Query纠错（1） - 原理

原创 WeData WePlayData 2020-09-22

收录于话题
#搜索引擎 2 #推荐系统 1 #query纠错 1

query纠错（qc）是搜索系统中感知较明显的模块，从产品形态上并不陌生，也比较影响用户体验。从技术角度，纠错是query分析中难度比较大的模块，类似小型搜索系统，其可以分为两个阶段，一个是召回，即获取可能正确的候选纠错串，其次是排序，在候选纠错串中挑选纠错概率最大的作为纠错结果。本文先介绍一些query纠错的原理。

如果把纠错当成一个黑盒，其输入可称为原串：用户query（q），输出可称为纠错串：纠错后的query（r）。纠错算法就是选出r使得q纠成r的概率最大，用贝叶斯公式表示就是：

$$\begin{aligned} &\text{argmax } p(r|q) \\ =\ &\text{argmax } p(q|r)p(r)/p(q) \\ =\ &\text{argmax } p(q|r)p(r) \end{aligned}$$

所以纠错可以分解成：

p(q|r)：计算原串到纠错串的转移概率，常见方法有编辑距离，q和r的共点击概率，以及分别抽取q和r的相应特征，用模型预测p(q|r)等；
p(r)：衡量纠错串作为正常query的概率，比如语言模型，高频query，实体知识库等，即如果r的语言模型得分很高或者是个搜索次数比较多的query，其作为正确纠错串的概率就越大。

纠错后续的流程和不同方法大都是围绕着这两个概率来计算。

还有个问题是纠错是和资源相关的，如果某个错误query下的资源很多，虽然能正确纠成正确的query，但此时也可能不会去纠错。所以**p(q|r)、p(r)、p(q)**在纠错中都会使用。由此可见，纠错是个比较复杂的系统，后续将分别介绍以下内容：

1、文本错误类型
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

喜欢此内容的人还喜欢

Query纠错（2） - 文本错误类型