GDELT 事件知识图谱

1652792 罗吉皓

简要概述

GDELT项目是有史以来建立的最大，最全面，最高分辨率的人类社会开放数据库。 仅2015年的数据就记录了近四分之一的情感快照和超过15亿的位置引用，而其总存档时间超过215年，使其成为现有的最大的开放式时空数据集之一，并推动了全球人类社会“大数据”研究的边界。 GDELT项目每时每刻监控着每个国家的几乎每个角落的100多种语言的各种不同形式的新闻媒体，它的全球知识图谱将全世界的人们，组织，位置，主题，数量，图像和情感连接到整个星球上的单个整体网络中，为全球提供了一个自由开放的计算平台。

GDELT 知识图谱框架分析

事件知识图谱概述

知识图谱是由Google公司在2012年提出来的一个新的概念。从学术的角度，我们可以对知识图谱给一个这样的定义：**“知识图谱本质上是语义网络的知识库”**。

在知识图谱里，我们通常用“实体”来表达图里的节点、用“关系”来表达图里的“边”。**实体指的是现实世界中的事物比如人、地名、概念、药物、公司等，关系则用来表达不同实体之间的某种联系**，比如人-“居住在”-北京、张三和李四是“朋友”、逻辑回归是深度学习的“先导知识”等等。

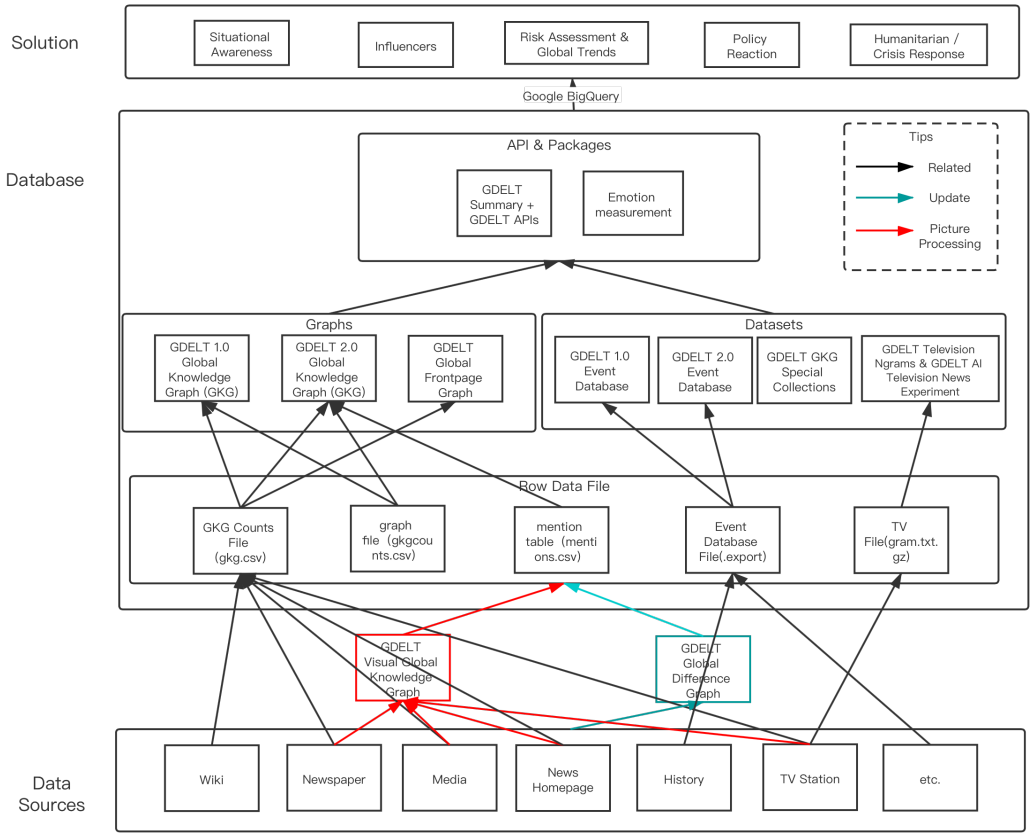
而随着知识图谱应用领域越来越广泛，单纯的实体知识库，单纯的实体-关系或者实体-属性-值类型的知识过于静态，不能满足日益复杂的需求和应用领域对知识图谱越来越高的期望。不论是在金融投资领域对根据事件之间的因果和顺承关系进行推理、沙盘推演和预测未来事件的要求，亦或是客服及咨询领域中，精确客户咨询事项中的状态变化捕捉需求，现有的实体产业链模型都无法胜任。而这种通过事件的表示和处理来推演、预测和预警未来，校准实体知识库，与用户就其正在参与或拟参与的事件进行深度沟通，具有巨大的应用价值。

