳机 \001 encyclopedia
五年高考三年模拟 \001 book

文本匹配

- 方案成熟，落地方便。
- 下限高。
- 不需要训练模型。

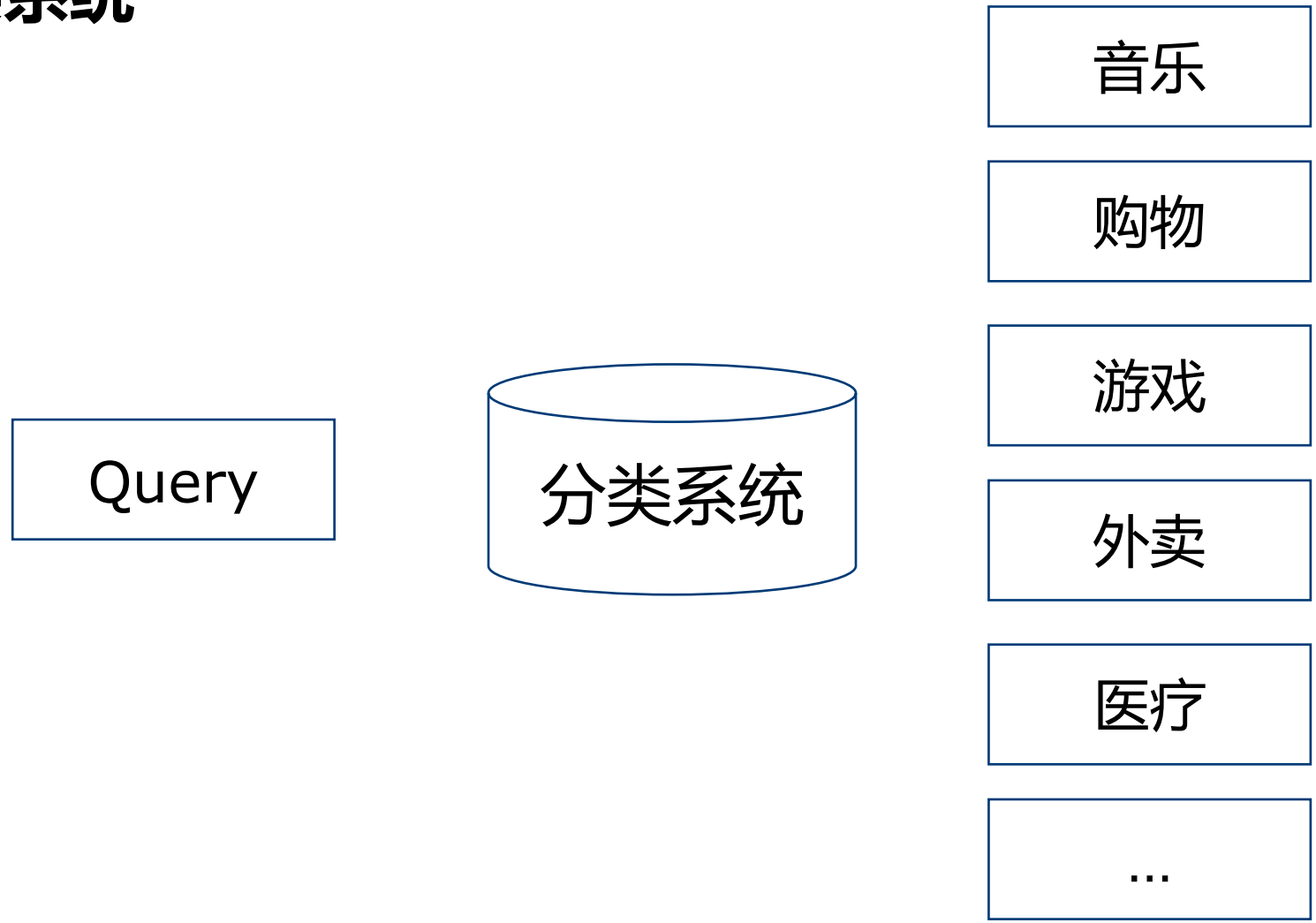
语义匹配

- 泛化能力强。
- 降低库的维护成本。
- 向量化预处理方便。

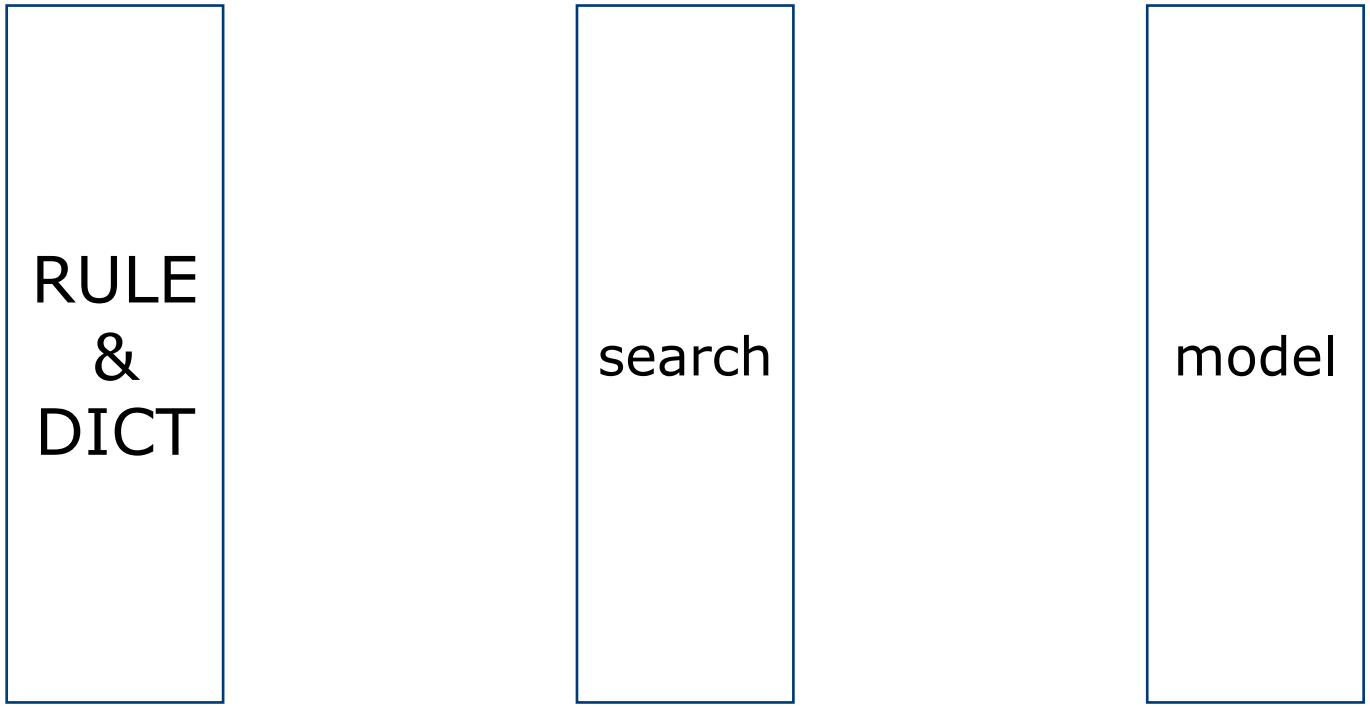
PART FOUR

分类系统

分类系统



分类系统



快速判断，快速拒绝
快速使用的干预能力

处理长尾、分散的情况

高泛化的模糊层

真正的工业界落地分类，并不是一招解决所有问题，而是多种方式有机组合完成。
兼顾干预能力、高性能等多个方面才是技术落地的关键。
