基于资讯数据的事件模型

【摘要】基于资讯数据的事件模型，依托于四象限理论框架进行构建，将事件定义为实体域的边际变化，并进一步将事件的构成要素分解为实体域和事件域。该理论描述了如何从实例泛化和抽象到本体域的过程。在实体域和事件域中，模型对实体和事件的关键要素进行了明确定义，并通过事件角色等投影机制实现了实体与事件之间的连接。标准预研的事件模型是对各类事件体系共性的高度提炼，在面向大模型时代人工智能技术的发展，仍能满足对资讯数据的事件化生产和管理的需求，具有推动证券期货业数智化的巨大潜力。

关键词：事件；实体；人工智能

正　文

一、四象限模型

我们创新性提出“四象限模型”，从方法论上提出一套通用的事件定义方法。如图 1 所示，本体域为实例域泛化、抽象而来，比如，“恒生电子”为实体域的实例，抽象为“公司”作为实体本体；“俄乌冲突”为事件域实例，抽象为“地缘冲突”成为事件本体。本体域应涵盖实例层的通用要素。这里讲述的是具体实体（第三象限）泛化为抽象实体（第二象限），具体事件（第四象限）泛化为抽象事件（第一象限）的过程。

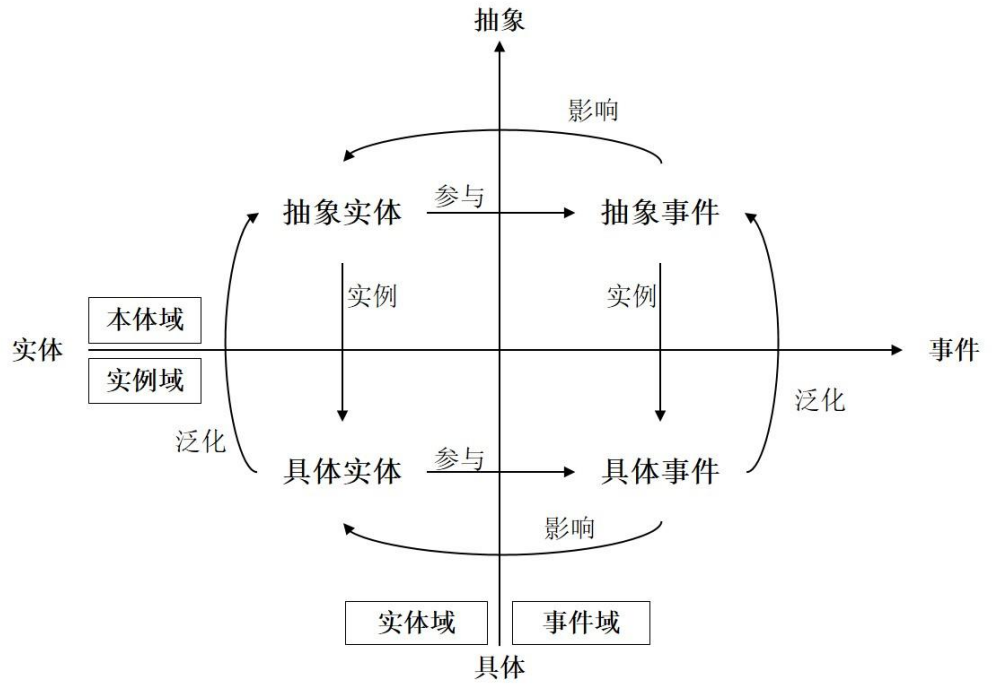
在本体域形成完整的事件模型，用以制定行业性事件标准。“公司”“地缘冲突”，分别作为本体域中实体、事件的代表，在本体域层面，可反向对实体实例、事件实例进行规范指导。也即，公司具有哪些通用的要素，可以参与公司事件的构成，需要在标准中予以定义。与此同时，地缘冲突事件，应具有哪些要素，可以关联实体域，并影响实体域，也需要在标准中予以定义。

事件作为一种驱动，最终是嫁接于宏观分析、行业分析、微观分析、基金分析、衍生品分析等分析框架之上，对具体的金融品种、监管对象产生影响。因此，事件模型中也要求设计资讯数据的关联要素。

事件，在明确原实体域各要素的变化后，可以对实体域

施加影响。如企业发生设立、注销等事件，则可反向对主体所涉及的数据进行变更；如涉及三元组中关系变化的，如恒生电子-任职关系-XXX，则反向对恒生电子为主体的数据进行更改，或在图数据层面予以数据更新。

图 1 基本思想



资料来源：恒生电子 技术平台总部

图 1 中所示四个象限，分别从两个维度绘制：从具体到抽象，从实体到事件。本体域的完备性决定了实例域是否可以广泛使用，反过来，也要求从本体域制定事件标准模型，应充分分析事件实例的特点，抽象其共有属性。

二、事件标准

（一）事件定义

在“四象限模型”思想的指导下，参照 ACE 2005、GDELT 等通用的事件库，以及业内事件体系供应商、学界多要素事件定义（如事件六要素方法、事件七元组方法等），我们提出事件的定义：

实体域的边际变化即为事件。

实体域，在定义上，我们做了一定拓展，定义中，除公司、自然人、产品等实体类型，也将宏观指标、行业等概念化实体纳入其中。常见知识图谱的三元组表达，如“实体-关系-实体”“实体-属性-属性值”均定义在实体域。