随着近几年机器学习等技术的发展，从大数据中抽取事件已变成可能。而事件抽取不同于静态关系的抽取。关系是静态的，**事件是动态的**，需要建立事件间因果、顺承、细分、概括等关联关系的复杂网络。关系抽取只需要考虑实体对实体，事件抽取则必须要考虑事件对实体、事件对时空属性、事件对事件等多种复杂情况。

事件知识图谱不同于普通的知识图谱，不仅在于它的刻画对象是事件，还在于它在刻画事件的过程当中，不可避免地会与实体知识库之间产生互动，形成包括实体、关系、属性、事件、事件属性、事件参与角色（论元）和事件之间的特殊关联关系等在内的全新数据结构和知识表示框架。

	知识图谱	事理图谱
描述知识	万物本体	逻辑社会
研究对象	名词性实体及其属性、关系	谓词性事件及其内外（空间、时间域）联系
构建目标	万物互联	全逻辑库，逻辑演化模型
回答问题	When 、Who、What、Where	Why、How
组织形式	有向图	有向图
知识形式	<实体，属性，属性值>,<实体，关系，实体> 三元组	<事件，论元集合，逻辑关系> 多元组
知识确定	事实是确定的	逻辑不确定，有转移概率
知识状态	相对静态，变化缓慢	动态的
知识敏感	精确性要求极高，实时性要求极高	可一定容错，参考逻辑
构建难点	知识本体的搭建、知识抽取与融合	事件的表示、事件的抽取；与知识图谱的融合

GDELT 知识图谱构建



类比普通知识图谱的构建，目前关于事理图谱的构建方式上，主要包括自上而下与自下而上两种方式。自上而下，即领域专家手动构建，准确率高但构建成本较大，且规模难以快速增长；自下而上，即基于海量文本自动化获取，构建成本较低，规模可快速扩充，能够迅速挖掘出海量逻辑，但缺点是精确度受多方面因素影响，准确率较前者要低。事件知识图谱GDELT并没有明确的说自己的构建方式，但基于大数据数据驱动搭建完成的过程还是比较类似自下而上的知识图谱搭建方式。

