025 | “查询关键字理解”三部曲之解析

2017-11-29 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 07:56 大小 3.63M

这周我分享的核心内容是查询关键字理解（Query Understanding）。周一介绍了查询关键字分类（Query Classification）的基本概念和思想。今天，我来讲一个更加精细的查询关键字理解模块：**查询关键字解析**（Parsing）。

如果说查询关键字分类是对查询关键字的宏观把握，那么，对查询关键字的解析就是微观分析。其实，查询关键字解析是一类技术的统称，我今天就来聊几个比较热的话题。

查询关键字分割

首先，让我们设想这么一个场景，在英文的搜索引擎中，如果一个用户输入的是“White House Opening”这个查询关键字，这个用户的意图（Intent）是什么呢？要想理解用户的意图，我们就得知道用户输入的单词的涵义。

那么，在上面这个查询关键字里，我们到底是分别理解每一个单词“White”、“House”和“Opening”呢，还是“White House”和“Opening”呢，还是有可能“White House Opening”是一个整体呢？这里说的其实就是“**查询关键字分割**”（Query Segmentation）这个概念。

在刚才的例子中，如何把“White House Opening”进行分割直接关系到搜索结果的质量。试想在一个比较标准的现代搜索引擎里，一般来说，都会有一个模块根据查询关键字来提取“**倒排索引**”（Inverted Index）中的文档。这个阶段的提取数目一般是几百到几千，这个过程常常被称为“**检索流程**”（Retrieval Phase）。

当有了这些文档以后，现代搜索引擎会利用比较复杂的排序算法，通常就是我们之前提到过的基于机器学习的排序学习模型，来对文档进行重新排序（Re-Rank）。

你可以看到，在这样两个阶段的流程里，如果好的文档没有在第一个阶段被提取出来，不管第二个阶段的功能有多强大，搜索的整体结果都不可能有多好。而对于“检索流程”而言，在“倒排索引”中进行查询的关键就是使用什么“单词”或者“词组”进行查找。

用刚才的例子来说，就是看文档究竟是符合“White House”，还是“White 或 House”，还是“White House Opening”。很明显，这三种情况得到的文档集合是不尽相同的。如果用户的真实意图是搜索美国总统府白宫的开放时间，那么把这个搜索关键字给分割成“White 或 House”，很明显就会影响提取的文档集合。

那究竟该怎样做查询关键字分割呢？

这里我介绍一篇论文《重新审视查询关键字分割》（Query Segmentation Revisited）。在这篇论文里，作者们集中介绍了一些主流的“查询关键字分割”技术，文章非常值得精读。下面我为你归纳一下要点。

第一种技术就是尝试从查询关键字里面产生“N 元语法”（N-Grams）。所谓 N 元语法其实就是从一组词语中产生连续的子词语。比如刚才的“White House Opening”的例子，我们就可以从这个词组里面产生“White House”和“House Opening”两个二元语法。

而第一种基于 N 元语法的方法，就是通过这些 N 元语法在一个大语料中出现的词频来判断这个“分割”是否有意义。当然，直接采用词频可能会比较偏好短的单词，所以在论文中，

作者们分别介绍了两种矫正词频的方法。

一种是基于词频本身的矫正，一种是基于维基百科，作为一个外部资源的矫正方式。两种方法的目的都是为了让长短语的打分（Scoring）有机会高于短的单词。文章中所需要的词频采用了谷歌 2005 年发布的“N 元语法”语料，也就是说，所有单词出现的频率都是直接在这个语料中获得的。

第二种技术是基于短语“互信息”（Mutual Information）的方法。“互信息”计算了两个随机事件的相关程度。在这里，就是计算查询关键字中每两个相邻短语的“互信息”。当这个“互信息”的取值大于某一个预设阈值的时候，我们就认为相邻的两个单词组成了短语。“互信息”的计算需要知道某个单词出现的概率，这些概率是从微软发布的一个“N 元语法”语料获得的。

第三种技术则是基于“条件随机场”（Conditional Random Field）。“条件随机场”是机器学习著名学者乔治·拉菲迪（John D. Lafferty）、安德鲁·麦卡伦（Andrew McCallum）和费尔南多·佩雷拉（Fernando Pereira）在 2001 年发表的“序列学习”模型（Sequence Model）中提出的。条件随机场的基本思想是对输出的复杂标签进行建模，尝试从特征空间建立到复杂标签的一个对应关系。

