前沿重器[2] | 美团搜索理解和召回

原创 机智的叉烧 CS的陋室 2020-09-19

收录于话题
#搜索 10 #自然语言处理 14

所向无前 (Take Over)

英雄联盟 - 所向无前 (Take Over)

【前沿重器】

全新栏目，那么栏目主要给大家分享各种大厂、顶会的论文和分享，从中抽取关键精华的部分和大家分享，和大家一起把握前沿技术。具体介绍：仓颉专项：飞机大炮我都会，利器心法我还有。

往期回顾

- 前沿重器[1] | 微软小冰-多轮和情感机器人的先行者
- NLP.TM[38] | 对话系统经典：检索式对话
- SIGIR20最佳论文：通往公平、公正的Learning to Rank!
- NLP.TM[37] | 深入讨论纠错系统
- NLP.TM[36] | NLP之源：n-gram语言模型

搜索做了很多年，但是在各种技术革新下也还总有东西做，总有提升点，虽然现在媒体炒的少了，但是至今仍然各种公司仍花费大力气来做这个搜索。这次和大家介绍的东西，来自于美团技术团队分享的一篇文章，这篇文章讨论了搜索的理解和召回，有意思的是他还对整个他们的现状分析进行了讲解，这个收获挺大的，链接摆出来，其实可以看到，是一个比较垂直领域——旅游：https://tech.meituan.com/2017/06/16/travel-search-strategy.html。

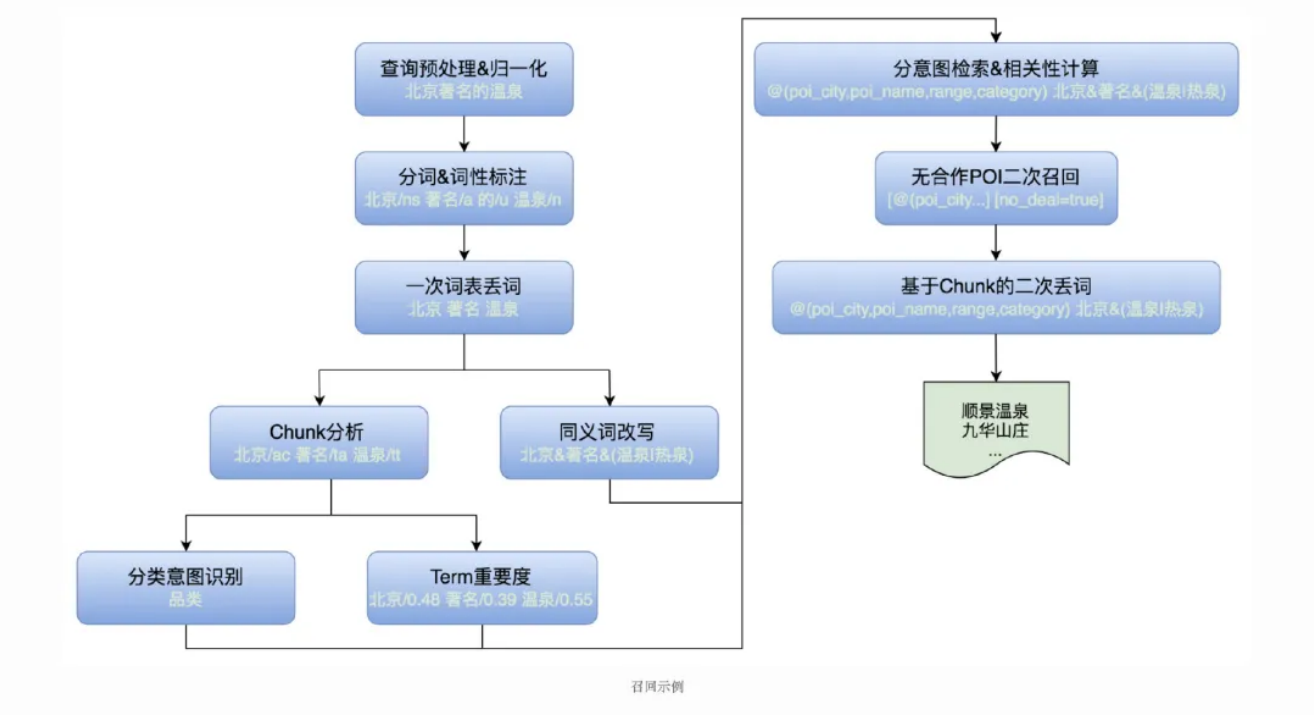
原文是按照探索迭代过程来讨论的，我不打算再这么讲一遍，意义不大，大家看原文就完事了，所以我打算整理成几点来讨论。

