090 | Word2Vec算法有哪些扩展模型?

2018-04-30 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:20 大小 2.90M

从上一期的分享开始，我们进入到文本分析的另外一个环节，那就是介绍一个最近几年兴起的重要文本模型，Word2Vec。这个模型对文本挖掘、自然语言处理等很多领域都有重要影响。我们讨论了 Word2Vec 模型的基本假设，主要是如何从离散的词包输入获得连续的词的表达，以及如何能够利用上下文从而学习到词的隐含特性。我们还聊了两个 Word2Vec 模型，SG（SkipGram）模型和 CBOW（Continuous-Bag-of-Word）模型，讨论了它们都有什么特性以及如何实现。

今天，我们就来看一看**Word2Vec 的一些扩展模型**。

Word2Vec 的扩展思路

在列举几个比较知名的 Word2Vec 扩展模型之前，我们首先来看看这个模型怎么进行扩展。

首先，我们来回忆一下 Word2Vec 的一个基本的性质，那就是这是一个语言模型。而语言模型本身其实是一个**离散分布模型**。我们一起来想一想，什么是语言模型？语言模型就是针对某一个词库（这里其实就是一个语言的所有单词），然后在某种语境下，产生下一个单词的模型。也就是说，语言模型是一个**产生式模型**，而且这个产生式模型是产生单词这一离散数据的。

既然是这样，如果我们更改这个词库，变成任何的离散数据，那么，Word2Vec 这个模型依然能够输出在新词库下的离散数据。比如，如果我们把词汇库从英语单词换成物品的下标，那 Word2Vec 就变成了一个对物品的序列进行建模的工具。**这其实就是扩展 Word2Vec 的一大思路，那就是如何把 Word2Vec 应用到其他的离散数据上。**

扩展 Word2Vec 的第二大思路，则是从 Word2Vec 的另外一个特性入手：上下文的语境信息。我们在之前的介绍中也讲过，这个上下文信息是 Word2Vec 成功的一个关键因素，因为这样就使得我们学习到的词向量能够表达上下文的关联所带来的语义信息。这也是传统的主题模型（Topic Model）例如 LDA 或者 PLSA 所不具备的。那么，我们能不能对这个上下文进行更换，从而使得 Word2Vec 能够产生完全不一样的词向量呢？答案是肯定的，这也是 Word2Vec 扩展的重要思路。

除此以外，还有一个重要的分支，那就是很多研究者都希望往 Word2Vec 里增加更多的信息，比如文档本身的信息，段落的信息以及其他的辅助信息。**如何能够让 Word2Vec 对更多信息建模也是一个重要的扩展思路。**

Word2Vec 的三个扩展

我们要介绍的第一个扩展是由 Word2Vec 作者本人提出的，就是**把学习词向量的工作推广到句子和文章里**，在论文《句子和文档的分布表示》（Distributed representations of sentences and documents）[1] 里进行了详细的阐述。这个扩展主要是解决如何能够更加“自然”地学习到比词这个单位更大的单位（比如段落或者文档）的隐含向量。

当 Word2Vec 被发明之后，很多研究者都发现这是一个**能够把离散的词表达成连续向量的利器。**然而，一个应用场景很快就成为了大家的拦路虎，那就是 Word2Vec 仅仅是在词一级数据上进行建模，却无法直接得到文档一级的隐含信息。

有一种做法是这样的，比如针对一个句子或者一个段落，我们就把这个句子里的词所使用的词向量加权平均，认为这个加权平均过的结果就是段落的向量了。很明显，这是一种非常不精确的处理方法。

那么，这篇文章的核心则是如何能够在模型本身上进行修改，从而可以学习到比词更加高一层级单元的隐含向量。具体的做法，就是修改原始 Word2Vec 的上下文信息。我们回忆一下 SG 模型和 CBOW 模型都有一个关键的信息，那就是利用上下文，也就是一个句子周围的词来预测这个句子或者上下文中间的一个词。这就是 Word2Vec 能够利用上下文信息的原因。那么这里的修改就是让这个上下文始终都有一个特殊的字符，也就是当前段落或者文章的下标，从而这个下标所对应的隐含向量就是我们所要学习到的段落或者文档的向量。在这样的情况下，作者们通过实验发现，学到的段落向量要比单独用加权平均的效果好得多。

我们要看的第二个扩展，来自论文《线：大规模信息网络嵌入》（LINE: Large-scale Information Network Embedding）[2]，就是**把 Word2Vec 的思想扩展到了另外一种离散数据图（Graph）的表达上。**

刚才我们提到，只要是离散的数据，Word2Vec 都有可能被应用上。那么，图的数据建模的场景是什么呢？我们设想一个社交网络（Social Network）的数据。每一个用户都有可能和其他用户相连，而两两相连的用户所组成的整个网络就是社交网络的庞大的用户信息。很明显，如果我们把用户看作单词，那么整个社交网络就是一个单词和单词的网络。如果我们把两个单词（在这里是用户）之间的连线看成是单词出现在一起的上下文，那么，我们其实就可以利用 Word2Vec 这样的模型对社交网络所表达图进行建模。这就是这篇文章里作者们利用 Word2Vec 对社交网络建模的核心思想。

当然，和在文档中不同，在图里面，上下文这一关系其实是比较难以定义的。因此，很多后续的工作都是关于如何更加有效地定义这一上下文关系。

最后，我们结合论文《用于支持搜索中查询重写的上下文和内容感知嵌入》（Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search）[3] 来看另一个 Word2Vec 的扩展。这个扩展是**尝试在查询关键词和用户点击的网页之间建立上下文关系，使得 Word2Vec 模型可以学习到查询关键词以及网页的隐含向量。**

这也就是我们提到的，巧妙地搭建上下文关系，使得模型可以学习到离散数据的隐含表达。你可能比较好奇，这里的离散数据是什么呢？这里有两组离散数据：第一组就是每一个查询关键词，这完全可以按照 Word2Vec 原本的定义来走；第二组离散数据，就是每一个网

页。注意，这里不是网页的内容，而是某一个网页作为一个下标。那么，从模型的角度上来说，这里我们要做的就是利用查询关键词来预测网页出现的概率。

总结

今天我为你介绍了 Word2Vec 模型扩展的一些基本思路和一些实际的案例。

一起来回顾下要点：第一，我们讨论了 Word2Vec 这个模型需要扩展的思路，比如从离散数据入手或者从上下文入手；第二，我们分享了三个比较经典的 Word2Vec 扩展。

最后，给你留一个思考题，Word2Vec 能否扩展到连续数据中呢?

欢迎你给我留言，和我一起讨论。

参考文献

1. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14), Eric P. Xing and Tony Jebara (Eds.), Vol. 32. JMLR.org II-1188-II-1196, 2014.

2. Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1067-1077, 2015.

3. Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM, New York, NY, USA, 383-392, 2015.

上一篇　089 | 为什么需要Word2Vec算法?

下一篇　091 | Word2Vec算法有哪些应用?

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。