在大致了解GDELT整体框架后，我简要的将整体GDELT进行一定的分层，如上图所示，自下向上分别是数据来源层，数据库层，解决方案层：

1. GDELT的数据来源非常广，从Wiki百科，新闻报道，到历史信息，书籍，甚至最新推出了观看电视频道的视频分析渠道，在谷歌强大的算法基础上，能够实现实时的对于不同渠道的数据进行收集整理与分析。通过engine对于文本，图像等信息进行处理，提取出事件的时间，地点，人物，事件等主要元素，转变成为两个主要的输入流：GDELT event stream和daily Counts File。事件的构成在这里并不多赘述。
2. 在原数据层数据提取完成以后，大量的知识数据主要存储在数据层的源文件中。为了保证存储事件的完整性和实时性，谷歌的研究员们预先制定了非常完善的数据格式来存储这些知识，极大程度的保证了知识的唯一性和相关性，方便计算机去识别。
3. 在原文件的基础上，GDELT拓展成为Graphs和Dataset两个方面。Dataset中存储了大量事件，记录着世界上各个媒体报道的事件。而Graph主要是以图的形式存储了时间，人物，事件等信息以及他们之间的关系。在途中也是预先定义了许多不同的类别，方便数据的导入。
4. 图数据库最为主要的表现方式是节点和连接节点的关系。在原数据中主要是以graph file的形式存储。知识中的人物，事件等信息自然而然的构成了图谱中的节点，而值得一提的是节点之间的关系的建立。在GDELT中，实体与实体的共同出现次数决定联系的紧密程度，加上GDELT采用了相邻日期捆绑式存储，可以很好的分离同名同姓的人在不同的事件中出现的情况。
5. 事件知识图谱GDELT追求准确性和实时性。实时已经可以通过谷歌的算法实时解析处理媒体数据来完成，而准确性在制定严谨的文件格式的基础上，GDELT还加入了Difference Graph来实现实时的更新。由于有些报道，新闻会存在一定的误差，媒体会对于这些信息进行删帖更新的操作，而通过不断的将同一个数据来源下的的最新的信息与原信息进行对比，数据信息的实时性和准确性都得到了保障。
6. GDELT的一个亮点是它的内嵌式机器学习以及算法系统，在内部直接处理格式化数据能够提高系统的性能，数据的应用面也大大拓展了。GDELT 提供了GDELT Summary和API方便用户使用，在GDELT 2.0推出时在语义情感方面大大增强，也提供了很多emotion package给用户尝试。
7. 如果说GDELT内嵌式的算法能够完成数据的初步分析，Google bigquery便可成为无所不能，在以亿每日为单位的大数据库中，一个高效的信息处理工具必不可少，而Google bigquery中强大的算法在查询和分析方面都能够提供很大的便捷，正如官网上宣传的“GDELT + BigQuery = Query The Planet”
8. 在数据层以及Google bigquery中强大的算法能力的基础上，GDELT一共提供五个主要的解决方案，分别是Situational Awareness，Influencers，Risk Assessment & Global Trends，Policy Reaction，Humanitarian / Crisis Response，在此不做过多的介绍。

GDELT 数据格式分析

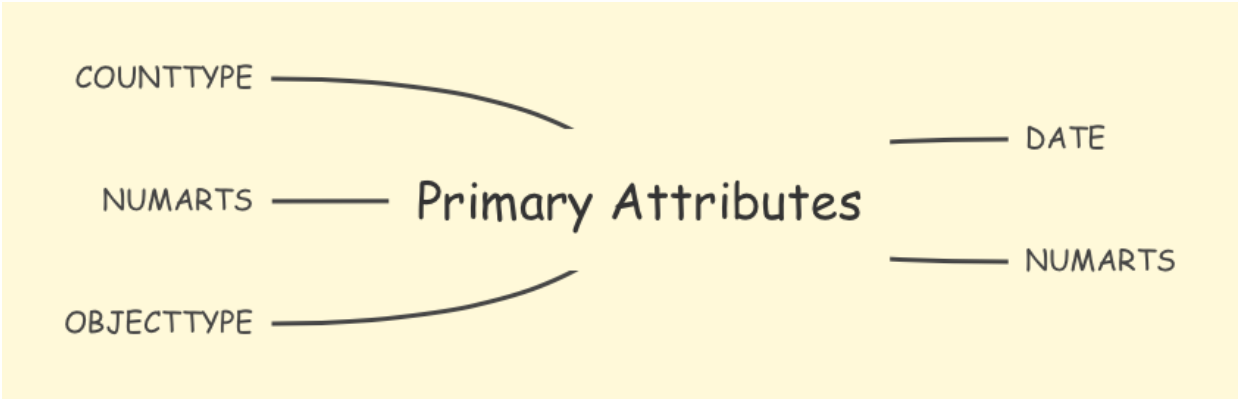
GDELT 1.0 Global Knowledge Graph

GDELT 1.0 知识图谱主要有两个输入流：每日的统计文件以及GKG图形文件。每日的统计文件主要记录了每天信息的收集。通过engine对于文本，图像等信息进行处理，提取出事件的时间，地点，人物，事件等主要元素，以预先定义类别的方式对于这些数据进行收集整理。而GKG图形文件基于每日的统计文件，将事件中包含的所有人员，组织，位置，情感，主题，计数，事件和资源以网络的结构连接在一起。在对数据文件进行分析之后，我用图的形式将分表字段表现出来，而数据文件中的字段我主要分为两类，一类是直接可以得到的，在图的右边部分显示，另外一类是需要进行一定处理后得到的字段，将其放置在左部。

