达观数据搜索引擎的Query自动纠错技术和架构详解

高翔 达观数据 2016-02-05

达观数据搜索引擎
Query自动纠错技术和架构

1 背景

如今，搜索引擎是人们的获取信息最重要的方式之一，在搜索页面小小的输入框中，只需输入几个关键字，就能找到你感兴趣问题的相关网页。搜索巨头Google，甚至已经使Google这个创造出来的单词成为动词，有问题Google一下就可以。在国内，百度也同样成为一个动词。除了通用搜索需求外，很多垂直细分领域的搜索需求也很旺盛，比如电商网站的产品搜索，文学网站的小说搜索等。面对这些需求，达观数据(www.datagrand.com)作为国内提供中文云搜索服务的高科技公司，为合作伙伴提供高质量的搜索技术服务，并进行搜索服务的统计分析等功能。（达观数据联合创始人高翔）

搜索引擎系统最基本最核心的功能是信息检索，找到含有关键字的网页或文档，然后按照一定排序将结果给出。在此基础之上，搜索引擎能够提供更多更复杂的功能来提升用户体验。对于一个成熟的搜索引擎系统，用户看似简单的搜索过程，需要在系统中经过多个环节，多个模块协同工作，才能提供一个让人满意的搜索结果。其中拼写纠错（Error Correction，以下简称EC）是用户比较容易感知的一个功能，比如百度的纠错功能如下图所示：

图 1：百度纠错功能示例

EC其实是属于Query Rewrite（以下简称QR）模块中的一个功能，QR模块包括拼写纠错，同义改写，关联query等多个功能。QR模块对于提升用户体验有着巨大的帮助，对于搜索质量不佳的query进行改写后能返回更好的搜索结果。QR模块内容较多，以下着重介绍EC功能。

在搜索引擎中，我们将用户输入的关键字查询叫做query，用户希望得到和输入query相关的质量较好的网页或文档，这个“好”字定义有多种衡量方式，最简单的标准就是那些对用户帮助最大最具吸引力的结果能够排到前列，搜索工程师们也在努力通过各种算法的提升来达到这个目的。但是往往出于各种原因，用户输入的query本身质量不高或是错误的，如果搜索引擎不对这种错误进行修正弥补，会导致召回错误的结果，或者结果数少甚至没有结果。

当用户看到搜索结果较差较少时，如果能意识到自己的query错误，对query进行修正再次检索，也许能找到想要的结果。但有时用户也不知道自己的query错在哪里，这个时候就会非常着急。笔者之前从事搜索相关工作时，刚开始搜索系统不支持纠错功能，结果收到用户大量的吐槽和投诉，说明没有纠错功能的搜索系统会大大降低用户体验，不仅如此，这些错误query检索还浪费大量的流量。当开发完毕并在搜索系统中使用EC模块后，纠错成功的流量占到总流量的2%，不仅提升了用户体验，还能够挽回流量损失，提升用户粘度。

2 EC常见错误

EC应该怎么做？首先我们看一下常见的query错误都有哪些。

对于英文，最基本的语义元素是单词，因此拼写错误主要分为两种，一种是Non-word Error，指单词本身就是拼错的，比如将“happy”拼成“hbppy”，“hbppy”本身不是一个词。另外一种是Real-word Error，指单词虽拼写正确但是结合上下文语境确是错误的，比如“two eyes”写成“too eyes”，“too”在这里是明显错误的拼写。

而对于中文，最小的语义单元是字，往往不会出现错字的情况，因为现在每个汉字几乎都是通过输入法输入设备，不像手写汉字也许会出错。虽然汉字可以单字成词，但是两个或以上的汉字组合成的词却是更常见的语义元素，这种组合带来了类似英文的Non-word Error，比如“洗衣机”写成“洗一鸡”，虽然每个字是对的，但是整体却不是一个词，也就是所谓的别字。汉字也有类似Real-word Error的问题，比如加薪圣旨，加薪和圣旨都是正确的词，但是两个连在一起确有问题，因此很多情况下汉语query纠错实

