026 | “查询关键字理解”三部曲之扩展

2017-12-01 洪亮劼

AI技术内参

进入课程 >

讲述：初明明
时长 07:29　大小 3.43M

我们在本周的前两篇文章中分别介绍了查询关键字分类（Query Classification）和查询关键字解析（Query Parsing）的基本概念和思想。今天，我来讲一个稍微有一些不同的查询关键字理解模块：**查询关键字扩展**（Query Expansion）。

查询关键字扩展想要解决的问题和分类以及解析略微不同。其主要目的不仅仅是希望能够对用户输入的关键字进行理解，还希望能够补充用户输入的信息，从而达到丰富查询结果的效果。

查询关键字扩展的概念

为什么要提供查询关键字扩展？主要原因还是用户输入的查询关键字信息不足。还记得我们上次提到的“苹果价格”这个例子吗？在这个例子中，用户到底是希望查询“苹果”作为一

种水果的价格，还是“苹果”作为手机的价格，其实无法真正从这个查询关键字中得出。因此，作为搜索引擎，如果为用户提供一些“扩展选项”，也就是一个被改写（Reformulated）过的查询关键字，会提供更加好的用户体验和更加精准的搜索结果。

查询关键字扩展除了显示出来能够让用户有更好的体验以外，还有一个作用是增加文档的“召回”（Recall），从而为提高搜索结果奠定基础。设想这样一个例子，用户搜索“iphone 6 backup”，希望了解如何备份 iPhone 6 的信息。因为苹果手机的绝大多数机型的备份流程都大同小异，因此，如果把“iphone 6”给扩展到“iphone”其他机型，然后看是否有比较好的介绍备份的网页可以显示。

值得注意的是，在扩展的过程中也有可能失去“精度”（Precision）。比如假设苹果对 iPhone 7 的备份流程做了很大的改进，那么其他机型的流程也许就不适用了，所以当用户搜索“iphone 7 backup”的时候，如果我们扩展到了其他机型，那让用户看到的很可能就是不那么相关的信息了。因此，**对“精度”和“召回”的平衡，成了查询关键字扩展的一个重要的权衡点。**

查询关键字扩展的另外一个重要应用就是对同义词和缩写的处理。比如，唐纳德·特朗普（Donald Trump）是美国现任总统。那么，如果用户在搜索“Donald Trump”、“Trump”、“US President”、“POTUS”（这是“President Of The United States”的简称）等类似词汇的时候，搜索引擎应该提供相似的结果。而从词汇的直接联系上，这些词汇在表面形式上可能有很大的差异（比如“Trump”和“POTUS”），因此需要其他手段学习到这些词语内涵的同义。

查询关键字扩展的技术

知道了查询关键字扩展的含义以后，我们就来看看有哪些技术可以为查询关键字扩展提供支持。

根据上面提供的一些例子，你可以看到，这里的**核心就是找到搜索结果意义上的“同义词”**。那么，在搜索中，如何挖掘“同义词”呢？

今天我在这里分享两种思路。

第一种思路，是根据查询关键字和查询结果之间的自然结合产生的同义效果。这需要对用户的搜索行为数据进行大规模挖掘。这里的基本假设是这样的，假设我们有两个搜索关键字，A 和 B。从 A 的搜索结果中，用户可能点击了一些网页，从 B 的结果中，用户可能点击了

另外的一些网页。如果这些被点击的网页恰好非常类似，那么，我们就可以认为 A 和 B 其实是同义的查询关键字。

更加完整的做法是把查询关键字和网页分别表示成“图”（Graph）中的两类节点（Node）。每个关键字节点和多个网页节点建立联系（Edge）或者边（Link），象征这些网页对应这个关键字的相关页面。而从每个网页的角度上看，多个关键字节点又和同一个网页节点相连，表示这些关键字都有可能和某个网页相关。

拿上面提到的特朗普的例子来说，美国白宫的首页作为一个节点的话，就有可能会和“Trump”、“US President”以及“POTUS”这几个查询关键字相关。因此你可以看到，寻找同义词的工作就变成了如何在这个图上进行相似节点，特别是相似关键字节点的挖掘工作。

如果把查询关键字的节点放在一边，把网页节点放在一边，我们就看到了典型的“**二分图**”（Bipartite Graph）。二分图的特点是同边的节点之间没有连接（比如关键字和关键字之间没有连接），而所有的连接都发生在不同边的节点之间（关键字和网页之间）。

