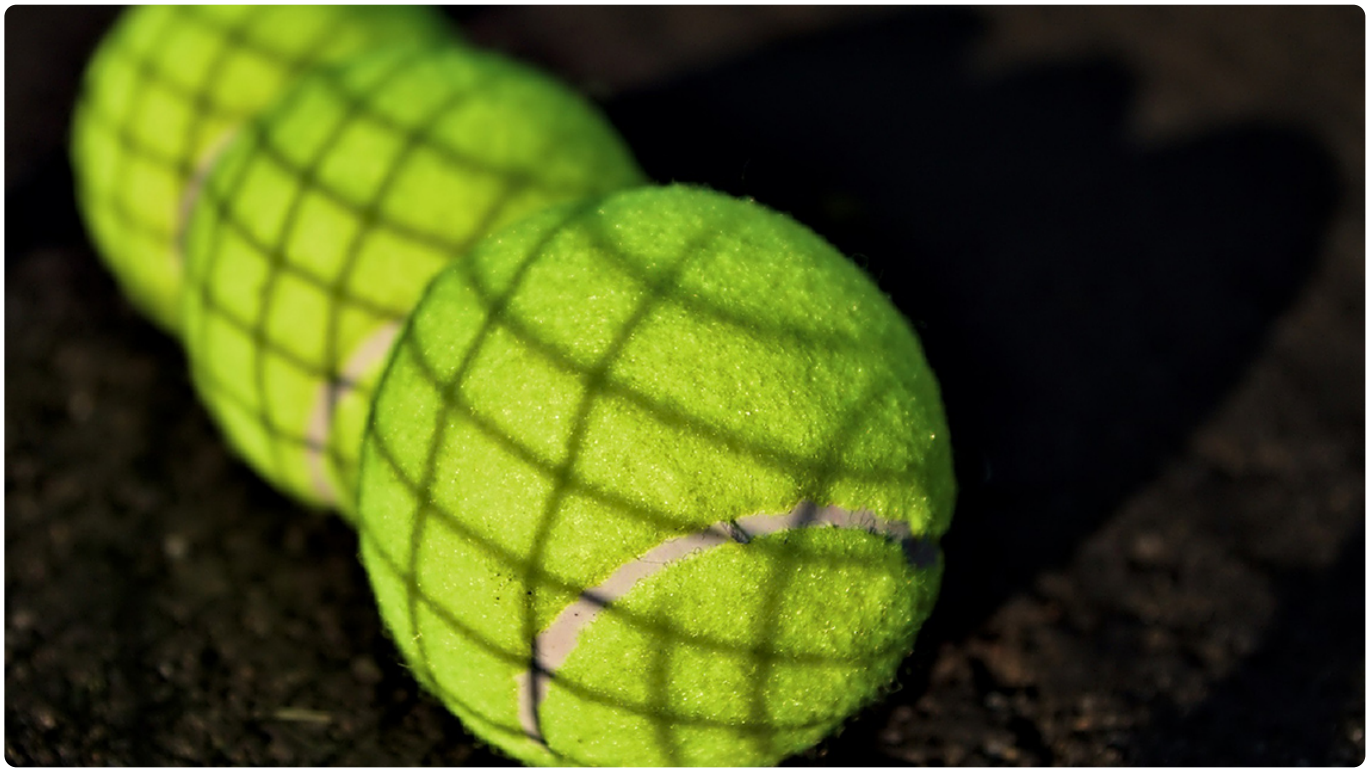
024 | “查询关键字理解”三部曲之分类

2017-11-27 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 09:07 大小 4.18M

我们在前两周的专栏里主要讲解了最经典的信息检索（Information Retrieval）技术和基于机器学习的排序算法（Learning to Rank）。

经典的信息检索技术为 2000 年之前的搜索引擎提供了基本的算法支持。从中衍生出的 TF-IDF、BM25 还有语言模型（Language Model）以及这些方法的各类变种都还在很多领域（不限文本）里继续发挥着作用。

另一方面，排序学习算法引领了 2000 年到 2010 年各类基于机器学习的搜索算法的产生和发展，也带来了搜索引擎技术的进一步成熟。

这周我们从排序算法转移到排序问题中一个非常重要的部分：查询关键字理解（Query Understanding）。也就是说，我们希望通过查询关键字来了解用户种种行为背后的目