际上是短语纠错问题。Query除了纯汉字外，现在还会出现中英文混拼错误，中文拼音混拼等错误。下图是笔者在搜索日志中找到的一些常见错误query：

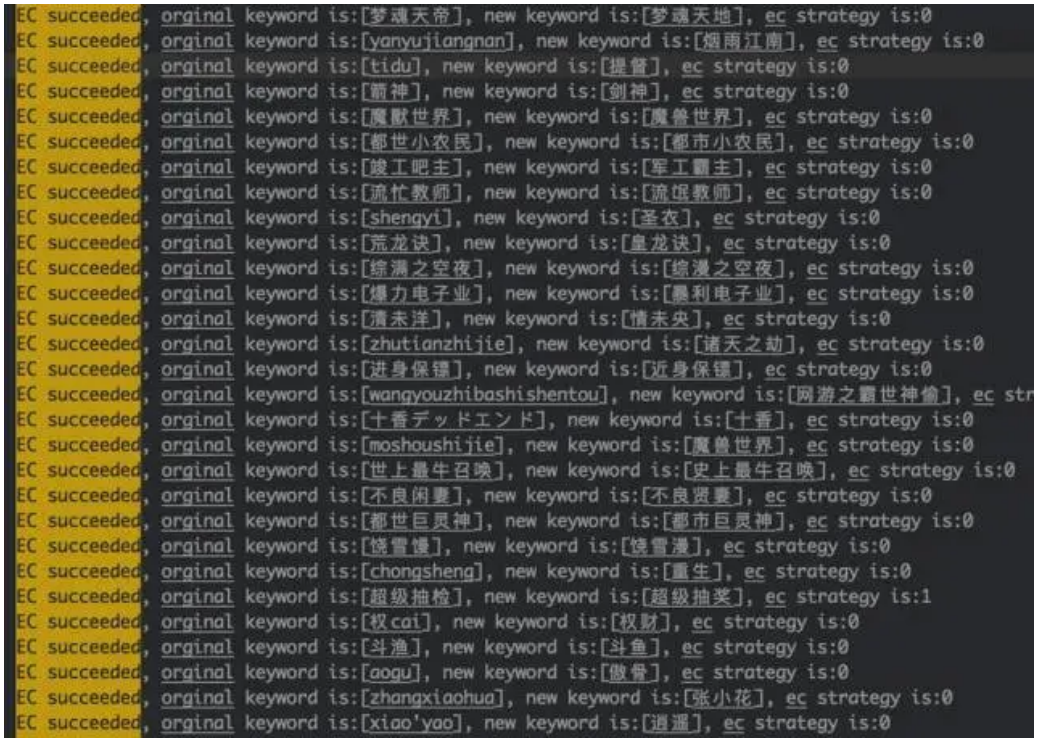

图 2：搜索日志中的错误query

从上图可以看出，中文搜索中常见的错误query主要包括别字，纯拼音，模糊音，拼音汉字混合，拼音其他符号混合等多种问题。

3 Query出错的原因分析

目前最普遍的中文输入方式是拼音输入法，用户输入拼音，输入法给出候选词，但是由于用户误选或无需要候选词时，query就有可能出错。虽然相较之前智能输入法现在已经足够强大，但仍有一些新的产品、小说、影视作品，输入法可能会覆盖不到。比如一些新奇网络词汇的出现，传统的词典已经无法包括这些词。还有一些较为陌生的词，比如“芈月传”，很多人都是听朋友介绍很好看，结果去搜索引擎中搜索相关信息，很多人只知道第一个字发音是“mi”，但实际是哪个字却不确定。

除此之外，用户搜索时也会从网页或其他文档上复制粘贴文字来搜索，导致搜索query不完整或带入其他字符，甚至打字太快也是错误query输入的原因。

4 Query纠错方案

英文拼写纠错已经有较长的历史，对于英文纠错的研究较多。英文纠错是中文纠错的重要基础，其中很多算法思想同样适用于中文。因此首先介绍一下英文纠错问题。在介绍

具体纠错方案前，先介绍两个重要的的概念：编辑距离，n元语法模型。

4.1 基础概念

4.1.1 编辑距离

编辑距离是两个字符串之间，由一个转换成另外一个所需要的最少操作次数，允许的操作包括字符替换，增加字符，减少字符，颠倒字符。举例来讲，apple和apply的编辑距离是1，access和actress的编辑距离是2，arrow和brown的编辑距离是3，编辑距离的计算操作过程如下图所示：