技术从不谈高低端，而是能怎样更好地完成任务。
规则也会有准招，随着技术迭代，可以把模糊错误的逐步下放到search或model层。

PART FIVE

分类标签

多分类和多标签

长城：

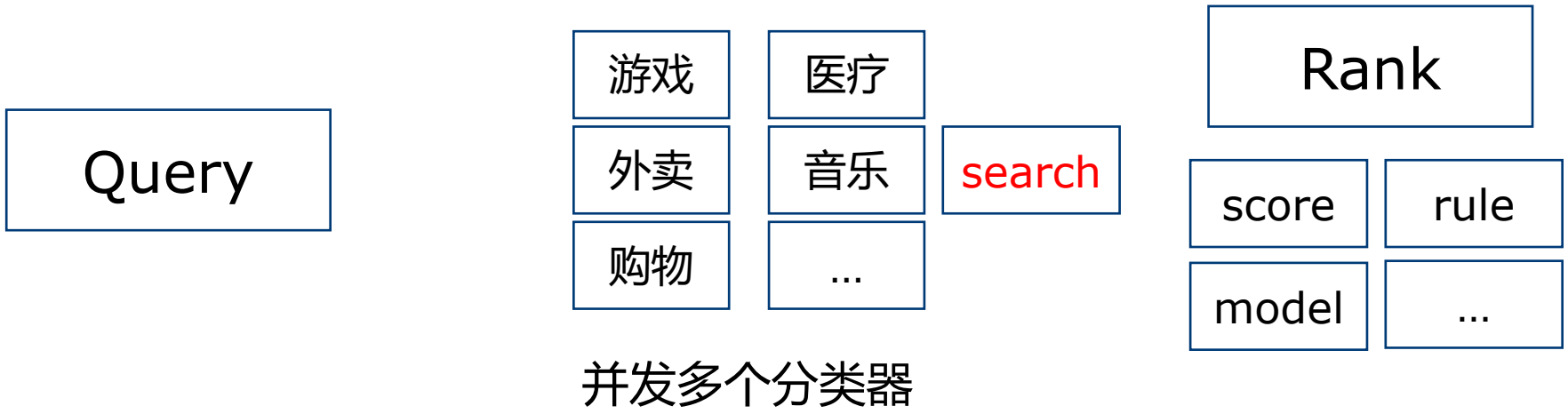
景点、汽车、电影、百科、宽带...

- 粗暴的方式：一个模型解决。

可控性？

迭代或新增一个类目，导致巨大的变化。

多个二分类+rank的方式处理



THANKS

叉烧，OPPO数据挖掘工程师。目前负责对话系统Query理解模块，先后在去哪儿网、美团、OPPO实习，毕业加入OPPO。

北京科技大学理学、经济学学士，理学硕士。至今累计发表论文7篇。

CS的陋室号主，累计原创文近300篇，超过50w字。

CS的陋室

我的微信