Query纠错（2） - 文本错误类型

原创 wedata24 WePlayData 2020-12-09

本文介绍一下常见的query错误都有哪些。首先看下为什么会出现query错误。常见的输入方式是拼音输入法，用户输入拼音，输入法提示候选词，由于误选或无需要候选词时，query就有可能出错。尤其当新的网络词汇出现时，如"耗子尾汁"，或者一些陌生的词，如"芈月传"，很多人都知道其拼音，实际哪些字并不确定。此外，通过复制粘贴来搜索，也可能导致搜索query不完整或带入无关字符。

关于纠错类型，最常见的是query中存在错别字，比如“尿酸高通风 -> 尿酸高痛风”。除此之外还会存在多字、少字、顺序交换、中英文混拼，中文拼音混拼等错误类型。下面列举一些常见的错误类型：

1、少字：微信跳一 -> 微信跳一跳
2、多字：微信跳一跳跳 -> 微信跳一跳
3、错字：微信 挑一挑 -> 微信跳一跳
4、拼音：tiaoyitiao -> 跳一跳
5、中英文混拼：held住 -> hold住
6、中文拼音混拼：跳yi跳 -> 跳一跳
7、知识错误：南山平安金融中心 -> 福田平安金融中心
8、音转：灰机 -> 飞机
...

这里有个细节值得注意，是否是错误还和是否被很多人来用有关，比如“西红柿首富”先验看是个错的电影名，正确的应该是“西虹市首富”，但是由于最开始的很多用户、自媒体都使用了“西红柿首富”，反而使得其是个正常的query，现在百度也专门有个“西红柿首富”的电影词条。这种问题对具体的纠错处理以及纠错应用都有比较大的影响。

从前面结果来看，qc的错误类型是非常多的，并且比较细，使得准确完成这些纠错本身就是个很困难的事情，所以对不同query，纠错系统产生的纠错结果的置信度也不一样。因此考虑到置信度，会对纠错结果标记为不同的纠错类型，并且影响下游的搜索策略和结果展示。具体纠错类型将在下一篇文章中介绍。后续将分别介绍以下内容：

2、纠错结果类型
3、纠错的召回方法
4、片段纠错
5、生成式纠错

6、先检后纠

7、纠错如何用于排序

相关阅读

1. Query理解 - 搜索引擎“更懂你”
2. 从搜一搜中检“相关性排序”的排序结果说起...
3. 搜索排序 = 相关性排序?
4. 搜索引擎新的战场 - 百度、头条、微信
5. 搜索引擎的两大问题（1） - 召回
6. 搜索引擎的两大问题（2） - 相关性
7. Query词权重方法（1） - 基于语料统计
8. Query词权重方法（2）- 基于点击日志
9. Query词权重方法（3）- 基于有监督学习
10. Query词权重方法（4）- beyond 词粒度
11. Query意图方法（1）- 基于片段意图
12. Query意图方法（2）- 基于文本分类
13. 搜索系统的评测方法
14. 搜索系统的架构设计
15. 你说百度更懂中国人，我说微信也挺懂中国人的
16. 在query理解中能ALL IN BERT吗?
17. Embedding搜索能代替文本搜索吗?
18. 【邢波】机器学习需多元探索，中国尚缺原创引领精神
19. 说一说视频搜索
20. Query纠错（1） - 原理