本周报告

> 2020/02/28 1652792 罗吉皓

本周主要做了毕业设计的前期调研，分为以下几个部分：

1. GDELT数据结构分析
2. 数据来源CAMEO（Conflict and Mediation Event Observations）调研

背景知识

本次毕业设计的主题是搭建一个中文政治事件的数据库构建，经过初步分析之后，我将毕设分为三个部分：

1. 数据获取；数据获取这一部分在之前的实验室工作中其实已经有所涉及了，实验的方法主要参考事件数据系统的研究方法，通过定义字典对于文本进行词性，词义分析，但原系统的语言为英语，其主要分析也是针对英语的文本，对于中文文本的分析需要我们进一步思考
2. 数据库数据结构设计；预定义的数据格式是非常重要的，一个好的数据格式设计能够在数据导入之前就很好的解决数据冗余等问题，对于实验的顺利开展有很大的帮助。数据格式中不仅要包含一些固定字段的设计，也可以通过组合多个元素，组成集合字段，来帮助解决数据冗余的问题，还需要设计一些字段来帮助实现推理等功能。
3. 简易前端展示及应用；在数据库构建完成后，可以通过一个简易的前端应用来进行数据展示，目前考虑的是希望能用知识图谱的形式来进行展示，相关推理的应用需要在前端搭建完成后进行探索。

数据获取

CAMEO调研

之前数据获取这一部分在实验室的工作中已经有所涉及，具体的思路也比较清楚。数据的主要来源是新闻媒体报道，数据获取的算法主要依赖于Kansas Event Data System的算法。

CAMEO

Kansas Event Data System是基于Macintosh的机器编码系统编写的，用于使用模式识别和简单的语言解析来生成事件数据。系统代码来自描述国际事件的机器可读文本，NEXIS数据服务，光学字符识别和CD-ROM提供初始文本数据。

CAMEO（Conflict and Mediation Event Observations）是Kansas Event Data System系统中的一部分，它结合第三方调解研究开发的编码方案，并引入了一些新功能：

- 编码方案针对调解研究进行了优化，并包含许多特定于调解的子类别
- 扩展了“武力使用”的类别，可以更好的分析事件中的暴力程度
- 结合了许多WEIS类别，帮助这些类别在机器编码中得到更为可靠地区分。
- 开发了系统的分层编码方案用于处理不同状态的用户
- 为宗教团体和种族团体开发更广泛的分类法

政治事件本身就是充满冲突和解，宗教种族等信息的，新功能的引入帮助算法能够更好的分析政治事件，“武力使用”类别的拓展，细分帮助我们更为准确判断政治事件的严重程度，对于宗教团体和种族团队的细分能更好的挖掘事件中的人物相关背景信息。编码系统的完善确保了数据的准确性。

通过调研我们可以发现CAMEO对于信息的分类，字典的编撰是相当细致的，相关文档地址：http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf

CAMEO对于信息的处理主要有两部分，一个是对于句式，文本的分析，如下图所示

CAMEO	**014**
Name	**Consider policy option**
Description	Review, reflect upon, or study policy option.
Usage Notes	This event form is typically, although not exclusively, a verbal act. There is no limitation on types of policies that could be under consideration.
Example	Europe's leading security forum is exploring the possibility of international patrols to monitor the former Yugoslav republic of Macedonia's border with Serbia, its envoy said on Friday.
Example	Malaysia is considering giving money to 20,000 Vietnamese boat people in the country to entice them to return home, foreign minister said on Tuesday.

CAMEO非常细致的列出不同状态，情形下句子的分析方式。

第二种是编码：

01: MAKE PUBLIC STATEMENT
010: Make statement, not specified below
011: Decline comment
012: Make pessimistic comment
013: Make optimistic comment
014: Consider policy option
015: Acknowledge or claim responsibility
016: Deny responsibility
017: Engage in symbolic act
018: Make empathetic comment
019: Express accord

CAMEO中的编码非常细致，从表达内容到人物的属性都有各自的编码，编码的存在一方面起到索引的作用，另外一方面也为后面数据存储做了很好的基础，预防数据冗余。

相关难点

1. 信息量相对比较小

文本分析算法存在一定的局限，在目前的实验中，对于一篇文章的分析往往只能得出1-2个核心事件，相比较而言，Google GDELT核心文本分析也是基于目前Kansas Event Data System中的算法，核心文本字典也是引用了CAMEO的字典，其强大的文本收集能力能够实时获取海量数据进行分析，源源不断的数据补充帮助补充其事件知识库，即便如此，google依旧在收集了分析了10年数据后才正式开放整个系统，可见如果要对于事件知识图谱进行分析，推理的话，海量的数据必不可少。而在实验室阶段的话我们没有必要去获取像谷歌一样的海量的数据，但是一定的数据量还是必不可少的，或许我们可以通过缩小整体实验范围，精确某些点深入分析来减轻数据量的压力。