的。查询关键字产生的特征（Feature）往往是很强的指导因素，也是个性化搜索结果非常重要的源泉。因此，深入了解并掌握查询关键字理解方面的技术就变得很有必要。

查询关键字理解最基本的一个步骤就是给查询关键字分类（Classification），看这些查询关键字有什么用户意图（Intent）。今天我就来聊一聊查询关键字分类的一些基本概念和技术，让你对这方面的开发和研究有一个基本认识。

查询关键字分类的历史

从商业搜索引擎开始面世的第一天起，人们就发现，可以从查询关键字中得到很多用户的信息，特别是理解用户的意图。早在 1997 年，商业搜索引擎 Excite 就开始了百万级别查询关键字的研究工作。然而，真正对查询关键字分类进行系统阐述的是安德烈·布罗德（Andrei Broder）的论文《网页搜索分类》（A Taxonomy of Web Search）。

安德烈很有名头，在斯坦福大学攻读博士期间师从图灵奖得主高德纳（Donald Knuth），然后在曾经名噪一时的第一代搜索引擎公司 AltaVista（后被雅虎收购）担任首席科学家，之后加入位于纽约的 IBM 研究院组建企业级搜索平台，2012 年后加入 Google，担任杰出科学家（Distinguished Scientist）。他还是 ACM（Association of Computing Machinery，计算机协会）和 IEEE（Institute of Electrical and Electronics Engineers，电气电子工程师学会）的双料院士。

安德烈的这篇论文可以说是奠定了查询关键字分类的坚实基础。这之后研究人员的很多工作都是围绕着如何自动化分类、如何定义更加精细的用户意图来展开的。

查询关键字分类详解

我就从安德烈这篇非常有名的文章说起。在网络搜索（Web Search）成为比较主流的咨询查询手段之前，传统的信息检索认为，查询的主要目的是完成一个抽象的“信息需求”（Information Needs）。在传统信息检索的世界里，最主要的应用应该是图书馆检索或者政府学校等企事业单位的检索。因此，在这样的场景下，假定每一个查询主要是满足某个“信息需求”就显得很有道理了。

然而，早在 2002 年，安德烈就认为这样的传统假定已经不适合网络时代了。他开始把查询关键字所代表的目的划分为三个大类：

1. 导航目的（Navigational）；

2. 信息目的（Informational）；
3. 交易目的（Transactional）。

此后十多年里，查询关键字的这三大分类都是这个方向研究和实践的基石。我们先来看这个分类的内涵。

第一类，以导航为意图的查询关键字，这类查询关键字的目标是达到某个网站。这有可能是用户以前访问过这个网站，或者是用户假设有这么一个关于所提交查询关键字的网站。这一类查询关键字包括公司的名字（如“微软”）、人的名字（如“奥巴马”）或者某个服务的名字（如“联邦快递”）等。

此类查询关键字的一个重要特点就是，在大多数情况下，这些查询关键字都对应唯一的或者很少的“标准答案”网站。比如，搜索“微软公司”，希望能够找到的就是微软公司的官方网站。另一方面是说，某些“信息集成”网站也是可以接受的“答案”。比如，查询“奥巴马”，搜索返回的结果是一个列举了所有美国总统的网站。

第二类，以信息为意图的查询关键字，这类查询关键字的目标是搜集信息。这一类的查询和传统的信息检索非常接近。值得提及的是，从后面的研究结论来看，这一类查询关键字所包含的目标不仅仅是寻找到某类权威性质（Authority）的网页，还包括列举权威信息的俗称“结点”（Hub）的网站。

第三类，以交易为意图的查询关键字，这类查询关键字的目标是到达一个中间站点从而进一步完成“交易”（Transaction）。这一类查询关键字的主要对象就是“购物”。现在我们对“电子商务”的态度可以说是非常自然了，但是十多年前，在传统信息检索界统治的搜索研究领域，提出“交易”类型的查询关键字可以说是很有新意的。

当然，这样的分类如果仅仅是概念上的区分那就没有太大的意义。安德烈利用搜索引擎 AltaVista 进行了一次调查研究，这次调查有大约 3 千多的用户反馈。想到这是在 2001 年的调查，可以说已经是大规模的研究了。