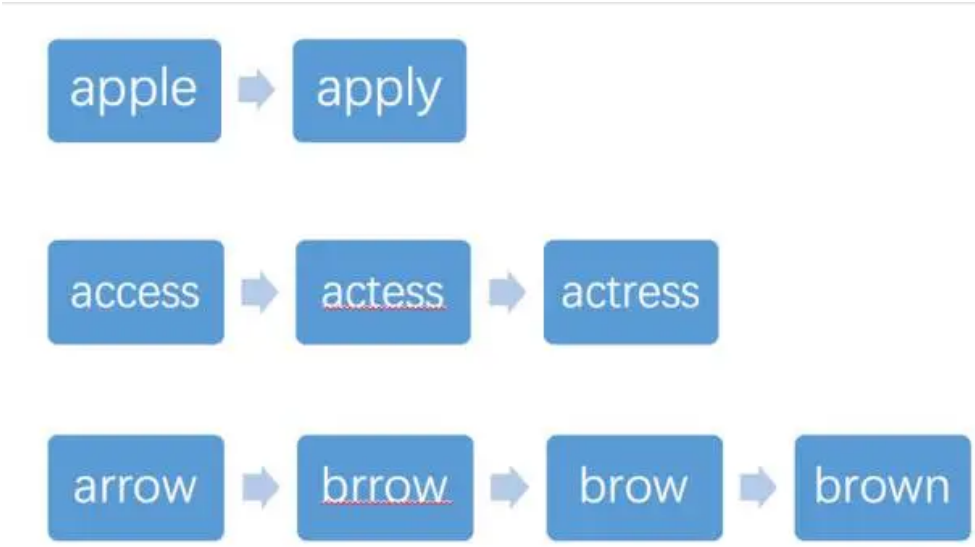

图 3：编辑距离计算过程

4.1.2 n元语法模型

语言模型（language mode）在基于统计模型的语音识别，机器翻译，汉语自动分词和句法分析中有着广泛的应用，目前采用的主要是n元语法模型（n-gram model）。

一个语言模型构建字符串的概率分布*p(W)*，假设*p(W)*是字符串作为句子的概率，则概率由下边的公式计算：

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

公式 1：语言模型

其中w1表示第一个词，w2表示第二个词，以此类推。*p(w4/w1w2w3)*表示前面三个词是*w1w2w3*的情况下第四个词是*w4*的概率。

*w1w2...wi-1*称作历史，如果w共有5000个不同的词，i=3的时候就有1250亿个组合，但是训练数据或已有语料库数据不可能有这么多组合，并且绝大多数的组合不会出现，所以可以将*w1w2...wi-1*根据规则映射到等价类，最简单的方式就是取*wi*之前n-1个历史，根据马尔科夫假设，一个词只和他前面n-1个词相关性最高，这就是n元语法模型。

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i \mid w_{i-k} \dots w_{i-1})$$

公式 2：n元语法模型

通常用的n元语法模型包括unigram，bigram，trigram，其中unigram表示这个词和前面的词无关，彼此独立，计算公式如下：

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i)$$

公式 3：unigram语法模型

Bigram表示一个词只和它前面一个词有关，计算公式如下：

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

公式 4：bigram语法模型

4.2 英文纠错

4.2.1 Non-word纠错

纠错首先要检测出错误。检测错误的方法有很多种，对于Non-word错误可以使用语料库字典，如果输入词不在字典中，即可以判定为错词。

纠错过程就是找出和错词最相似的一些候选词，然后从中选出正确的纠错词。候选词可以使用上面介绍的编辑距离从语料库中找出。统计指出，80%的错误词的编辑距离是1，并且几乎所有的错误的编辑距离在2以内。