2. 语言转换

   CAMEO是一个针对英语政治事件的数据库，其字典的搭建，语法，内容的分析都是基于英语文本及相关概念的。而本次毕业设计的目标为搭建一个中文政治的数据库，中文文本相对于英语文本分析来说更为复杂，不论是在文字，句式，语法等各个方面，中文文本分析需要考虑的方面要比英语文本分析要多得多。其中字典的搭建尤为复杂，不像英语文本，其相关的研究非常多，有很多现成的英语字典来帮助我们划分英语文本，而在中文领域，这一块的研究相对比较匮乏。实验室项目目前在这一块已经有一定的研究，如下图所示：

```
---  作出评论   [010]  ---
传递
保卫    [074]
得到
获得    [061]
取得
忽视    [120]
重申
重说
重述
揭露
揭示
透露
确认
批准
说明
声称
主张
陈述
宣称
讲
发行    [080]
拒绝    [120]
检验    [090]
表示
宣告
宣布
透漏
传达
- * 提名                                                    [010]         # REVEAL
- * 主张                                                  [010]         # RELEASE
- * 断言                                                  [010]         # RELEASE
- * 名字                                                    [010]         # REVEAL
- * 解职令                                                 [010]         # CONFIRM
- * 报告                                                    [010]         # ISSUE
- * 指控                                                  [010]         # CONFIRM
- * 辩护                                                  [010]         # CONFIRM
- * 立场                                                    [010]         # RESTATE
- * 观点                                                    [010]         # RESTATE
- * 优先                                                    [010]         # GET
- * 领先                                                    [010]         # GET
- * 投机                                                 [010]         # DISMISS
- * 外视图                                                  [010]         # SET
```

这是部分的动词字典，学长们将中文的文本与英语文本之间形成映射来帮助构建中文字典，结果如下：

```
2018-08-02 00:00:00     ---MIL  ---GOV  192
        joined_issues   null
        ids     339764-e_0001
        StorySource     NULL
        content  在为期10天的演习中，马美两国海军舰队将在南中国海展开舰炮射击、水下作战、潜水与援救及后勤支援管理
        Source  海军
        Target  管理
        actortext       海军     管理
        eventtext       展开
        actorroot       ---     ---
        eventroot       192 占领类
        location3       在南中国海

2018-08-02 00:00:00     PERSON_GATES    ---     192
        joined_issues   null
        ids     339765-0004_0002
        StorySource     NULL
        content  随后，美国防部长盖茨在接受哥伦比亚广播公司专访时忧虑地表示，美国每天都在遭受网络攻击，五角大楼不得不把从事网络攻防任务的
专家扩充4倍
        Source  盖茨
        Target  ---
        actortext       盖茨     ---
        eventtext       扩充
        actorroot       盖茨     ---
        eventroot       192 占领类
        location3       美国
```