GDELT Count File

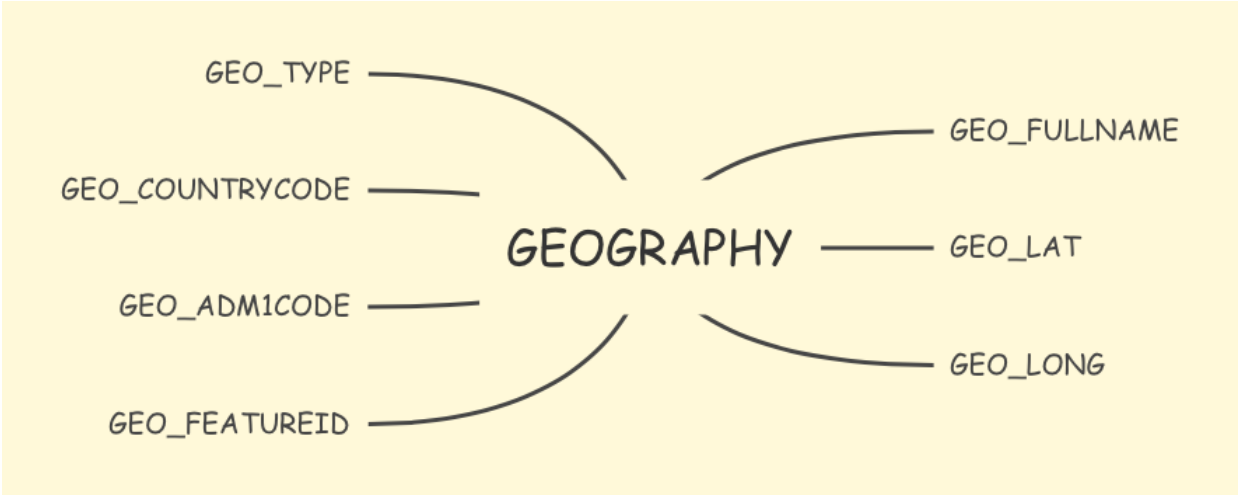
在Count File中，字段主要分为两部分，除去一些基本字段之外，主要是对于地理信息的处理。

Primary Attrubutes



基本字段中，Date，Numarts分别代表了事件发生的日期以及报道的次数。Count File中有一些预先定义好的类别，主要包括“AFFECT, ARREST, KIDNAP, KILL, PROTEST, SEIZE, WOUND“等。当输入事件进行预处理判断后，会对于事件中的一系列数据按照以上类别进行分类，数据统计并记录在CountType，Numarts和OnjectType字段中。

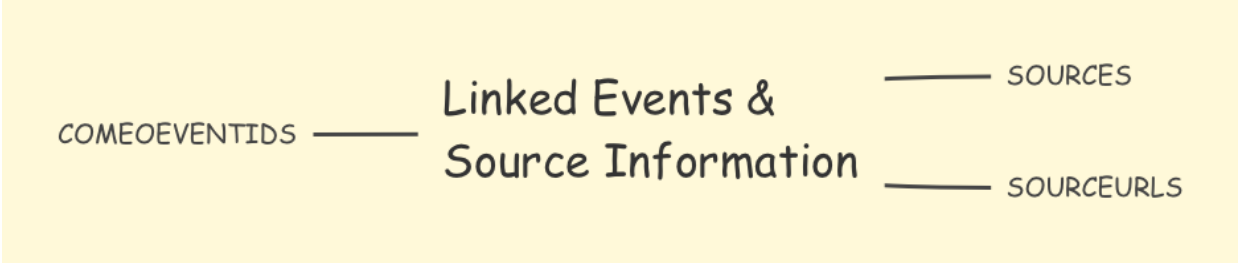
Geography



在地理信息中，位置的全名以及相应的经纬度都可以直接从原数据中得到。对于得到的位置，我们进一步处理得到该位置的类型（城市/国家等）以及相应的country code。 FeatureID的引入帮助我们确定位置的唯一性，防止同一个名称的位置来干扰数据的判断。

在GDELT项目搭建过程中，对于地理信息的处理主要是通过Leetaru教授在2012年提出的全文地理编码器算法，两种CountryCode的引入也是帮助算法的实现，具体的论文在http://www.dlib.org/dlib/september12/leetaru/09leetaru.html。

