曾经的知乎搜索为什么这么"烂"？来看看搜索技术有多复杂

全球创客会 2015-11-09

↑点击蓝字，轻松关注

昨天，搜狗、知乎公布了两家最近的战略合作，其中最主要的内容便是搜狗技术支持知乎，大大提升了后者的搜索体验。那么在这以前，知乎的搜索为什么这么烂呢？来听听搜狗搜索工程师怎么说。

图为知乎吉祥物"刘看山"。

特定站点搜索相比通用搜索的技术局限

搜索是技术方向辐射相当广的一个复杂系统，其技术门槛之高，在众多的互联网产品中能与搜索比肩的是少之又少。要想玩转这套系统，拥有一批最优秀且懂搜索的工程师和研究员是必不可少的。我看到之前 @熊辰炎同学也提到说想解决的话，知乎可能需要5个熟练工干大半年。在我看来，这种团队配置作为站内搜索差不多能解决大部分基础问题，即达到不被“到处”抱怨。但如果要求再高一点点，能稍”智能”地处理用户查询，那么这种团队配置恐怕还是望尘莫及。

当然搜索也绝不仅仅是一个人力问题，支撑搜索的人工智能技术正在”经验主义”(以统计学为代表)的道路上享受着大数据（特别是用户行为数据）的红利。从一个特定站点出发，即使是一个格调高、深受用户喜爱的站点，其能够接触到的数据无论是用户群体行为数据还是全网的信息资源都是十分有限的。**用户对于全网通用搜索和站内搜索的期望**

差别仅在于搜索范围从全网变为这个特定站点，但搜索用户天生的"懒惰"、表达含糊以及对搜索结果智能的期待从未改变过。而且由于用户对他所喜爱的站点的了解、熟悉程度远远超出其对全网的了解，所以用户对搜索服务所存在的各种问题更为敏感，从而也有更高的要求。正是这种数据局限所带来的技术水平局限与用户需求之间的矛盾，使得原生站内搜索注定就是一件不太可能成功的路。

扯远了，回到作为一个技术人员解释为什么知乎站内搜索没有通用搜索(例如百度、搜狗)的site查询好用吧。

通用搜索背后的技术

@张前川 关于搜索效果的评测解释得已经比较完善了，下面我就以这几个case为例解释一下通用搜索是如何解决背后的技术问题吧。主要分为NLP/相关性计算/排序这几个方面。

1. NLP

1.1 分词(Word Segmentation)

搜索中的分词是指将文本切成多个独立的语义单元以作为检索的最小单元，然后分词后的词串建立倒排索引以加快检索服务的速度。这是信息检索最基本最重要的架构，这里不详细展开。

先看看张前川提到的"避谷"这个case，正如张前川所说，避谷应该切成一个独立的词。为了解释后面的算法，我把case改成"避谷方法"，更容易说明问题，它的正确切词方法是【避谷】【方法】。如果把避谷分成【避】【谷】两个单字，就容易出现知乎站内搜索这种【避】【谷】两字分开出现的结果，也是我们常说的结果发生语义漂移。那么如何知道【避谷】应该是个独立的词呢？

最经典的分词方法有基于词典的前向/后向最大匹配或基于语言模型的分词等等，其中如何构建准确而全的词典，用什么语料统计适用的语言模型都是算法成功的关键所在。

问：通用搜索如何解决这个问题呢？

答：挖掘网络语料或用户行为数据！

a. 对于基于词典的方法，由于"避谷"是个道家的一个术语，有可能分词词典里不包含这个词。那么通用搜索通常可以通过挖掘网络语料（例如百科词条）来补充词典。

b. 对于语言模型或其他统计方法，用户群体历史的行为数据就是一种非常有价值的数据。这里仅提一个思路。历史上搜索"避谷方法"的用户，所点击结果的标题中"避"与"谷"很大概率彼此紧邻，"方"法"很大概率紧邻，而"谷"与"方法"很小概率紧邻。由此可以推断【避谷】【方法】应该相互连接组成一个词，而"避谷"与"方法"之间切分开来更合适。利用用户历史行为数据的方法还有很多，大家也可以打开思路。

1.2 查询纠错(Query Correction)

再看"什么名字haoting"这个case，非常直观，大家都能看出来是用户把查询词的一部分敲成拼音了，需要系统自动纠错。当然这是个简单的纠错，只要找到haoting对应的上下文语言模型概率最大的汉字"好听"即可纠正过来。

有些需要纠错的case就不那么容易了。例如"哦泡手机"，原意是找"oppo手机。"人脑能够非常快速准确的完成这一个纠错过程，但对于不具备智慧的机器，这个转换过程并不

那么容易。

针对这个case算法纠错的过程大致应该是这样：首先把“哦泡”转换成拼音“opao”，然后计算“opao”和“oppo”之间的编辑距离（一种度量文本串之间相似程度的方法），然后通过多种数据和模型计算出来“哦泡”纠错成“oppo”的概率，特别是在上下文为“手机”的条件下“哦泡”纠错成“oppo”的概率。这里面的每个步骤都同时需要算法与数据的支撑，通用搜索面对更多的数据和更更多的用户，显然有非常大的优势。

1.3 查询理解(Query Understanding)

查询理解这个概念比较广，广义上前面提到的分词、查询纠错也可以纳入查询理解的范畴，这里我们主要用查询理解来概括查询改写、词间紧密度、词赋权等一系列的对查询的理解以帮助获得更好的搜索结果。前川前面给出的“101大厦”就是一个比较综合的例子，但是这个case我有些不同看法。