在“查询关键字分割”的场景下，我们其实可以把复杂标签看作是从一个查询关键字到多个短语的多个二元决策问题。这里的二元决策是指某一个备选短语是否可以作为分割的短语。条件随机场可以比较直观地对这类问题进行建模，而传统的二分分类器则很难对序列信息进行建模。我在这里就不详细展开条件随机场的介绍了，有兴趣的话可以翻看相关的论文。

查询关键字标注

刚才我聊了查询关键字理解最基本的“分割“问题。可以说，“分割问题”是查询关键字理解的第一步。那么，下一步则是更细致地分析查询关键字。

回到刚才的例子“White House Opening”，我们其实不仅是想知道这个查询关键字可以分割为“White House”和“Opening”，而且希望知道“White House”是一个建筑物的名字或者一个地理位置的名字，而“Opening”则可能是一个名词，暗指“开门时间”。也就是说，我们希望为查询关键字中的词组进行“标注”（Annotation），来获取其“属性”（Attribute）信息。希望为查询关键字中分割出来的词组进行标注的组件就叫做“**查询关键字标注**”。

那么，标注信息又是怎样帮助搜索结果的呢？试想一下“苹果价格”这个查询关键字。这取决于用户搜索的场景，如果“苹果”代表“水果”这个属性，那么这个查询的结果是希望找到水果的价格，可能还需要搜索引擎返回附近超市的一些信息。但如果“苹果”其实代表的是“手机”，那这个查询的结果也许最好是返回苹果公司的官方销售网站。你看，“苹果”所代表的属性不同，最优的返回结果可能会有非常大的差别。

对查询关键字进行标注的方法也有很多。我这里再推荐一篇经典的论文《使用伪相关反馈针对搜索查询关键字进行结构化标注》（Structural annotation of search queries using pseudo-relevance feedback），这篇论文**利用一个叫做 PRF（Pseudo-Relevance Feedback）的方法来进行标注**。这里面的一个技术难点是，查询关键字的信息实在是太少，需要利用大量的辅助信息来进行标注，因此 PRF 作为一个技术在这里得到了应用。

另外一个主流的查询关键字标注的方法，依然是利用条件随机场。我前面讲了，条件随机场是很好的序列建模工具。那么，在这里，以“苹果价格”为例，条件随机场是需要预测标签是否是“手机名词”还是“水果名词”这样的组合输出结果。而传统的二分或者多类分类器很难捕捉到这里的序列信息，条件随机场就是解决这方面的利器。

于是，我们需要做的就是为查询关键字构建特征（Feature），然后直接放入条件随机场中。有一点需要注意，条件随机场的应用成功与否与数据的多少有很大关系。因此，**构建一个有标注信息的数据集就变成了查询关键字标注的一个核心挑战。**

小结

今天我为你讲了现代搜索技术中的一个重要环节，那就是查询关键字理解中的查询关键字解析问题。你可以看到查询关键字解析从大类上分为查询关键字分割和查询关键字标注两个比较重要的模块。

一起来回顾下要点：第一，简要介绍了查询关键字分割的场景和三种主要技术，分别是“N元语法”、“互信息”和“条件随机场”。第二，详细介绍了查询关键字标注的场景和主要技术，包括利用 PRF 和利用条件随机场两种主流的标注方法。

最后，给你留一个思考题，我举了英语的查询关键字的解析问题，那么对于中文而言，又有哪些特殊的挑战呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. Query segmentation revisited. Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 97-106. 2011.
2. Michael Bendersky, W. Bruce Croft, and David A. Smith. Structural annotation of search queries using pseudo-relevance feedback. Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, New York, NY, USA, 1537-1540. 2010.

上一篇　024 | “查询关键字理解”三部曲之分类

下一篇　026 | “查询关键字理解”三部曲之扩展

精选留言 (3)

写留言

金晓烨
2018-10-16

1

有关思考题, 中文的处理相对英语主要是分词方面复杂度会高很多

展开

沛沛
2018-06-01

您好，想问下关键词解析用rnn如何

展开

颛顼
2017-11-30

想请教下对于查询关键字想去对他进行聚类，然后分析，最好能产生一个结构化的意图树出来，这有什么方法嘛？或者有什么类似的论文参考嘛？

> 作者回复: 建议参考Query to Knowledge: Unsupervised Entity Extraction from Shopping Queries using Adaptor Grammars。不完全一样，但是是一个参考。