如何预测用户query意图

发布于2012-4-18

有一个朋友问，一个用户搜索一个query是“百度”，怎么知道用户真正是想找什么呢。

我回答说，分析之前搜索这个query的用户点了些什么结果啊。

朋友继续问，如果没有用户点击呢。

呃，如果没有点击，这个问题就比较复杂了。整理了下思路，于是写成了本文。主要描述了关于如何预测用户query意图。希望会有所帮助。

首先我们的明确一个标准，如何判断我们对用户意图的猜测是正确的?

用户的思维是很发散的，也许今天搜索“葛优”，是想找“让子弹飞”，明天搜索相同的query，就是想找“非诚勿扰”。 我们确定了要在某个方面的query预测上做一个改进，那么我们首先的把标准定下来，依照这个标准来进行改进。

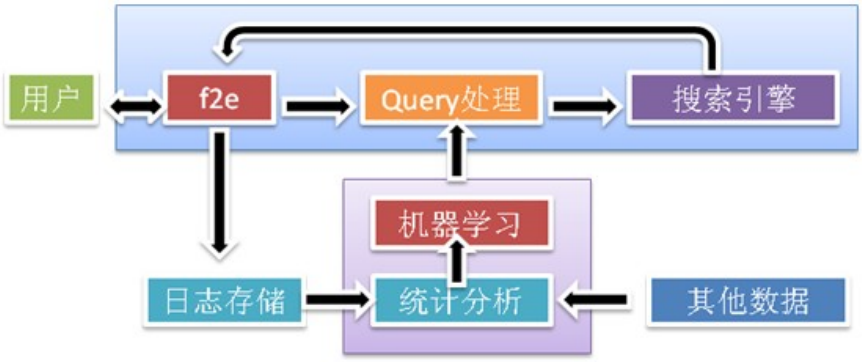
现在有很多对搜索系统的评价指标，如pv，ipv，ctr，搜索引导的后续转化率等可以量化的指标，这些指标是对搜索系统总体的评价。具体到用户意图预测上，标准很难确定，对于排序比较直观的就是进行side by side的评测，比较原有的效果和改进的效果，看是否会排序更优；对于导航，那我们可以看我们预测的类目和用户实际点的类目的占比，是否能有效降低用户点击非推荐类目的比率。

接下来，我们从2种情况下来回答这个问题，

如果我们已经有了一套完整的系统，有大量的用户访问

先从简单的说起，假设我们已经有了一个完整的搜索系统，有大量的用户访问，我们希望通过对用户query的预测来提高搜索体验。这样的系统的大概架构如下。

包括图所示的几个部分，



1 前端（f2e）

前端负责直接和用户进行交换，当收到用户搜索请求之后，往后端系统传递请求，并接收搜索引擎返回的结果，组织到网页上，展示给用户。

前端还肩负着一个重要的记录日志的工作，这个日志的记录，并不是apache的访问日志，这样的日志内容过于简单。如果要前端记录过多的日志，又会给服务器带来不小的压力。所以目前主要的手段是通过用户在页面上进行搜索或点击等行为时，调用javascript向指定的日志服务器，发送特征url来记录，这种url不会返回内容，仅仅为了给日志服务器添加记录。发送的url会包括从cookie中解析出的用户特有的数据。

2 Query处理

相关文章

从手机登录页面设计想到的
如何把无意识引入交互设计中
交互设计的真相
当视觉设计师遇上产品经理
手机交互设计原则
用户体验之网页板块设计
更多...

相关培训课程

以用户为中心的设计
可用性评估
Desktop及Web-based视觉设计
认知原理与设计应用
手机用户界面设计
更多课程...

成功案例

北京 以用户为中心的界面设计
北京 用户体验& 界面设计
上海 华为 用户体验& 界面设计
深圳 用户体验& 界面设计
爱立信 以用户为中心的设计
北京 用户体验与界面设计
福州 以用户为中心的界面设计
更多...

