

带复杂计算的金融领域自然语言查询的 SQL 生成

陶益怡

郑州财税金融职业学院 河南 郑州 450000

[摘要]在金融领域，一个比较重要的应用就是数据的查询，比如我们想查询一家银行的开户情况，这就是一个典型的 SQL 查询，但是现在很多公司在做的都是无监督学习或者半监督学习，比如在一些招聘网站上可以看到很多公司在进行一些数据分析，这类工作其实对自然语言查询有一定要求，对于业务人员来说，自然语言查询比纯 SQL 查询更容易理解，所以对于金融领域来说，自然语言查询也是非常重要的一个应用。

[关键词]Text-to-SQL；自然语言查询；金融领域；行计算查询；分治方法

[中图分类号] G641 **[文献标识码]** A **[文章编号]** 1647-9265 (2024)-0019-18 **[收稿日期]** 2024-01-17

一、什么是自然语言查询

自然语言查询其实就是一种用自然语言描述查询条件的查询，比如我们想查询某家银行的开户情况，这就是一个典型的 SQL 查询，我们可以把它分为几个步骤：

- 1.数据导入：如果是纯 SQL 查询，我们需要将数据导入到数据库中，而自然语言查询只需要将输入的 SQL 语句转化成自然语言描述即可；
- 2.解析：如果是纯 SQL 查询，我们还需要解析这个语句，获取它所需要的信息；
- 3.执行：如果是自然语言查询，我们就可以直接执行该语句了。

前面的两个步骤都是基本不需要人工干预的，只要将输入的 SQL 语句转化成自然语言描述即可。但是后面两个步骤就相对比较复杂了。

上面这张图是一个典型的自然语言查询示例，首先我们可以看到这句 SQL 语句其实是非常长的，在这种情况下要想把它转化

成自然语言描述其实比较困难。这里需要引入一个新的概念叫分词，它是一种对自然语言进行分词处理的技术，所以我们可以看到这个 SQL 语句中有很多汉字。下面我们来详细讲解一下分词。

二、如何做

第一步，首先我们要把这张表进行结构化，我们要把所有的字段都输入到一个表格中，因为 SQL 查询是一种关系型数据，我们可以先创建一个列，然后在这个列中输入一些字段，比如“账户”、“客户”、“账户余额”等，然后我们需要对这些字段进行统计，比如一些值是很小的字段，可以直接删除不需要统计的字段，然后通过公式来统计这些字段的值。在这里我们需要做的是对字符串进行编码，就是将字符串转换为一个编码，比如将“账户”转换为 `digits.df` ("clean")，然后在查询中用 `digits.df` 来匹配这个字段。如果有一些字符串在前面没有出现过或者是很长的字符串可以直接删除掉不

需要统计。

第二步，我们要根据 SQL 查询的结果生成自然语言查询语句，比如查询“银行账户余额”，就会在第一行输出“银行账户余额”。

三、数据准备

1.数据集：我们使用了一个有标注的数据集，有两个字段，一个是银行名称，另一个是开户时间。这个数据集应该说是比较好的，因为没有什么特殊的要求，不需要额外进行预处理。

2.数据预处理：由于是无监督学习，我们首先要进行清洗。清洗方法大致有三种，一种是手工清洗，一种是使用 `cnn` 工具清洗，还有一种就是使用 `stackoverflow` 工具来做预处理。其中手工清洗比较复杂，但是可以得到非常干净的数据；`cnn` 工具和 `stackoverflow` 的预处理方式差不多；`stackoverflow` 可以把预处理和清洗一起做，效率更高。我选择了使用 `stackoverflow` 来进行数据的预处理。

3.根据规则对数据进行预处理

四、模型设计

本文的模型结构比较简单，主要包括两个子模型，一个是基于规则的规则引擎，它用于处理从一句话到另一句话的逻辑组合。

一个是基于统计的统计引擎，它用于处理从一个字符串到另一个字符串的语义组合。两个子模型通过将查询转化为语句实现对句子的逻辑组合的处理。所有代码都是在 Python 中实现的，这需要有一定的编程基