在候选词中找到最终的纠错词，比较简单的方法可以根据候选词的权重进行排序，给出权重最高的词作为纠错词，这个权重可以是人工标注的结果，也可以是语料库统计的词频或其他方式。相对复杂的候选词选择方法可以使用统计模型计算，比如噪声信道模型。

噪声信道模型（Noisy Channel Model）最早是香农为了模型化信道的通信问题，在信息熵概念上提出的模型，目标是优化噪声信道中信号传输的吞吐量和准确率。对于自然语言处理而言，信道噪声模型如下图，其中*I*表示输入，*O*表示经过噪声信道后的输出，*I'*表示经过解码后最有可能的输入。

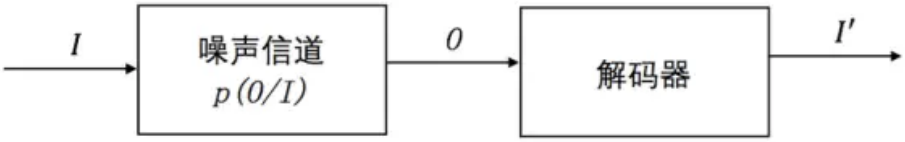

图 4：信道噪声模型框图

在自然语言处理中，很多问题都可以归结为给定输出*O*（有可能包括错误信息），在所有可能的输入*I*中找到最可能的那一个作为输入*I'*。

自然语言处理中的机器翻译，词性标注，语音识别等多个问题都可以使用信道噪声模型来解决，对于纠错问题也可以使用信道噪声模型来解决，相应的求解问题可以用公式表达：

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

公式 5：噪声信道模型纠错公式

其中$p(x|w)$是正确的词编辑成为错误词x的转移概率，包括删除（deletion）、增加（insertion）、替换（substitution）和颠倒（transposition）四种转移矩阵，这个转移矩阵的概率可以通过统计大量的正确词和错误词对来得到，转移矩阵的计算公式如下：

```
del[x,y]:    count(xy typed as x)
ins[x,y]:    count(x typed as xy)
sub[x,y]:    count(x typed as y)
trans[x,y]:  count(xy typed as yx)
```

$$P(x|w) = \begin{cases} \frac{\text{del}[w_{i-1},w_i]}{\text{count}[w_{i-1}w_i]}, & \text{if deletion} \\ \frac{\text{ins}[w_{i-1},x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i,w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \frac{\text{trans}[w_i,w_{i+1}]}{\text{count}[w_iw_{i+1}]}, & \text{if transposition} \end{cases}$$

公式 6：转移矩阵公式计算

将转移矩阵计算公式代入公式5的噪声信道模型公式中，根据不同候选词和纠错词之间的变换关系选择转移矩阵类型，就能得到概率最大的候选词。

4.2.2 Real-word纠错：

有研究报告指出指出有40%~45%的错误属于Real-word Error问题。Real-word问题中，每个词都是正确的，但是组合在一起成为短语或句子时意思却不对。因此纠错策略和Non-word有些不同。首先是候选词集合的生成，对于句子或短语中每个词生成候选集合，这个集合包括：1，这个词本身；2，所有和这个词编辑距离为1的词；3，同音词。集合选定后，选择最佳候选对象或组合时，可以使用的方法有噪声信道模型及特殊分类器。

噪声信道模型和Non-word纠错类似，只是计算目标从某个候选词的最大概率变成不同位置候选词组合形成的句子$p(s)$的最大概率，这个问题可以使用HMM（Hidden Markov model，隐马尔可夫模型）求解。

