之前的检索模型, 如果一个词没有出现过, 就检索不到, 比如说, 如果搜索"电脑", 那么包含"计算机"但不包含"电脑"的结果是得不到的. 我们希望能检索到近义词. 对此可以用word2vec, 但我没用这种做法. 由于此语料的特点是短, 我用了sentence embedding. 采用预训练好的BERT模型. elasticsearch中存储的doc增加了一个字段, 存储文档向量. 搜索时, 比较余弦相似度.

实验结果如下:

Rank	News	Score
1	081203计算机应用技术；	1.9432029
2	0计算机通用软件。	1.9401981
3	0微型计算机；	1.9392992
4	0计算机考试。	1.9268427
5	0服务器程序。	1.908815
6	0与互联网概况；	1.9052893
7	0或者<N>数据库服务器。	1.903332
8	0<N>网管系统软件。	1.9025323
9	0接入Powerline网络。	1.9013579
10	0电子计算机及其外部设备中（<N>）集成电路卡；	1.8986117

可以看到, 结果与计算机网络关系很大, 比如结果8, 在原来的搜索中是不会得出这样的结果的.