首先“101大厦”合在一起表示一个完整语义的实体，所以相关的结果中101和大厦应该紧邻在一起。前川说应该分成一个词，但出于搜索查全率的考虑，即尽可能找到更多的相关结果，它们还是分开比较好，因为“101大厦”还有很多种其他的叫法，例如“台北101”“101大楼”等等。挖掘出101大厦的这些等价(或同义)说法对于搜索效果至关重要。这种等价或同义的算法用在搜索中就是查询改写一种最常见的形式。

但是“101”和“大楼”之间又存在非常紧密的关系，两者如果在文档中相距太远，结果通常是不相关的。这里涉及的是另一个概念——紧密度，即既需要切成两个独立的词，但又要求结果中这两个词之间的距离足够近，某些情况要求一定紧邻。

查询改写、紧密度同样依赖于网络资源的挖掘以及历史用户行为的挖掘，例如用户在同一个session内的主动改写、用户查询后的点击、具有相似点击结果的多个query等等...每种数据的合理应用，都能让搜索效果有所提升。通用搜索正是利用其数千亿网页索引库以及每日数亿次的用户查询及后续行为，在大数据上逐渐积累对查询理解的智慧。这些恐怕任何一个站点都无法触及的。

2. 相关性(Relevance)

前面提到的都是NLP相关内容，我们再来看看搜索里另一个核心技术—相关性计算。相关性计算通常指给定一个查询和一篇文档，计算两者是否语义相关。语义相关是个非常大的挑战，从技术的发展历程来看，从早期的统计词出现的频率，例如tf.idf、BM25、到language model、proximity等等都试图从查询词在文档中出现的次数、位置、词的权重、文档的长度等等多个角度去估计查询与文档之间的相关度。近来在深度学习的影响下，基于深层神经网络的词嵌入、语义表示、语义匹配等新兴技术的涌现，正在带领相关性计算由匹配统计迈入“语义计算”的大门。搜狗、百度已经在这这方面取得了阶段性的成功，同时这个方向还有很多问题待解决，让我们拭目以待吧。

就前川提到的“为什么要来北京”这个case，可以从多个角度解决。例如通过查询理解，我们可以知道“北京”在这个查询中是个非常重要的词，而标题包含重要的词的文档相比于仅正文包含重要词的文档中有更大概率与查询词先关。前川提到的第二条结果不相

关，”北京”即仅仅出现正文里。解决这个问题的思路还有很多，要想做个搜索，需要从多个维度去阐述查询与文档之间的关系，这是一项需要相当深积累的工作。

3. 排序（Ranking）

排序，望文生义即将搜索结果按照满足用户需求的程度从高到低排序，以便最满足用户需求的结果能够排在搜索结果列表的最前面，让用户能够最先浏览到。排序主要涉及两大问题：用于排序的多维特征以及多维特征的融合以决定最终的顺序。

相关性无疑是搜索排序的一类非常重要的纬度，我们前面也提到相关性自身也需要从多个更细纬度去剖析。正如很多用户提到的，知乎是问答社区，有人提问、有人回答、还有人点赞、关注，为什么知乎返回的结果很多都零回答、零关注。其实问题的回答数、关注数、点赞数都是衡量一个文档质量非常客观的指标，这些对于衡量问题是否能够满足用户需求都是非常有价值的，也就是说这些都应该成为排序所考虑的特征。

那么这么多特征相互如何融合来决定最终的顺序呢？有很多基于规则或线性融合的方法，近年来排序学习(Learning to Rank)的方法已经无数次在各种竞赛、学术论文、工业界产品中将排序多特征的融合的结果带入或逼近局部最优解或全局最优解。

无论是排序特征的准确与丰富还是排序融合，都是搜索工程师们孜孜不倦地不断优化的方向，经验与积累也是非常重要的。

4. 搜索架构

大部分用户会认为搜索效果和搜索性能没有什么关系，但实际上两者是紧密联系在一起的。由于服务负载的压力、用户响应时间的限制，分给每次用户查询的计算资源和时间是非常有限的。底层的检索的性能越好，所能查找的候选文档越多，所留给排序优化的时间越多，越能使用更丰富的特征和更复杂的算法，达到更好的排序效果。简而言之，性能越高，效果提升空间越大。

除了最基本的倒排索引，架构上还有很多可以优化的点。例如对历史数据的批量倒排和针对新数据或更新数据的实时倒排的设计，其次针对标题、正文等重要度不同字段的处理、倒排的压缩，快速交并算法、灵活的多机分环架构等等这些都是一个好的搜索架构需要考虑的问题。而好的架构的设计也是来源于对于搜索这个任务足够深刻的理解，如果没有对搜索多年的打磨，一名再优秀的架构师也是不可能设计出一套完美的搜索架构的。

啰嗦很多，总结一下，知乎搜索体验不理想，存在多种问题，但这些问题绝不是知乎仅有的问题，也不仅仅是人力投入的问题。搜索一个异常复杂的系统，好的搜索体验需要技术的沉淀与积累，需要海量数据特别是海量用户行为数据的支撑。站内搜索就于其在搜索方向的积累、其能接触到的数据，像知乎这样面对高标准严要求的用户，注定不易做到用户满意。

当然凡是问题，是都能够被解决的。

（文章来源：雷锋网）