这次调研的结果是这样的：在用户提交的信息中，导航类型的查询关键字占 26%，交易类型的查询关键字占到了 24%，而剩下的将近 50% 是信息类型的查询关键字，用户的日志（Log）分析进一步证实了这一数据。

你可以看到，**这种把查询关键字进行分类的研究是对用户行为进行建模的必要步骤。**于是，很快就有不少研究人员嗅到了查询关键字分类的价值。然而，完全依靠用户直接反馈来获取这类信息则变得越发困难。

这里主要有三个原因。第一，不可能寄希望于用户汇报自己所有关键字的意图；第二，面对亿万用户输入的查询关键字，手工标注也是不可能的；最后，安德烈的三类分类还是太粗犷了，在实际应用中希望得到更加细颗粒度的用户意图。

把查询关键字分类问题转换成为标准的机器学习任务其实很直观。确切地说，这里需要做的是**把查询关键字分类转换成为监督学习任务。**这里，每一个查询关键字，就是一个数据样本，而响应变量，则是对应的类别。具体情况取决于我们的任务是仅仅把查询关键字分为几个类别，并且认为这些类别之间是互相独立的，还是认为这些类别是可以同时存在的。

在最简单的假设下，查询关键字分类就是一个普通的**多类分类问题**，可以使用普适的多类分类器，比如支持向量机（SVM）、随机森林（Random Forest）以及神经网络（Neural Networks）等来解决这类问题。

对于绝大多数监督学习任务而言，最重要的一个组成部分就是选取特征。随后很多年的研究开发工作中，有一部分就集中在尝试使用不同的特征，然后来看对提高分类的精度是否有效果。

过去的研究反复证明，以下几类特征非常有效。

第一类特征就是查询关键字本身的信息。比如，查询关键字中已经包括了已知的人名或者公司名，这种时候，分类结果就不太可能是交易意图的类别。也就是说，查询关键字，特别是某些词或者词组和类别有某种关联信息，而这种关联很大程度上能被直接反映出来。

第二类特征是搜索引擎返回的查询关键字相关的页面本身的信息。你可以想象一下，假如搜索“奥巴马”这个关键字，返回的页面都是维基百科的页面以及奥巴马基金会的页面，那么这些页面上面的内容可能很难包含任何商业的购买信息。而对于“佳能相机”这个查询关键字而言，返回的页面很可能都是电子商务网站的商品信息，从而能够更加准确地判断“佳能相机”的分类。

第三类特征则是用户的行为信息，那就是用户在输入查询关键字以后会点击什么网站，会在哪些网站停留。一般来说，哪些网站点击率高、停留时间长，就表明这些网站在返回结果中

可能更相关。于是，采用这些网站来作为查询关键字所代表的内容，就可能更加靠谱。

在实际的应用中，查询关键字的分类往往还是有很大难度的。因为在普通的现代搜索引擎上，每天可能有三分之一、甚至更多的关键字是之前没有出现过的。因此，如何处理从来没有出现过的关键字、如何处理长尾中的低频关键字，就成了让搜索结果的精度再上一个台阶的重要因素。我今天就不展开相应的话题了，如果你有兴趣，可以查看相关论文。

小结

今天我为你讲了现代搜索技术中一个非常基础但是也在实际应用中至关重要的环节，那就是查询关键字理解中的用户意图分类问题。你可以看到**查询关键字从大类上分为信息意图、交易意图以及导航意图三类。**

一起来回顾下要点：第一，简要介绍了查询关键字分类提出的历史背景，安德烈·布罗德的论文奠定了查询关键字分类的坚实基础。第二，详细介绍了主要的分类以及如何通过多类分类器的构建来达到自动化的目的。

最后，给你留一个思考题，在机器学习排序算法中，我们应该如何使用查询关键字分类的结果呢？

欢迎你给我留言，和我一起讨论。

拓展阅读：A taxonomy of web search

上一篇　023 | 机器学习排序算法：列表法排序学习

下一篇　025 | “查询关键字理解”三部曲之解析

精选留言 (2)

写留言

罗马工匠
2017-12-07

低频关键字的论文能列几篇么?

展开

作者回复: 不太明白你需要哪方面的?

张岩kris
2017-11-30

符合分类的搜索结果赋予更大的排序权重吧?

展开

作者回复: 这是一种思路。