础，因为使用自然语言查询肯定会用到一些复杂计算。为了简化代码，我们没有采用 `TextRank` 这样非常复杂的方法，而是使用了最基本的 `textRank`。

根据上面描述的内容，我们可以进行如下设计：在表中增加两行字段，分别记录“开户”和“贷款”，以匹配金融领域常用的查询语句。因为我们想做一次 SQL 查询，所以我们将两行字段作为模型输入。为了便于处理，我们将所有字符串按照“姓名”、“性别”、“年龄”、“职业类别”等进行了处理，并通过字典实现了对文本内容进行拆分。

五、模型训练和预测

首先我们需要把我们的数据进行预处理，将 SQL 查询转换为字符串形式，然后用 Tensorflow 的 `TextCNN` 和 `ResNet` 模型进行训练。

我们需要将预处理后的 SQL 查询转换为 Tensorflow 的 `TextCNN` 模型，将每个 SQL 查询中的字符串转换为 Tensorflow 模型的向量表示，然后用 `tensorflow` 训练。

然后我们可以使用训练好的模型对 SQL 查询进行预测，来判断这个 SQL 查询是否符合我们想要的结果，这里我们使用了 Tensorflow 中的 `softmax` 层对字符串进行分类，然后使用分类结果作为最终预测结果。

最后，我们可以根据我们得到的预测结果来判断这个查询是否符合我们想要的结果，如果不符合那么就说明这个 SQL 查询是不符合我们要求的，如果符合我们要求则说明这个查询是符合我们要求的，如果不符

合那么就说明这个查询不符合我们要求。

这样一个完整的模型就训练好了，这其实是一个简单的模型生成过程，当然也可以根据实际情况进行调整，比如一个 SQL 查询中可能有很多个字符串需要转换成 Tensorflow 中模型能理解和接受的格式。其实这样一个过程是非常简单和常见的，如果需要处理大量数据的话，则可以进行模型训练和预测。

六、结语

对于金融领域的数据分析，一般情况下，对自然语言查询的需求是比较大的，所以我们要想办法将其实现出来，其实对于纯 SQL 查询来说，要想实现很容易，但是对于带有复杂计算的天然语言查询来说，要想实现其实很难，因为需要非常多的数据，而这些数据在系统中可能并不是一个统一格式，所以需要做一些预处理，比如分词、去停用词、词干提取等。同时在学习的过程中也

要不断地修正模型和代码。对于这种带复杂计算的天然语言查询，我认为是可以实现的。但是要想实现这种查询其实并不容易，因为 SQL 本身是比较复杂的语言，不像天然语言那么好理解，所以需要有大量的训练数据和模型。对于金融领域来说，业务人员更需要对数据进行分析和处理。当然从技术上来说也是可以实现的。下面我们将会介绍如何利用机器学习的方法来生成带复杂计算的天然语言查询。

参考文献：

[1]潘璇,徐思涵,蔡祥睿,等.基于深度学习的数据库天然语言接口综述[J].计算机研究与发展.2021,(9).DOI:10.7544/issn1000-1239.2021.20200209.

[2]曹金超,黄滔,陈刚,等.天然语言生成多表 SQL 查询语句技术研究[J].计算机科学与探索.2020,(7).DOI:10.3778/j.issn.1673-9418.1908025.

SQL generation of natural language queries in the financial domain with complex computations

Tao Yi yi

Zhengzhou Finance and Taxation Vocational College Zhengzhou, Henan 450000

Abstract: in the financial sector, a more important application is the data query, such as we want to query a bank account, this is a typical SQL query, but now many companies are doing are unsupervised learning or half supervised learning, such as in some recruitment sites can see many companies in some data analysis, this kind of work actually has certain requirements for natural language query, for business personnel, natural language query easier to understand than pure SQL query, so for the financial sector, natural language query is also a very important application.

Key words: Text-to-SQL; natural language query; financial domain; line calculation query; divide and conquer method.