LINKED EVENTS AND SOURCE INFORMATION



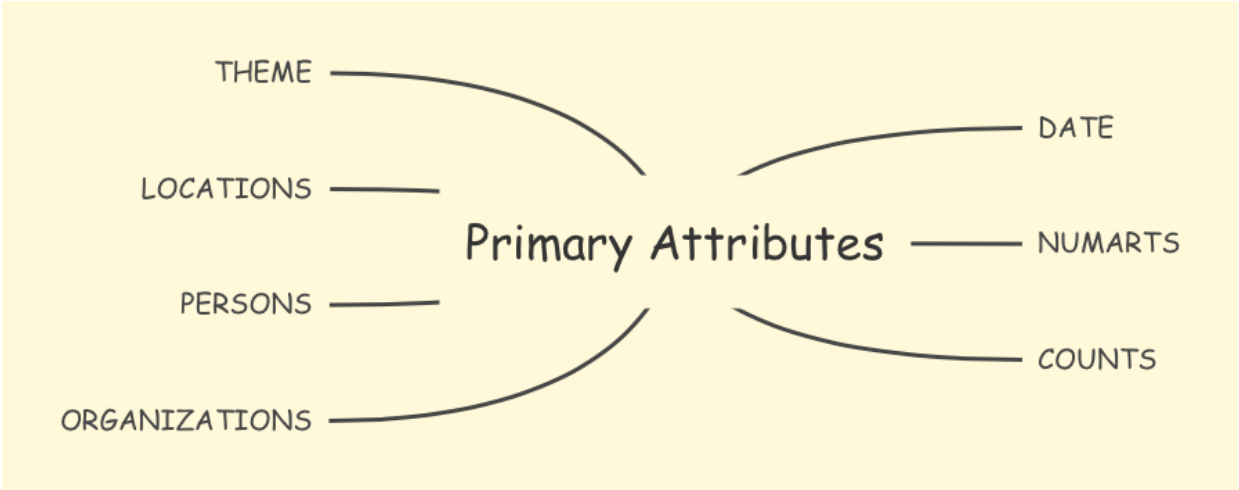
信息来源字段中三个字段皆为列表，值得注意的是comeoeventid，他将同一个类别的事件串联了起来，并记录下了来源和地址，帮助图谱关系的建立。

GKG Graph File

GKG Graph File是在Count File的基础上进行拓展的，与count file类似的，gkg graph file 也有其基本字段字段，地理信息字段被合并入基本字段中，不同的是它还有自己的Emotion相关的字段。

GKG围绕所谓的“Nameset”进行操作，这些名称集本质上是一组名称和其他信息的唯一配对，这些名称和其他信息一起出现在一组文章中。GDELT GKG引擎会汇总所有人员姓名，组织名称，位置，总体情感，提及某件事的列表一组预定义的类别，并将所有文章包含相同的一组人员，组织，位置，数量，事件和主题进行汇总。一组个人名称，组织名称，位置，人数和主题的唯一配对称为“Nameset”。