懒人目录：

- query理解的流程和操作。
- 美团旅游搜索用的技巧。
- 迭代方向调研思路。
- 小结。

query理解的流程和操作

常规的query理解流程其实我已经聊过几次了：R&S[24] | 浅谈Query理解和分析，这里讲的这里关注到的大小技术。首先看看整个流程美团搜索的安排：



美团旅游搜索流程

- 预处理和归一化主要负责的是规范化整个句子，大小写、异常字符之类的，这个大家应该都比较好理解吧。
- 分词和词性标注，这个是经典的NLP任务了，jieba之类的能给个基线，当然想提升也可以自己再整整。
- 一次性词表丢词，其实就是删除一些不必要的词汇，美团场景下就有一副、一张之类的，其实并不需要放入搜索引擎进行搜索了。
- 后面是一系列的chunk（实体）、意图、改写等一系列的工作。
- 相关性计算。
- 召回和二次召回等。

我从上面抽关键点来详细聊一下吧。

分类意图识别

意图识别的初衷在于对用户划分需求来进行定制化服务，对数据进行分块化管理：R&S[25] | 搜索中的意图识别。在这个基础上，美团对旅游这块的需求进行了类目体系的构建，设计了8类意图：

- poi。经典、游乐场，度假村等。
- 行政区。国家、省、市、区、县等。

- 品类：POI品类体系的品类词，如公园、体验馆等。
- 线路游：一日游、跟团等。
- 旅游关键词：旅行、游玩等。
- 旅行社。
- 门票：门票、套票、成人票等。
- 非旅游：美食、住宿等。

这8个类，针对的其实不是query本身，而是对应query内的词汇进行处理标注，文中叫chunk，我们会把它叫做实体，所以这个任务其实就是命名实体识别了。对query进行实体识别以后，我们就可以根据这个识别，结果，设立特定规则，如线路游>POI>品类>门票，比如“北京故宫一日游”是线路游意图，“北京故宫”是POI意图，“北京动物园”是POI意图，“动物园”是品类意图。

此处我们也可以看到，借助实体识别结果来完成意图识别任务其实也是一种方法。

Chunk识别

既然上面提到了chunk识别，那肯定要好好说到的说道事情了，其实就是一个实体识别任务（NER：NLP.TM[18] | 搜索中的命名实体识别），这么聊其实就是一个非常常规的NLP任务了。

NER任务上模型方法都算简单的，困难的反而是样本的收集，文章中提到的方式值得借鉴，来源的其实就是用户一段时间以内的搜索日志，借助词表规则进行标注（这个我之前讲过，见：NLP.TM[29] | ner自动化打标方法），然后再进行人工的校验。这里的词表则来源于各种模板规则从用户query中的挖掘，可见这里其实需要花费很多的人力物力。

在模型方法上，使用的是简单的CRF，工具是CRF++，有意思的是，他们并没有直接把文本扔进去就完事了，而是放入了大量的特征，主要分为3类：

- 边界特征，如左右熵、互信息。
- tag特征，如词汇长度、term中的tag类别等。
- 组合特征。

term重要度

之前我就已经有文章提到了词权重问题（NLP.TM[20] | 词权重问题），其中就参考了这篇文章，。这这篇文章后则能够，词权重被分为3个等级：

- 超重要。都是一些核心词，如欢乐谷之类的。
- 必要。行政区、品类词，如温泉、北京。
- 重要。一些不太关键的品类词、旅游关键词等，如门票。

- 非必要。一些口水词泛需求词，线路游、一张等。

这里面使用的方法稍微高级，使用的是监督学习的方式进行，涵盖如下特征：

- 统计特征，PMI、IDF等。
- 语言模型特征。query语言模型概率/去掉term后的语言模型概率。
- chunk识别结果。

模型则使用的是XGBoost，至于样本，这里没有详细说，只是说了人工标注，感觉上这个样本量基本w级别应该足够了，所以人工应该还是可以接受的（人工特征的泛化性不错，所以好效果可以很快出来）。

文本相关性

文本相关性这里指的不是语义的相关性，而是一个文字的匹配结果，这块一般用于粗排。我已经不止一次强调了在搜索里文本层面相关的重要性了，此处就不赘述啦。

TF-IDF计算相关性是比较经典的做法：

$$R_{Q,D}=\sum_{t\in Q}(\sum_{f\in Q}\frac{tf_{t,f}}{l_f}*w_f)*idf_t$$

Q是query中的词汇集，H表示t命中的文档集，$tf_{t,f}$词t在文档f中出现的次数，l_f是f的长度，w_f是文档权重，这个与文档性质有关，idf_t表示t的逆文档频率。

然而TF-IDF存在易受到文本长度影响、无法使用动态权重等问题，因此在BM25的基础上，引入动态权重（上面求的term重要度）后，使用了下面方法进行计算：

$$R_{Q,D}=\sum_{t\in Q}(max_{f\in Q}\frac{tf_{t,f}(k_1+1)}{tf_{t,f}+K}*w_f*i_f)*idf'_tK=k_1(1-b+b*\frac{l_f}{avgl_f})$$

k_1和b都是调节因子，这个含义与BM25中的类似，能够降低文本长度对相关性的影响，i_f是对应命中文本里面的文本动态权重，可以根据词在query中的占比和权重进行计算，最后的idf'_t则是query中的动态权重了，来自上面term重要度计算得到。