Query处理是线上服务系统，它是对用户意图进行预测后，对用户的搜索结果进行改进。在接收到前端的请求之后，会利用线下对query分析得到的数据，对用户的query和上下文环境进行分析，附加更多的条件到搜索引擎的请求命令之中。常见的Query处理，会有以下的一些类容，query改写，query分类预测，query的导航等。

Query处理这部分主要的意义在于，将用户的搜索query，翻译为对搜索引擎更适合查询串。在大多数情况下，用户使用搜索引擎是为了解决自己的问题，如果能直接获得答案，用户是不大愿意进行搜索的。

用户也许的问题是，“非诚勿扰2里面说的廖凡是谁”这样的一个问题，这样的问题直接搜索是不太会有会令用户比较满意的答案，（除非有向百度知道这样的系统已经存在了类似的问题）。有些用户就会考虑换个关键词试试，搜索下“廖凡”，看是否会有一些答案可以让自己满意。所以很大程度上是搜索引擎在教用户如何使用自己。但是并非所有的用户都对搜索系统如此的熟悉，那我们就需要考虑看看在我们搜索的结果里面效果不太好的query，分析它是怎么构成的。我们也许无法准确回答“非诚勿扰2里面说的廖凡是谁”，但是可以把其中最关键的信息抽取出来“非诚勿扰2”“廖凡”，并且，我们需要回答“是谁”这样的疑问问题。把这些信息传递给引擎，才会有更好的结果。

再例如，用户想找，“1000元左右的手机”，那么对于淘宝来说，可以把搜索的条件转化为800-1200价格限制范围的，手机类目下的宝贝，或者更进一步，把各种型号的手机，列在一起，进行参数的比较。

再深入一步，用户想找“舒淇在非诚勿扰2中用的手机”，如果我们可以把这个问题转化为对“朵唯S920”的搜索，那就是非常非常好的效果了，至于这个query如何对应到这个结果，也许后面的一些分析，能提供一些线索。

具体的实现，可以参考下面几点，

对query的线上处理，如果是较为hot的query，可以以查表为主，可以用hash表，trie树等进行查表，把在线下计算好的数据，通过查表的方式找到对应的结果，附加到给引擎的搜索条件上，并返回。

另外，可以把线下训练好的模型，在线上进行预测，一般的分类算法预测速度都比较快。可以对长尾的query，进行及时的预测。

也可以做一些规则，如我们上面举的例子，“1000元左右”，可以通过正则表达式进行识别，将其转为对应的搜索条件。这些规则如何来定呢，这是比较麻烦的一点，像这类的query，肯定是pv比较低的，属于长尾的query，这些query效果提升可能比较明显，但是对总体搜索系统效果影响会较小。这个问题比较尴尬，如果我们这类query处理的效果好的话，那用户会使用的更多；用户知道了这样的query效果不好，所以就换成了效果好的query。如果要做好规则，那就把长尾的这些query都拿出来，多看看，分下类，再结合实际的问题分类，总结出一些通用的规则，来进行优化。

3 搜索引擎

搜索引擎主要负责检索和排序，一般由一些倒排表和正排表组成。倒排表用于查找对应的文档id，能快速的检索出命中query的文档，在根据正排表来查对应id的数据。

一般将需要字符串类型的文档字段作为倒排表来进行检索，字符型的字段可以放在正排表中，在通过倒排表找到了满足条件的文档，再在正排表中进行过滤。

找到满足条件的文档后，再进行过滤，统计，并根据排序参数进行排序。

排序分为2个部分，一部分是文档自身的静态分，每个文档会有类似pagerank这样分数，另外一部分是还有和query相关的部分，会计算文档和query的关系，例如，query中出现的词的在文档中是否距离较近，query是否为文档的中心词。

4 日志存储

日志存储系统收集前端记录的日志，存储在数据仓库中，解析后用分布式文件系统来存放。有几类日志比较重要，

A、 搜索日志，搜索日志一般会包括以下一些信息，用户id，session id，用户搜索query，用户当前搜索的分类，用户搜索时间，

B、 点击日志，用户id，session id，用户搜索query，用户当前搜索的分类，用户点击的item，用户点击时间

C、 当然可能还有其他的如交易记录等，

有了以上几个部分之后，我们就可以通过以下2个部分来进行用户意图的预测，

5 统计分析

日志分析主要是一种统计分析，数据源来自于访问日志。另外还可以分析数据库中存储的用户的购买，收藏等行为。

可以从日志中分析出用户搜索query，“nike”最想找的是运动鞋呢，还是运动服。

常用的应用有下拉提示，相关搜索等，

下拉推荐是一种比较常用的用户意图分析的系统，通常是统计日志中，表现比较好的query，将这些query按照pv和数据表现等指标进行排序，然后把query转化为英文和中文对应的前缀，把相同前缀的建成统一索引，在用户输入关键词后，推荐相应的query。