Primary Attribute

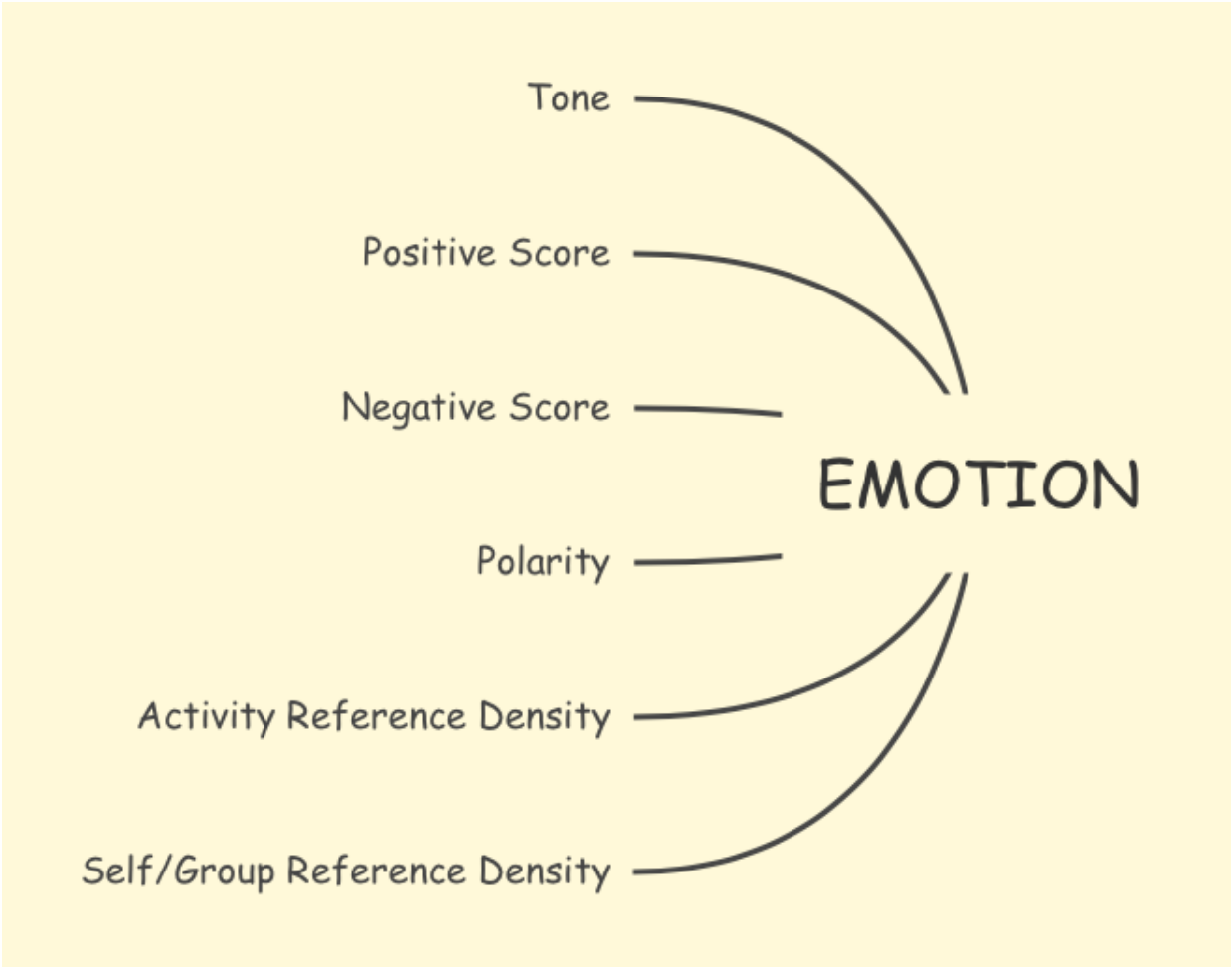


在count file的基础上，gkg graph file新增了Theme，Location，Person，Organization字段，记录的是 GKG引擎汇总事件时，判断所需要的人员，组织，位置，数量，事件和主题等信息。

值得一提的是，这里的Theme值的是关于事件的讨论而不是事件本身的措施等,Theme的产生也是通过预先定义的类别以及文本分析算法的引用来帮助完成的。

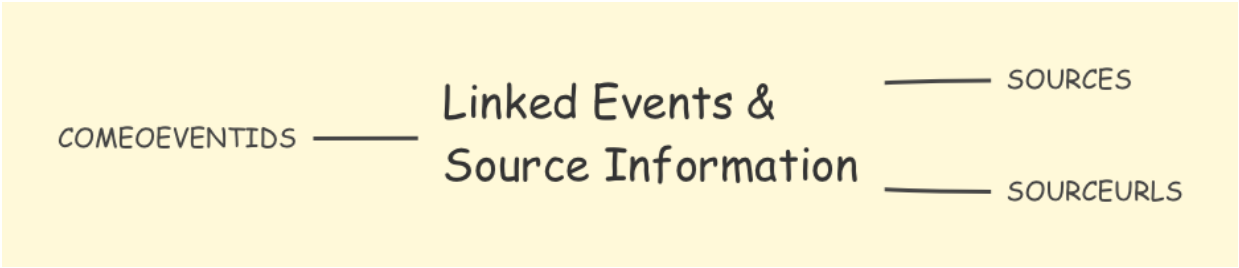
> Themes should be sort of as what the core topics of discussions are around a given entity, rather than an indication of action around the entity.

