100 | 文本情感分析中如何做意见总结和搜索?

2018-05-23 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:28 大小 2.97M

在文本“情感分析”（Sentiment Analysis）这个领域，我们首先介绍了最基础的文档情感分类这个问题。在绝大多数情况下，这是一个监督学习的问题。当然，我们也可以通过建立情感词库来进行简单的非监督学习的步骤。紧接着，我们讨论了文本情感分析中的另一个关键技术，即情感“实体”和“方面”的提取。这个任务可以说是很多情感分析的根基，我们需要从无结构的文本中提取实体和方面等结构信息，便于进一步的分析。我们讲了如何通过词频、挖掘配对信息以及利用监督学习来对实体和方面进行挖掘。

今天，我们来看文本情感分析的另外一个主题——**意见总结**（Opinion Summarization）和**意见搜索**（Opinion Search）。

意见总结

为什么“意见总结”这个任务会很重要的呢？

假如你希望在电商网站上购买一款数码相机。这个时候，你可能需要打开好几款相机的页面进行比较。对于相机的硬件指标，能够从这些页面上相对容易地直接得到，除此以外，你可能还比较关心对这些相机的评价。

在这个场景下，“意见总结”的重要性就凸显出来了。因为优秀的相机款式往往有上百甚至上千的用户评价，这些评价包括了用户对产品很多方面的评价，有褒义和贬义的情绪。如果对这些评价逐一进行浏览，很明显是一种非常低效的做法。因此，从购物网站的角度来说，如果能够为用户把这些评论进行总结，从而让用户看到总体的有代表性的评论，无疑能够帮助用户节省不少时间和精力，让用户获得更好的体验。

简单来说，意见总结就是从无结构的文本中提取出来的各种情感信息的综合表达。我们这里聊的意见总结主要是指“**基于方面的意见总结**”（Aspect-based Opinion Summarization）。也就是说，意见的总结主要是围绕着产品的种种方面来产生的。

概括一下，基于方面的意见总结有两个特点。第一，这样的总结主要是针对物体的**实体**以及对应的**方面**来进行的。第二，意见总结需要提供**数量化的总结**。什么是数量化的总结？就是总结里需要指出，持有某种意见的用户占多少百分比，又有多少百分比的用户有其他意见。很明显，这里还牵涉到如何表达和显示这些意见总结的步骤。

可以说，基于方面的意见总结成为了意见总结的主要任务。另外，基于方面的意见总结还可以**与其他文本技术相结合**，从而能够延展这个技术的效果。比如，总结语句的生成可以分为“**句子选择**”和“**人工句子生成**”这两种方案。

首先来说一下句子选择这个想法。句子选择的思路是，我们希望在最后的意见总结里，能够利用已有的非常有代表性的句子，这样用户看到的最后的总结会显得更加真实。那么这里有两个问题：一个问题是如何对所有的句子进行筛选；第二个问题是如果有重复多余的字句，又如何进行进一步的选择。

通常情况下，我们通过对句子打分来筛选，这个时候，一般需要设计一个打分机制，这个机制往往是看这个句子对某一个实体的方面是否进行了有情感的评价。然后，对所有句子进行聚类，这样所有评价类似的句子就可以被聚集到一起，从而能够过滤掉重复多余的字句。

那么，人工句子生成又是怎么运作的呢？首先，我们必须知道这个物品的哪些方面得到了用户的评价，而且都是什么样的评价，比如是正面评价还是负面评价。然后，把这些信息和一个语言模型，也就是语句生成器相连接，从而能够“生成”最后的总结语句。值得注意的是，这样生成的总结语句并不会出现在所有用户的原始评价中，因此也可能会对用户的最终体验有一定的影响。

除了基于方面的意见总结以外，还有一些类似的但是并不完全一样的总结方案。比如，有一种总结方案叫“**针对性观点总结**”（Contrastive View Summarization）。这个任务更加突出针对同一个主题的两种截然相反的观点。这种意见总结不仅可以针对商品，也针对新闻事件，比如某一个政策法规、选举结果等往往比较有争议的话题事件，“针对性观点总结”往往会有比较好的用户体验。

意见搜索

我们可以认为“意见搜索”是建立在意见总结之上的一个任务。通常情况下，意见搜索需要完成的任务是用户输入一个主体的名字，我们需要返回和这个主体相关的意见信息，这些意见信息有可能是通过意见总结而呈现给用户的。

意见搜索的难点，或者说是和传统搜索不一样的地方主要还是在于**针对意见信息的索引和检索**。

第一，我们需要在索引库中找到有哪些文档和字句包含了我们所需要查询的主体。可以说，这一点和传统的搜索是非常类似的。

第二，我们需要在找到的文档和字句中检查是否包含主体的某种意见，以及其褒义或者贬义的评价。这就是有别于传统搜索的地方。在找到了所有关于某个主体的情感评价以后，我们需要设计一个评分机制从而返回最有说服力的文档，并且还需要在这些文档的基础上进行意见总结。很显然，这些步骤都是传统的搜索中并没有的。

按照上面所说的这两点，我们可以把意见搜索分为两个阶段。

第一个阶段，就是利用现有的搜索技术，比如我们介绍过的文本搜索或者基于排序学习的搜索等方法，得到最初的一个文档的备选集。然后进入下一个阶段，就是通过一个模型，针对所有的文档进行基于意见的打分。这个模型可以是简单的分类器，用于分析当前的字句和主体的意见究竟有没有关系，也可以是一个更加复杂的模型，输出当前的文档和主体的哪一个方面有关系。在这里，任何一种文本分类器都可以被利用起来。

总体来说，意见搜索可以算是对于意见分析和总结的一个综合体现。

总结

今天，我为你介绍了一类比较高级的文本情感分析技术：意见的总结和搜索。至此，我们对于文本情感分析的分享就告一段落了。

一起来回顾下要点：第一，我们讲了意见总结的的重要性，基本概念和技术；第二，我们分享了意见搜索的基本概念和两个阶段的技术。

最后，给你留一个思考题，除了常见的观点和评分以外，用户对于产品的评价一般还在意哪些信息呢？

欢迎你给我留言，和我一起讨论。

上一篇　099 | 如何来提取情感“实体”和“方面”呢？

下一篇　101 | The Web 2018论文精读：如何对商品的图片美感进行建模？

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。