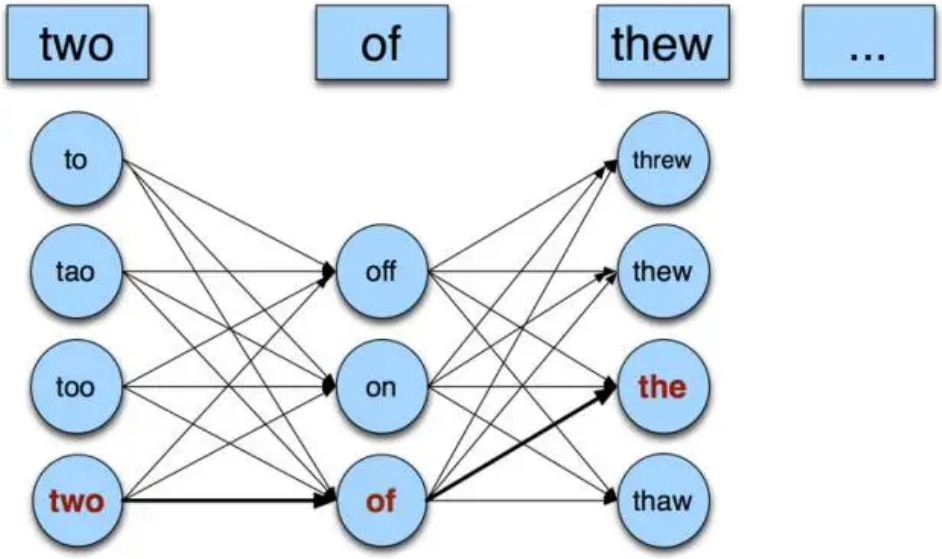

图 5：噪声信道模型纠错Real-word Error问题

上图中，每一数列是这个位置词的候选词集合，其中每个词的状态转移概率可以通过语言模型在语料库中统计求得。

基于分类方法纠错，分类器将会根据多个特征训练出一对Real-word之间的转移模型，常见的分类器包括SVM（Support Vector Machine，支持向量机）或者是基于规则的分类器，特征可以选择每个word的unigram，bigram概率等。

4.3 中文纠错

中文纠错以英文纠错为基础但却有所不同。在中文中，一般情况下错误词和正确词的长度是相同的，只是指定位置上的某一个字有误，因此状态转移矩阵只有替换一种。其次是中文词语往往较短，即使编辑距离只有1，就会有大量的候选词，存在较大的转义风险。中文使用拼音作为文字的发音，每个字都有固定的发音（多音字除外），而拼音输入法占据中文输入方式的主导地位，导致错误query中的别字同音但字形错误。因此中文纠错以拼音为基础，编辑距离等其他方式为辅的策略。

4.3.1 候选词集合的获取

对于错误的词的候选词集合，可以通过数据自动挖掘来生成。英文候选词集合一般通过编辑距离来获得，而中文候选词集合使用和错误词有相同的拼音的词组成，比如“嗒衣”这个错词的拼音是“dayi”，则可以通过事先挖掘好的拼音是“dayi”的词组成候选集，比如<dayi:大一，大衣，大意，大姨，搭衣，答疑..>。

4.3.2 候选词的选择

对于纠错候选词的选择就是一个对候选词进行排序，按照一定的排序规则，把排名最高的候选词作为最佳纠错结果返回。排序规则可以使用词频等多种特征，候选词按照这些特征规则进行排序，返回权重较高的词。

对于一个无上下文关系的词进行纠错，候选词的选择会比较困难，比如上面“嗒衣”这个错词的候选词有很多，无论按照哪一种方式进行排序，都存在较为严重的转义风险，这时可以使用编辑距离等其他方式辅助选择。

相对于单独一个词的query，由多个词组成的query纠错相对更加精准，每个词存在上下文关系约束，整个query的意图更加明确。通过对query分词，查找每个词的候选词集合，然后使用和英文Real-word纠错类似的方式纠错。

除了搜索日志query和语料库的统计挖掘，搜索系统中的session分析和点击模型提供的数据也能够为query纠错服务。搜索session指的是用户在某一个时间段内的搜索行为，如果把搜索日志按照时间排序，对于某一个用户的搜索日志来说，可以看到用户的搜索行为是分段的，每段之间往往有较为明显的间隔，每一段我们称为一个搜索session。一般来说，用户在一个session内的搜索行为都是为了解决一个问题，因此在此session内用户输入的query往往都是相关的。

点击模型中的一些统计数据可以判断一个搜索query质量的高低，质量高的query往往会给出较好的结果，用户点击的欲望更高。举例来说，“度假”（正确）“渡假”（错误）这两个query，假设用户输入较多的是错误的“渡假”，系统给出结果会比较差。下图例子中“渡假”的搜索结果都没有命中标题，而标题往往是用户最为关注的信息，如果标题中不含有搜索query关键字，用户点击的欲望也会较低。