Emotion



与情感相关的字段主要是TONE字段，主要是由Tone，Positive Score，Negative Score， Polarity，Activity Reference Density 和 Self/Group Reference Density八个指标组合而成。通过提取相关文章中包含感情色彩的词，以算法将其量化，完成相关指标的计算。

LINKED EVENTS AND SOURCE INFORMATION



GDELT 2.0 Global Knowledge Graph

考虑到兼容性等问题，GDELT2.0保存了绝大多数1.0的特性，在存储上也保留了大部分的字段。而在这 基础上，2.0版本为了增强图谱数据分析能力，加入了更多的字段，新特性和能力，其中包括：

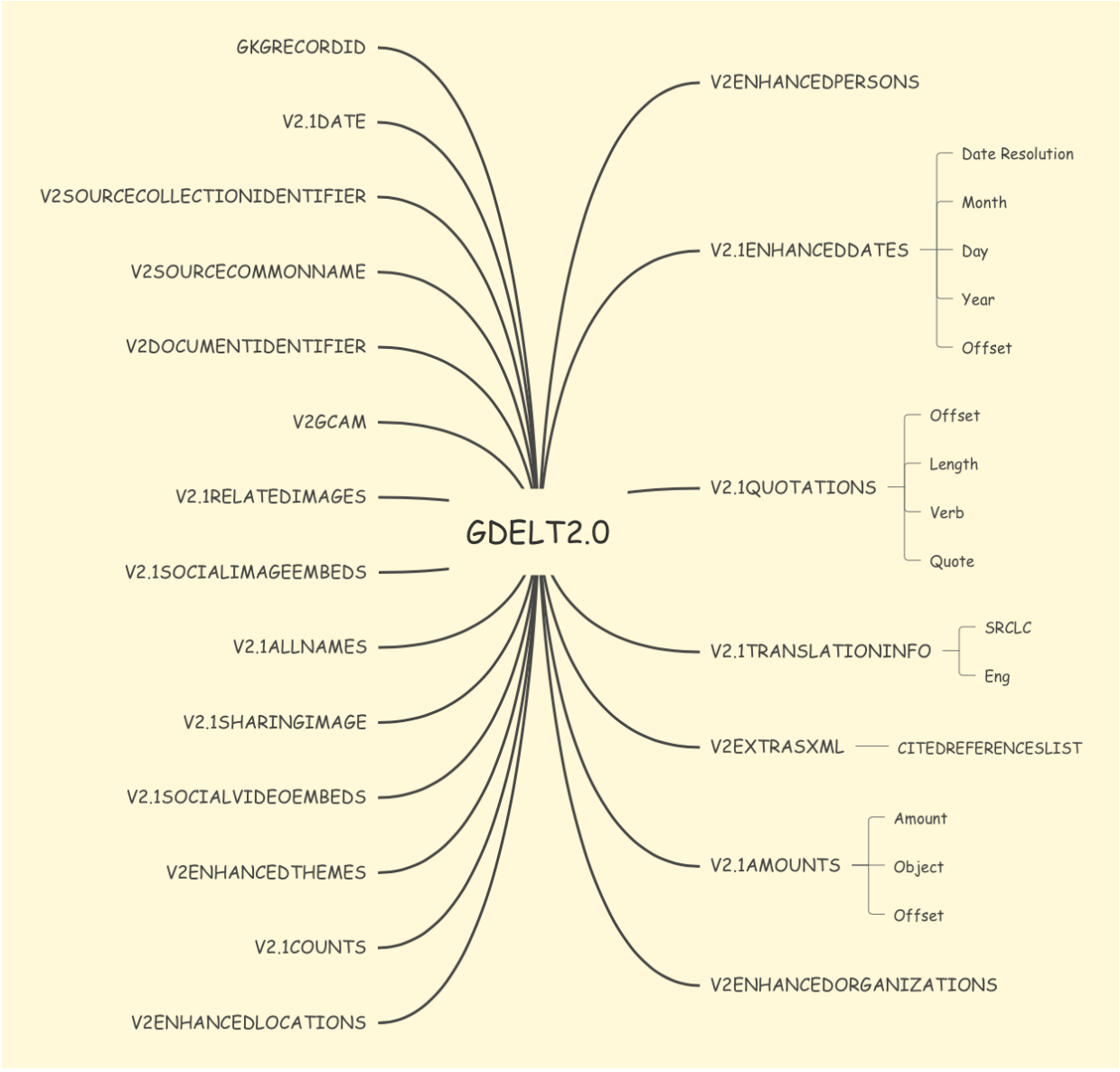
1. 2300种情绪和主题的实时测量
2. 65种语言的实时翻译
3. 嵌入式的图像视频社交分析
4. 100多种新的GKG 主题分类
5. 可拓展的XML语言