美团旅游搜索用的技巧

上面提到的一些常见的任务，这里还想要提的是内部所使用的的提升效果的小trick。

首先想说的是丢词，一般地我们都会把query中的所有词汇扔到搜索引擎（倒排索引）中进行检索，但实际上并不是所有词汇都需要这么做，一些类似“我想看”、“想去”、“一张”之类的词汇是完全没有意义的，还增加了这么多无效召回，因此这些词可以被忽略，从而提升效率，而丢词的标准，有两个：

- 词典&规则。
- term重要度。

多次召回。query千千万，总有一些query在当前策略下是无法召回所需结果的，所以我们可以放松检索条件，但是直接放松可能会带来过量的召回，增加下游粗排精排压力，因此策略就是进行再次召回。

迭代方向调研思路

无论是大系统的改进，还是小任务的提升，我们都要遵循的一个原则是发现问题-定位问题-解决问题，只有严格按照这个路径走，我们才能够达到最终的提升目标，我们来回顾美团整个分析的过程来从中吸收一些养分，这应该是我们每个算法都应该尝试具备的进阶能力。

虽说我们底层有大量的算法指标供我们参考，但最直接影响公司乃至我们的还是钱——收入，美团搜索需要衡量的是每搜索用户收入，通过提升用户的点击率、支付率、消费率等，才能有效达到提升收入的效果，说白了还是要增加用户在美团平台的任务达成量，也就是所谓的——满意度，满意度涉及的一方面是拥有的服务要保证好，另一方面则是对于没有的服务我们要尝试拓展，对于搜索，其实就要优化的就是无结果率了，保证用户尽可能可以搜到可靠的内容。

搜索领域比较好的一点是我们可以经过人工评判，可以快速发现明确的bad case，这个是相比推荐更舒服的一个点，所以搜索迭代方向调研的一个重要思路就是——bad case驱动，定期进行质量评估，根据评估结果分析问题，针对问题解决问题，然后经过实验迭代上线。我们来看看美团的迭代思路和方向是怎么演化的：

- 2015Q3：意图划分不明确导致用户需求无法被充分满足，尤其是poi的误召回过多，因此进行了意图划分等方式进行了优化，而针对无结果率，通过多次召回的方式进行了补充。
- 2015Q4：无结果率在进一步分析发现，POI的缺失（32%）、错误（27%）占比较多但算法不可解，而query表达多样性导致的误召回成为了主要原因（30%），因此query理解的优化成为降低无结果率的一大目标。
- 2016Q2：无结果率进一步分析发现，免费景区、不可网购这类型无法上线导致的无结果case占比最高（47%），但算法可解的只能是线上已有内容但是没有召回的才是算法可解的（29%），针对这个问题，又进行了新一轮的丢词、chunk分析优化，进一步提升效果。
- 2016Q3：扩召回的数量增加后，大量质量较差的结果也会被召回，此时比较精准的粗排显得更为重要，因此开始对粗排进行了新一轮优化，涵盖距离分、综合评分、新单照顾、因子组合方式等，尽可能考虑多方面需求，从而达到效果提升。

整个流程看来，搜索有了比较稳定的迭代提升，可见洞察系统问题再进行针对性解决，能最大限度降低优化风险，避免大起大落，“召回-排序”二元体系本身存在不稳定性，稍有不注意就会出现大开大合的情况，美团搜索多次优化召回也是步步为营、对症下药，这是我们需要吸收和学习的一个关键点。

小结

全文读完，无论是新入场的小白还是酒精沙场的老将，其实都会有不少的收获，从具体任务的使用方案到方向的分析探索都有很多养分可以吸收，我们也可以再次发现即使是大厂，也不是说非要高级前沿的方法才会去用，而是不拘一格地选择最快最优的方式来解决问题。

我是叉烧，欢迎关注我！

叉烧，OPPO搜索算法工程师，主做Query理解，NLP方向。19届北科技统计学硕士（保研），17届北京科技大学信息与计算科学、金融工程双学位毕业，论文7篇，学生一作3篇，参与国家级及以上学术会议4次，优秀论文一次，国奖金。曾任去哪儿网大住宿事业部产品数据，美团点评出行事业部算法工程师。

微信个人公众号
CS的陋室

微信 zgr950123
邮箱 chashaozgr@163.com
知乎 机智的叉烧

喜欢此内容的人还喜欢

属于算法的大数据工具-pyspark：10天吃掉那只pyspark
CS的陋室

不再拾人牙慧，来自己创造一个专家级别的神经网络吧！
CS的陋室

心法利器[13] | 任务方案思考：句子相似度和匹配
CS的陋室