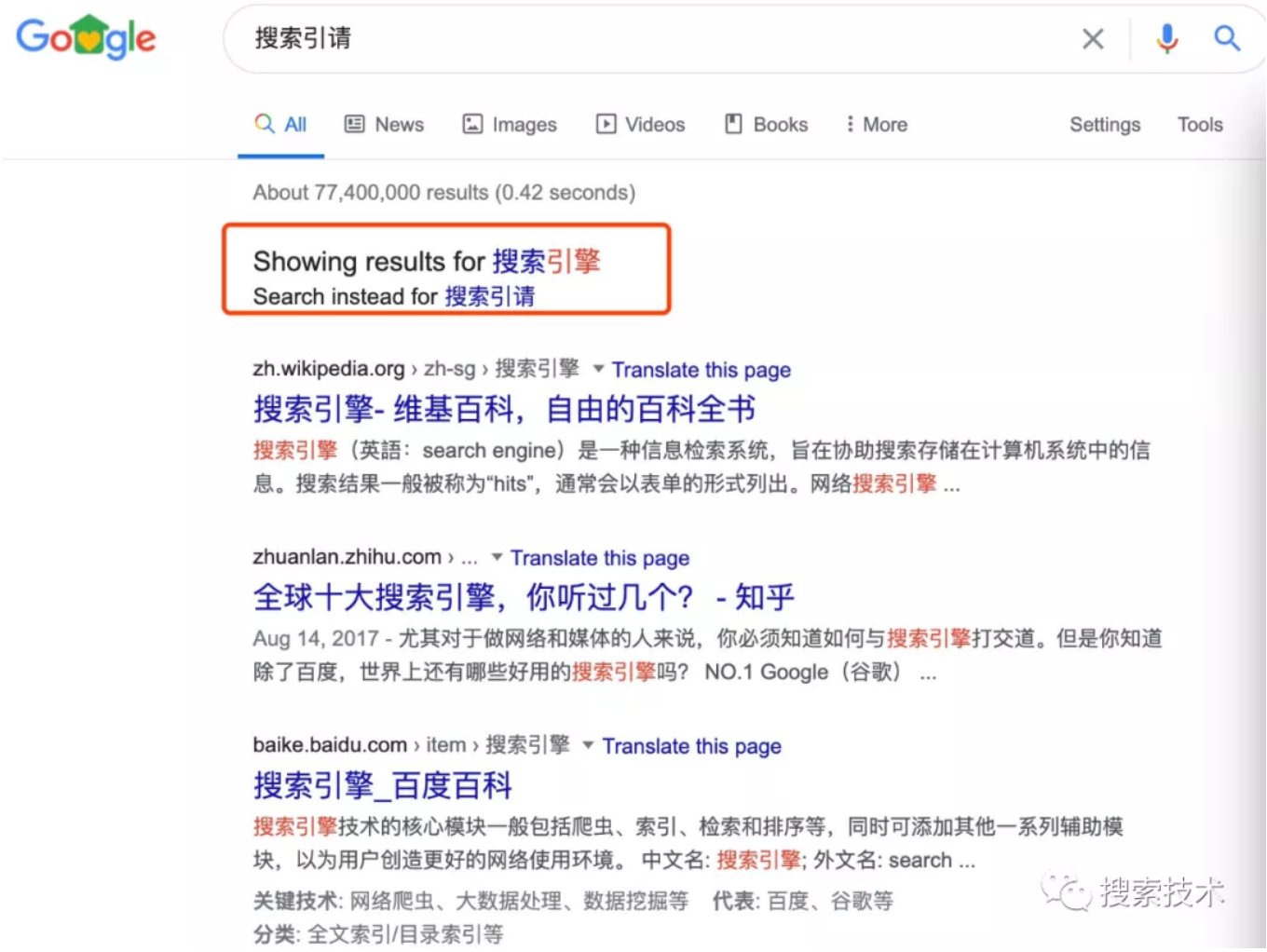
如何设计跟Google一样的搜索纠错引擎

原创 搜索技术 搜索技术 2020-04-30

第二代搜索引擎，以Google和百度为代表，发展了20多年，Google市值上万亿，搜索引擎技术在这20年是引领互联网整体技术发展的。

在国内IT圈都知道，百度出来的同学，技术都很厉害，在各个公司都是架构师和技术Leader或技术VP，年薪都是百万级别以上。所以学好搜索技术，走向人生巅峰，财务自由，未来可期（题外话）。



图一：Google搜索："搜索引清" 被纠错为 "搜索引擎"

搜索引擎是根据用户输入的关键字（Query）,给出一些用户想要的结果集，让用户选择。从用户输入关键字到搜出结果，在搜索引擎中，经历了一系列复杂的算法过程，才能找到用户想要的结果。**有时用户还输入错误的关键字，这样会得到不好的搜索结果或者相关性很差，这样会严重影响用户体验，纠错引擎就是来解决这个问题了，可以显著提高搜索体验。**

纠错的方法：

纠错一般放在query重写（query rewrite 简称QR）模块，QR模块主要包括拼写纠错，同义词近义词，关联query等多个功能。QR模块是个很重要的模块，拼写错误的query通过QR模块，改写成正确的query，显著提升用户搜索体验。

用户拼写错误分为两种，一种是Non-word Error, 是指单词本身错误，另外一种是Real-word Error, 单词本身拼写正确但组词或者短语就错误了，比如“天安门”被拼写成“田安门”。