等等。

新特性的引入帮助知识图谱拓宽了数据来源，从原来的文章报道，逐步拓宽到照片，视频，甚至电视直播分析，另一方面，更细致的分类增强了图谱的数据分析能力，内嵌式机器学习以及算法系统大大提高系统的性能。

从概念的角度来看，GKG 2.1 / 2.0格式与GKG 1.0之间的两个关键区别围绕条目的聚集方式以及将文章包含在GKG流中的最低标准。在GKG 1.0格式下，所有产生相同GKG元数据的文章组合在一起。因此，将列出同一位置，主题，人物和组织的两组文章放在一起，而NumArticles的值保持为2。而随着新的GCAM系统的推出，该系统可以评估2300多种情绪和主题每篇文章都清楚地表明，发表相同位置，主题，人员和组织的多篇文章可能会使用非常不同的语言来讨论它们，从而产生不同的GCAM分数，**因此在存储的时候也会以多条数据的形式进行存储**。另外，将实时翻译引入GDELT体系结构需要在文档级别识别元数据来源的能力。因此，GKG 2.1不再基于共享的元数据将文档聚集在一起-如果20篇文章都包含相同的提取位置，主题，人员和组织列表，它们将在GKG流中显示为20个单独的条目。但是，每日GKG 1.0兼容性流仍将继续执行群集。除了聚类更改之外，GKG 2.1还更改了名次出现在GKG中的最低标准。根据GKG 1.0和2.0，要求文章至少具有一个成功识别和地理编码的地理位置，然后才能将其包含在GKG输出中。但是，由GDELT监控的许多主题，例如网络安全，宪法讨论和主要的政策讨论，通常都没有很强的地理中心性，许多文章甚至都没有提到一个位置。这从GKG系统中排除了与许多GDELT用户社区高度相关的大量内容。

更改后的数据结构如下：



GDELT Event Database

相比起知识图谱，GDELT的数据库相对比较固定，其字段表示也更为细致，比较符合日常我们使用数据库时满足的规范，但是缺乏了图数据库“Nameset”的灵活性。相比较而然，Event Database更适合做数据的存储，而GKG比较适合做数据挖掘，发现事件之间的关系。

总结

经过本次的探索，我大致了解了GDELT 知识图谱，了解了GDELT中知识图谱的构成要素，比如包含有主题分析、情感分析等等，读懂GDELT提供的事件与知识图谱数据样例，也对于事件知识图谱有一个初步的认识。

在初步接触事件知识图谱中，我感觉事件知识图谱有几个方面是必不可少的：

1. 事件的获取和构建是事件图谱的第一个难点，及时获取并处理大量的数据需要很多的精力。在初步图谱搭建完成后，如何实时获取数据，如何过滤错误数据，保证图谱准确性也是非常重要的方向。
2. 在框架方面，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. 数据的类别定义也需要我们考虑，但是数据类别的定义或许可以通过数据分析而得到，而不是人为去定义。通过大数据的引入，分析，得到的分类可能更符合数据分析，推理的需求。
4. 在数据量逐渐增大的情况下，搜索查询工具的要求也会逐步增加。随着数据量的增加，单主机存储数据方案会逐步转变成为分布式数据存储的形式，搜索查询工具也会随之改变。