相关搜索是更为常用的用户意图分析，一般通过关联规则(Apriori，FP-growth)，统计同一sesion中，用户经常出现的相关的query，比如，可以发现同一个session里面搜索了nike的用户，很多都搜索了“nike dunk”这样的信息，我们就可以再搜索结果中进行改进。这一算法可以大量应用于数据挖掘。推广开去，我们要找某个类目下进行了购买的用户，还希望购买些什么类目的东西；看了一本书的用户，还会看什么书；搜索了一个“长款”属性，是否还希望“修身”这样的属性。

在往下深入，我们可以分析用户历史行为，进行个性化的预测。比如分析用户性别，喜好，来进行分类，推荐。

6 机器学习

统计的算法也是机器学习的一种，如果用户行为数据足够多，那直接使用统计分析应该是可以解决大部分问题。剩下的小部分问题是可以交给机器学习其他算法来完成。

举一个简单的例子来说明，用户搜索“nike”和“羽绒服”比较多，有了足够多的统计数据，我们知道“nike”对应的是运动鞋，运动服等。“羽绒服”对应的是服装。但是用户搜索“红色的nike羽绒服”次数很少，没有足够多的数据，我们统计到的结果也许是不准确的，偏差较大。

那我们可以将较好的数据进行训练，并对长尾的query进行分类预测。这里的训练数据的特征是用户query中每个词，词出现对应这一种分类。

训练数据的选择是分类算法最重要的一步，一般对文本的分类预测，可以使用信息增益，卡方，互信息等来作为训练特征。具体问题具体分析，例如使用loglinear算法进行预测，实验证明信息增益来作为特征选择会更加有效，另外也得分析应用的场景，根据需要来选择算法，选择特征，法无定法，对于淘宝的数据来说，用于搜索的限于宝贝的标题，非常的短，直接使用用一般的网页分类算法是不太可行的，所以，数据不一样，方法就不一样，重要的是了解数据，了解方法的原理。机器学习不是万能的，不能靠运气。

分好类后，对每个类中的文档的排序也可以通过机器学习来进行，如果每个文档有很多标准的特征，每个维度的特征有一定的分数。这个也可以通过机器学习的方法来进行好中坏分档，或者找出线性加权的最优化参数。

假设我们没有用户反馈数据

我们首先可以做的是把文档的自身的相关性做好，回到最开始的那个问题，一个用户搜索一个query是“百度”，怎么知道用户真正是想找什么呢。

先我们至少可以把文档按分词后的结果和query进行比较，文档中如果是“众里寻他千百度”这样的就可以过滤掉了，因为“千百度”和“百度”还是有一些区别的。这是从文档自有的相关性上来进行优化。

接下来，我们看这个文档是不是描述文档的，比如文档里面是讲“非诚勿扰2”的，里面提到“廖凡，如果你不知道廖凡是谁，请百度一下”，那么这种文档不是描述“百度”这个词的，而是描述“非诚勿扰2的”，我们可以通过给文档进行分类或者加上tag，来表示他的主题词，这样，这类的文档也可以过滤掉。

我们再讨论下如果进行分类，在有用户数据的时候，我们可以用用户的行为来作为文档分类的结果；没有的情况下，我们可以进行人为的标注，当然这部分工作量巨大。另外可能可行的是，在结构化比较好的数据里面，找到关键的字段进行分类，例如，品牌+产品型号，这样的字段作为聚类的关键key，把文档分为很多类。如果结构化不是很好，可以考虑用crf算法来抽取其中的关键字段进行聚类。同时把query对文档的直接搜索转化为对不同类文档的搜索。那么这时候，我们已经把搜索的所有文档进行了聚类，发现“朵唯S920”手机的描述中，常会出现“舒淇在非诚勿扰2中使用”这样的描述，是否就可以考虑把两者联系在一起了呢。

相关文章

产品UED流程及交付物

可用性测试

浅谈如何留住用户

QQ音乐2012设计总结