深度学习如何帮助搜索引擎提高效果？

原创 cheng 待字闺中 2017-04-02

最近，百度度秘事业部首席技术官朱凯华在上海计算机学会做了题为《AI赋能的搜索和对话交互》的主题报告。详细介绍了深度学习在百度搜索引擎的引用，以及神秘的DuerOS系统。其中有关深度学习在搜索引擎的应用，引起了我的兴趣。我本身也是搜索出身，在计算所的时候从事信息检索相关的工作，后来人民搜索、百度App搜索、招聘中候选人、职位的搜索都是搜索引擎技术的应用。尽管做了这么多的搜索应用，但技术的变化并不大。近几年随着深度学习越来越火，也在业余时间思考深度学习该如何提高搜索引擎的效果呢？不过一直也没有很好的思路，只是比较模糊的尝试方向。

这次看了百度的分享，再次激起我的思考。深度学习在NLP领域的应用越来越多了，取得了很多成绩，那应该是时候，再去尝试下这个方向了。对于深度学习在NLP领域的应用，大家可以阅读待字闺中之前发表的两篇文章：

- 深度学习如何在自然语言处理中大显身手？
- 97.5%准确率的深度学习中文分词（字嵌入+Bi-LSTM+CRF）

百度的分享中，很多点因为商业的原因不能说透，整体理解下来效果还是很明显的，主要有以下几点：

- 海量的点击数据，作为训练集。主要是依托大量的query/点击的title对。这个百度独有的优势，天然有大量的标注数据。
- 更好的调试理解DNN的训练结果，要有方法更好的调试DNN模型是否符合预期的工作。
- 通过CNN建模短距离依赖关系。
- 通过RNN建模长距离依赖关系。这一点比较关键，也是在传统搜索引擎过程中遇到的一个痛点，如何建模词之间有间隔的概念。

百度目前主要的思路是在Rank的层面做的。纵观搜索引擎的历史，Rank做得好的都有一个前提就是得有点击数据。然而，这有一定的局限，我也觉得这个并没有满足对深度学习的带来变革的期待，只是做了一些不错的改进。除了Rank，还有哪些方面会有可能的尝试呢？

我们对搜索引擎本身进行拆解，可以简单得到：

- 索引
  - 索引数据结构，空间时间的权衡
- 检索

- 检索的算法，要找到结果，更细的拆分，还有很多小点，比如Query分析，也是很重要的一点。
- Rank，找到结果进行更好的排序。

百度在Rank上做了改进，那我们就尝试下索引、检索，以及Query分析是否可以有想法上的突破。**深度学习在自然领域的应用，最重要、最基础的是word2vec，而且我认为是划时代意义的。其本质在于，很多原来不可以计算的“变量”，通过word2vec表示之后都可以计算，而且是具备客观物理意义的。**之前我们表示某一个词的时候，采用的是One-Hot的编码方式（向量中只有一列值为1，其他的值为0），极其稀疏，在计算过程中损失了大量的语义信息。word2vec则不同，充分考虑前后词之间的关系，得到一个稠密的、更有计算意义的向量。而且，这一思想应该不仅仅局限于自然语言理解这个领域，比如我们考量股票市场中股票之间的关系构成一个图，之后可以将每个节点表示为一个向量，可以充分建模这些股票，以及他们之间的关系。

word2ve的好处还有很多，百度的搜索Rank改进，基础也是word2vec。但word2vec确实给索引结构、搜索算法带来了了巨大的挑战。搜索引擎主流的索引结构是倒排索引，辅助树形结构的索引，综合权衡时间、空间、以及效果：

- 千亿规模的网页
- 100ms内返回结果
- 保证基本的文本相似度

已经是相当长时间内最好的解决方案了。但无论是倒排索引，还是树形结构的索引，都无法很好的解决word2vec的相近向量查找问题。之前的数据结构和算法，使用的都适用字面的常量、或者是离散的值。而word2vec是连续的值。无法利用这些结构。如果有同学实现过simhash的话，simhash构建索引查找相近文档的时候也是两个向量的计算。通过分块，舍弃一些精度。还是可以利用倒排索引进行提速和缩小查找范围的（一个具体实现见：https://github.com/sing1ee/simhash-java ）。simhash算法得到的向量是0或者1的，这是可以索引的基础。但word2vec是连续的值，那么能否解决呢？一定能的，这个方向我是看好的。**从索引结构，以及检索的算法角度进行改进，实现真正的语义搜索引擎。**

基于这样的想法，我开始搜集资料：

- https://github.com/DiceTechJobs/ConceptualSearch 使用word2vec，查找近义词，进行索引的扩展，然后开发Solr的插件提供功能，有一定的效果提升。但要注意word2vec的结果是不能直接用的，要区分出哪些是近义词。可以快速尝试的一个点。
- https://github.com/spotify/annoy 近邻算法的一种快速实现，可以看看找找启发。

- https://github.com/facebookresearch/faiss 重磅：Facebook发布AI搜索引擎Faiss：比最先进搜索算法快8.5倍。这是我今天刚刚看到的论文和code，还在了解中。

随着深度学习的不断进步，语义搜索的实现越来越近。单点的突破，会带动其他点的变革。还在不断的求知探索中，有想法就会在待字闺中和大家讨论，也可以在“待字闺中读