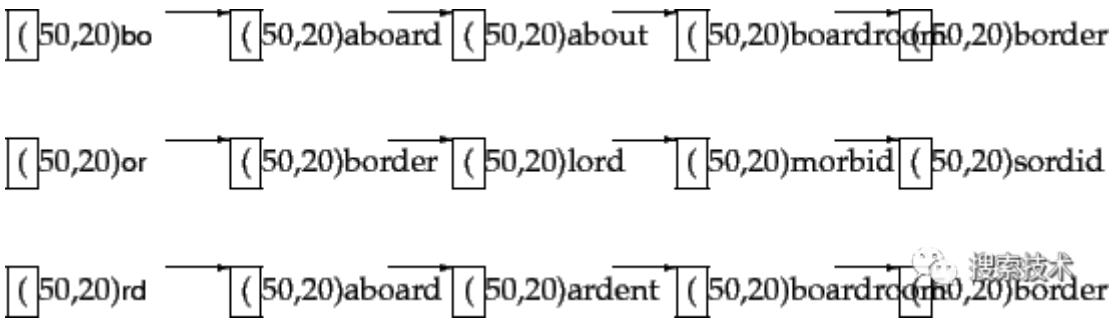
中文错误一般都是Real-word Error，因为中文都是输入法打出来，单个字都不可能错误。英文就不一样，比如hello被写成heloo，这就属于单词本身的错误，Non-word Error。

纠错方案选择：

1) 解决拼写错误，使用拼写纠错库，可以通过搜索日志挖掘出来。

2) 中文输入大部分是使用拼音输入法，很多都是相同的拼音，词选错了，所以需要构建同音纠错库。

3）n-gram词库，解决中英文输入错误问题召回的，定义为编辑距离词库。



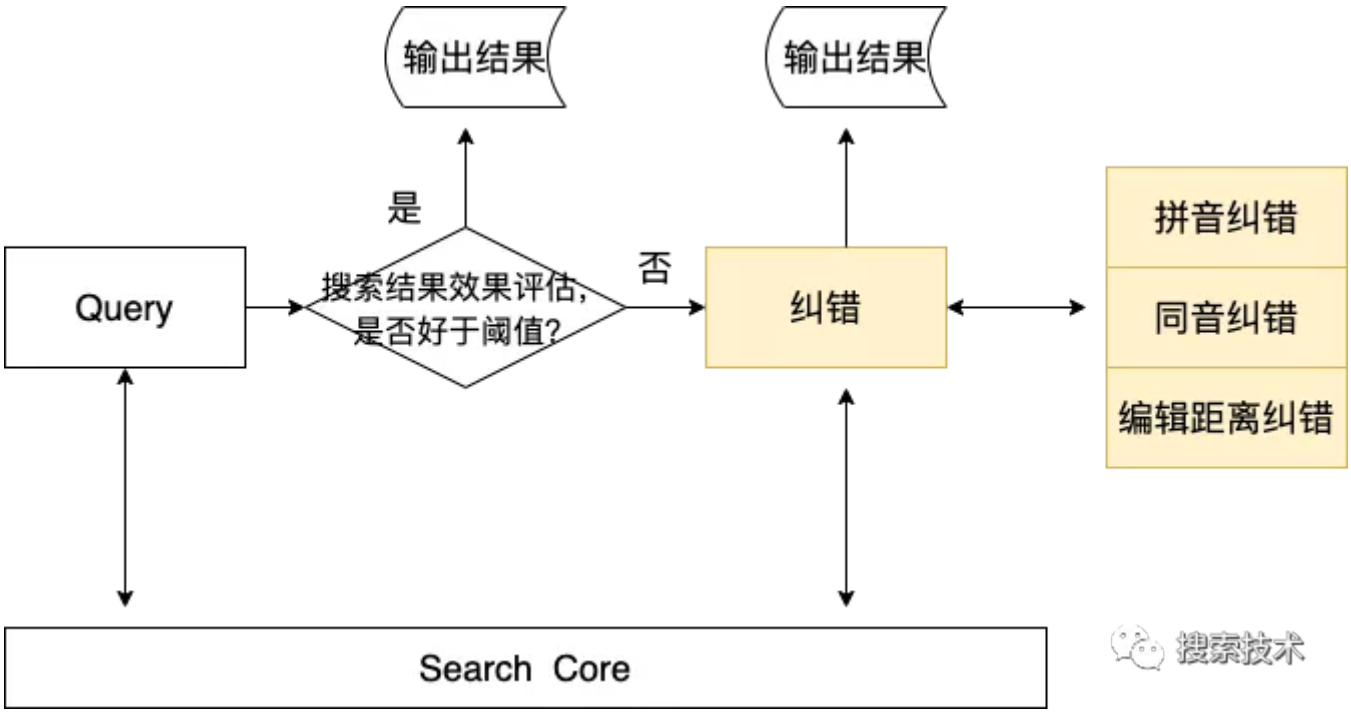
图二: query:bord, 使用 2-grams匹配过程.

纠错系统流程：

1）原始query调用搜索核心模块，获取到搜索结果，并对搜索结果进行评估打分。如果结果评分小于阈值，就调用纠错模块，否则输出搜索结果。

2）使用拼音纠错，同音纠错和编辑距离纠错等混合纠错方式，对纠错结果进行评估，选择最优纠错结果。

3）纠错结果去调用搜索核心模块，返回纠错后搜索结果。



图三：搜索纠错流程

总结：

纠错的好坏，可以从搜索日志和用户的点击行为中，进行评估，反过来优化策略，这个系统是个长期优化的过程，并且纠错词库需要线下挖掘出来的。