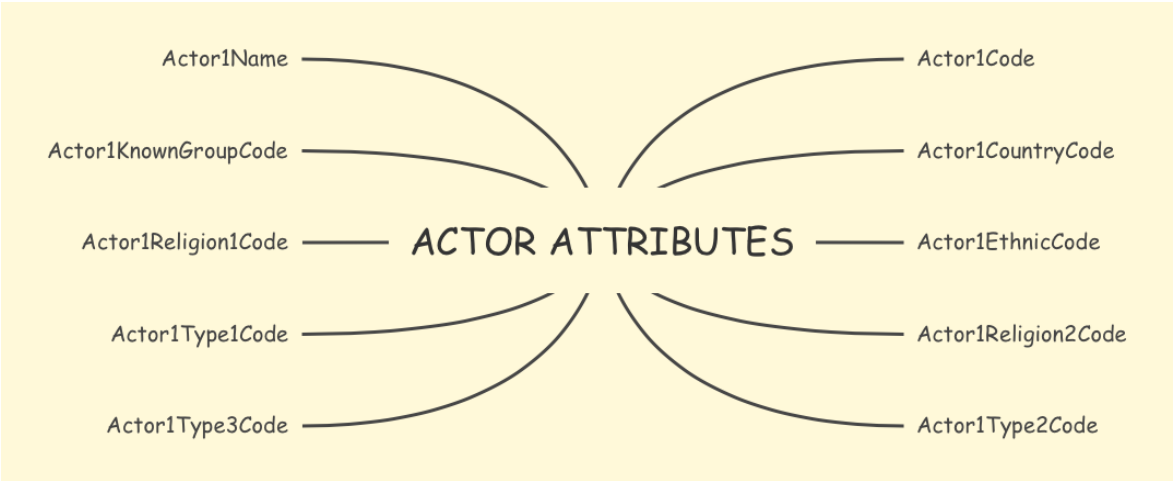
初步的文本分析已经可以完成，但是还有几个方面需要我们考虑：

1. **文本消歧**。中文中有很多词语其实表达的都是一个意思。比如说中国首都，北京，北平等词语在大部分条件下表达的都是一个意思，但是在计算机处理中，这些会被单独提取，标为不同的对象，如何将这些信息统一值得我们思考。
2. **字典的拓展**。字典可以理解为写代码时候的依赖库，字典越完善，其分析的时候就能更细致。对于字典的拓展也是一个比较好的研究方向。

数据库数据结构设计

我觉得数据库的设计可以参考一下GDELT event database的设计。在数据库的设计中，数据格式的定义是非常重要的，一个好的数据格式设计能够在数据导入之前就很好的解决数据冗余等问题，对于实验的顺利开展有很大的帮助。在目前的调研中，我觉得主要有两种方法能够增加数据的健壮性：

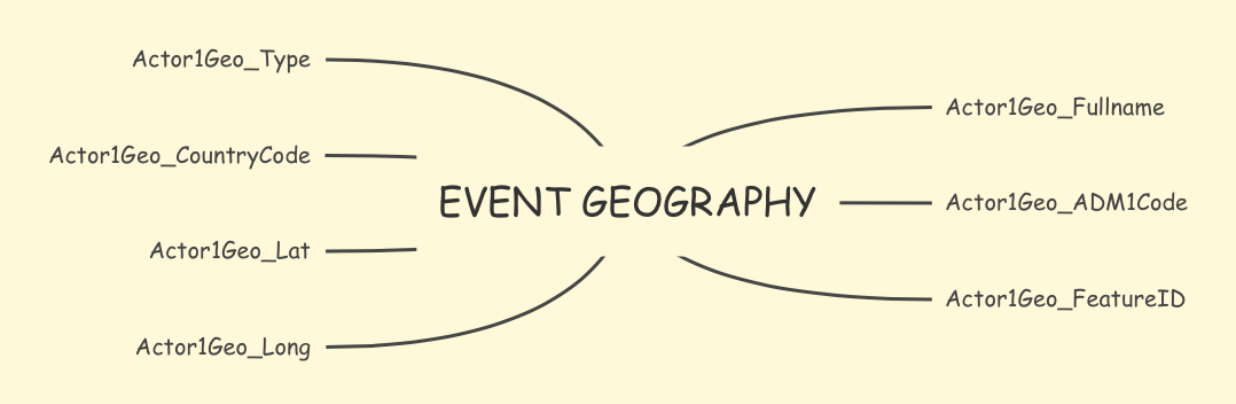
1. 引入编码。最显而易见的去除冗余的方法便是添加编码。比如在GDELT中的Actor Attributes字段：



在其中，Actor1TypeCode和ActorReligionCode相关的5个字段的设置都是为了防止重复。

> **Actor1Type2Code.** (character or factor) If multiple type/role codes are specified for Actor1, this returns the second code.

当然这些编码的引入并不是随机引入的，Actor字段主要编码的引入依赖于CAMEO中的相关编码。又比如Event Geography字段



其中部分编码来源于CAMEO，另外一部分基于leetaru算法。

因此我们在设计字段的时候，首先要确认需求，找到所有需要存储的字段，再对于这些字段进行调研，以CAMEO编码为依托，寻找相关算法的支持，尽力保证字段的健壮性，去除冗余。

2. 引入多位合成字段。多位合成字段的优势是比较容易处理，信息覆盖量大。多个字段合并而成的字段不太容易重复，处理信息的时候只需要一次提取就可以得到大量相关信息，减少sql查询的复杂度。而缺点是并不满足范式，字段长度较长，对存储的压力比较大。多位合成字段非常灵活，比较适合用作知识图谱搭建，数据挖掘等操作，对于推理等功能会有很大的帮助。

   在设计数据库的时候，以多位合成字段为核心单独设计一两个分表，对以后的业务，推理等需求是一个比较良好的准备。

简易前端展示及应用

由于初始数据的结构为三元组，比较符合知识图谱的存储方式，因此在考虑展现形式的时候，我也偏向于使用知识图谱的方式进行展示，相关推理的应用需要在前端搭建完成后进行探索。