图 6：错误query“渡假”的结果少，质量差

图 7：正确query“度假”的结果多，质量好

在这种情况下，虽然“渡假”的搜索次数更多，但是点击模型给出query分数会比较低，而候选词“度假”的query得分就会高一些，可以辅助其他纠错方式完成纠错。

4.4 存在的问题

搜索系统许多功能的召回率和准确率是矛盾的，但是在query纠错问题中，准确率往往要求更高。拼音query到汉字query的纠错，往往会存在较大的转义风险，不同的类型的拼音转换方式（全拼，模糊全拼，简拼，混拼）有着不同程度的转义风险，召回越大则准确率降低，因此使用全拼较为稳妥。（达观数据联合创始人高翔）

5 达观数据搜索系统query纠错技术介绍

达观数据在搜索引擎等大数据技术上有着深厚的积累，搜索引擎提供多种功能及服务，其中纠错模块是比较重要的功能之一。

5.1 纠错过程

对于搜索中的query纠错功能，纠错过程主要分为以下3个过程：

1， Query纠错判断。对于常见错误，例如常见的拼写错误，使用事先挖掘好的错误query字典，当query在此字典中时纠错。如果用户输入的query查询无结果或结果较少于一定阈值时，尝试纠错，可以根据不同领域的策略和容忍度，配置最少结果数阈值。

2，不同策略独立纠错。达观数据使用多种纠错策略，主要使用拼音纠错和编辑距离纠错，并辅助模糊音形近字二次纠错等其他纠错策略。同音策略是用户输入的错误query和候选纠错query有相同的拼音。编辑距离策略就是错误query和候选query之间编辑距离小于一定阈值，并配合其他条件进行过滤。

3，候选词结果选择。因为每个策略比较独立，不同策略会给出不同的候选词，因此对于候选词的选取，每个策略有所不同。不同策略之间，不同策略内部需要使用不同的评估方式，来选择最优结果。

达观科技搜索系统的纠错模块包括上述多个策略，每个策略独立运行，针对不同的领域和业务情况，策略优先级和权重可配置，纠错松紧度可调节。

5.2 系统设计

达观数据EC系统主要分为三部分：数据模块，离线建库端及在线检索端。

图 8：EC系统模块构成

5.2.1 数据模块

数据模块的主要作用是为后面的离线建库端和在线检索端提供数据。

数据模块对搜索log定期进行抽取和统计，对query进行归一化后给出query频次词典。对数据库信息整理给出自定义词典。通过爬虫系统爬取优质词条词典。

5.2.2 离线建库端

离线建库端使用数据模块准备好的各种词典生就纠错词典，包括拼音纠错词典，编辑距离纠错词典等。根据配置，对频次词典中对超出一定长度query上述操作不处理。

5.2.3 在线检索端

在线检索端负责query实时纠错，根据5.1节的三个步骤进行。如果第一次纠错query查询结果较差，使用扩大召回的方式，比如二次纠错、片段纠错等扩大召回重新纠错，进行二次查询并返回质量较高的查询结果。

5.3 纠错效果评估

纠错效果的好坏从微观上来讲，可以查看搜索日志中无结果或结果少的纠错query以及点击模型中点击较少的纠错query等方式发现bad case，在针对这些bad case出现的原因进行分类总结，后续改进算法。

在宏观上可以关注搜索效果评估系统中的MAP和MRR分数，使用AB test，查看使用纠错模块后或纠错算法升级后的带来的效果提升。

6结语

一个完善的搜索引擎系统中，纠错功能是重要的一环，对提升用户体验及用户满意度有很大的帮助，亦能补救大量错误query所带来的流量损失。达观数据在搜索引擎服务上有着丰富的行业经验，能够为合作企业提供高质量的搜索服务，充分挖掘企业的数据价值。（达观数据联合创始人高翔）

技术犹如满天繁星，指引我们前行。（*题图笔者摄于Mt Cook*）
了解更多大数据技术和服务尽请访问达观数据(www.datagrand.com)