二分图的聚类问题（Clustering）是机器学习界的经典的问题。而利用二分图的聚类问题来做查询关键字的同义词挖掘也是很多研究人员尝试的方向。文末我列了几个参考文献，比如参考文献 [2] 就是利用二分图上的“随机游走”（Random Walk）以及随机游走所产生的“到达时间”（Hitting Time）来挖掘出类似的关键字。如果你有兴趣，可以查看这篇经典论文。

说了基于用户行为信息和关键字挖掘的思路以后，我们再来看看第二种思路。

第二种思路的核心是从海量的文本信息中分析出词语之间的相关度。这里面需要注意的是，这些词语的相关度有可能是语言本身带来的。比如，单词“Male”和“Man”。也可能是语境带来的，比如谈论手机的网页中对于“iPhone 6”和“iPhone 7”的谈论。

总之，这一个思路的想法就是如何为每一个词组都建一个“表达”（Representation），从而通过这个表达找到同义词。近年来流行的一个做法是为单词找到数值表达，也就是通常所说的“嵌入”（Embedding）。如果两个词在“嵌入空间”（Embedding Space），通常是“欧式空间”中距离相近，那么我们就可以认为这两个词是同义词。

如何为词组产生“嵌入”向量呢？这里面也有很多做法。比较通用的有算法 Word2Vec（参考文献 [3]），目标是通过一个文档的每一句话中某一个词周围的词来预测这个词出现的概率。可以设想一下，在苹果手机的很多帮助文档或者帮助站点中，描述如何帮助 iPhone 6 或者 iPhone 7 来做数据备份的字句都是相似的，甚至，可能唯一的区别就是描述机型的名字。

因此在这样的情况下，通过文字周围的“上下文信息”（Contextual）来对单词本身的“嵌入向量”进行学习可以有效地学习到单词的语义。而通过语义，我们就能够找到其他的同义词。当然，要想真正应用到查询关键字扩展中，可能还需要有其他的调试，比如文末我列的参考文献 [4]，就是其中的一种。如果你感兴趣，建议去精读。

最后我需要说明的是，第一种思路需要已经有不少的用户交互数据，而第二种思路可以通过其他的语料（比如维基百科）加以学习，并不需要用户数据。这也是另一个值得参考的信息点。

小结

今天我为你讲了查询关键字理解中的查询关键字扩展问题。你可以看到，查询关键字扩展从技术上的两种流派，一个是通过用户的交互数据来产生一个图，并且利用图挖掘技术来得到查询关键字之间的关系；另外一个就是通过产生词汇的嵌入向量从而得到同义词。

一起来回顾下要点：第一，简要介绍了查询关键字扩展的内涵。由于用户输入的查询关键字信息不足，通过查询关键字扩展可以提供更好的用户体验和更加精准的搜索结果。第二，详细介绍了查询关键字扩展的两个主要技术。

最后，给你留一个思考题，如何来测试查询关键字扩展的优劣呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Claudio Carpineto and Giovanni Romano. A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys. 44, 1, Article 1 (January 2012), 50 pages.2012.
2. Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. Query suggestion using hitting time. Proceedings of the 17th ACM conference on Information and

knowledge management (CIKM '08). ACM, New York, NY, USA, 469-478. 2008.
3. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
4. Diaz, Fernando, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891 (2016).

上一篇　025 | “查询关键字理解”三部曲之解析

下一篇　027 | 搜索系统评测，有哪些基础指标?

精选留言 (2)

写留言

极客星星
2017-12-03

2

洪老师你好 刚好最近在做查询扩展的工作 您的文章真是雪中送炭。有两个问题想咨询下 1扩展后的词索引得到的结果 和原有词的结果如何放在一起排序?2我想用第一种方法来实现扩展 这种方案除了文中 提到的二分图方法 还有没有其他方法呢 譬如推荐系统中的协同过滤的方法是不是也可以用在这里.谢谢

作者回复: 协同或者也有用Matrix Factorization的。

老敖

2017-12-04

加个人工互动反馈的环节？看看是否扩展出来的结果是否满意，有点类似于推荐系统那种。

作者回复: 人工互动是一种思路，但是可能没法大规模化。