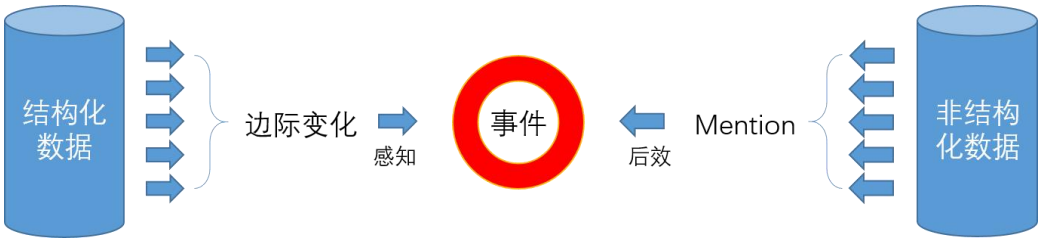
边际变化，体现的是实体的变化、动作、行为等，也可以体现在所隐含的后续变化上。

相比于传统的，以触发词+名词的常见形式之外，定义上也拓展了事件的数据源，也即，既包括文本数据为代表的非结构化数据，也包括公司财务指标、宏观指标、EDB 数据等结构化数据的边际变化。结构化数据的边际变化感知而来，文本数据通过事件抽取的方式，从非结构化数据中而来。

事件抽取：针对文本数据中提及（Mention）该事件，利用事件抽取的方法，从非结构化数据中抽取而来。

事件感知：结构化数据的边际变化形成事件，在事件标准中，我们结合指标域、变化域形成事件。

图 2 事件定义



资料来源：恒生电子 技术平台总部

事件的分类，我们应用笛卡尔积的方法和思路。即，在实体域和业务域联合形成事件的分类。

事件的构成，我们同样采用笛卡尔集的思路，通过结构事件中包含的实体域、事件域要素，形成事件的要素模型。模型所包含的要素分为通用要素、特定要素，通用要素在实体域、事件域的定义中予以呈现，特定要素则在不同业务域中予以定义。

通用要素，如实体、事件发生时间等，在标准中定义了 XX 个通用要素，可进一步拆分为必选要素和参考要素。

特定要素，与业务域高度关联，如在司法类事件中，文号是区别实例事件的核心要素，但文号在分红送股类事件中并不存在，所以，定义文号为司法监管这一业务域中的特定要素。

（二）实体域

金融行业在涉及事件的需求上，相关要素应与金融行业

的实体有明确关联，以满足金融投资、监管、风控业务的实际需求。

事件实体域 ＝{“基础域”:entity，

“地域”：location，

“指标域”：indicator

“时域”：time ，

“路径”：path}

实体域：如公司-时域-指标域的典型，如“恒生电子 2022 年三季度营业收入” “中国 2022 年 10 月 PMI” 构成事件的实体域。

（1）基础域：包含实体类型、实体中文名。将实体类型区分为宏观、政策、市场、产品、公司、自然人。

（2）地域：用于描述实体所在的国别/经济体、地区、地点等，和事件中的发生地点有差别。实际使用中，若缺失可省略。

（3）指标域：指标、变化幅度、变化速度等，

（4）时域：季度、年度、日度、月度、周度、分钟、小时，除此之外，还有一些比如，“1 个月内股价涨幅翻番”“近 2 个月来产品价格持续上涨”“自 1995 年 11 月以来的新高”等等表述

（5）路径：也即，实体域允许有多跳的情况出现。

1. 基础域

我们先看几个例子：

例 1：202X 年 X 月 X 日，MDI 价格上涨 5%。

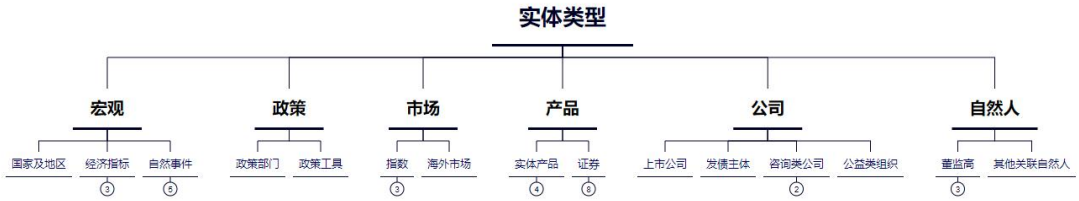
例 2：202X 年 X 月 X 日，恒生电子发布员工持股计划。

例 3：202X 年 X 月 X 日，受 XX 地 6.1 级地震影响，XXXX 公司停产。

例 4：2022 年三季度，中国 GDP 同比增 3.9%。

以上例子都视为事件，实体分别为产品、公司、地区、指标。我们在事件定义上，囊括以上例子中的所有类型，并进行如下分类：

图 3 实体类型



资料来源：恒生电子 技术平台总部

具体的说明如下：

（1）宏观：主要包括国家及地区类型的地域类实体。若宏观事件发生在国内，一般可省略国家作为主体，而直接以指标作为实体的分类。此外，自然事件在常规理解中，偏宏观面，因此划分为宏观实体类事件。

（2）政策：主要包括政策部门、政策工具类事件。政策部门，包括中共中央、国务院、工信部、发改委等部门；

政策工具，则包括 LPR、公开市场操作、财政赤字等。

（3）市场：包括指数类、海外市场类事件，着重强调全市场层面的事件。如上证指数、沪深 300、行业指数等。

（4）产品：

a) 实体产品：主要是企业提供的产品或服务，产品又包括实体产品、概念产品，聚焦于产业、企业上下游产品等。涉及企业所提供的产品或服务的，参考深交所联合深圳证券信息有限公司、国泰君安、国信证券、招商证券、博时基金、恒生研究院等机构牵头形成的对于业务、数据的定义。在品种上，保持与 SDOM 模型的一致性。产品，若国标对于产品的分类过细，可参考产品的概念类型。

b) 金融产品：主要是证券类事件，包括股票、信用债、基金、衍生品等的事件。实体分类，可参照 CFI 编码，并与 SDOM 模型保持一致。

（5）公司：包括上市公司、发债主体、会计师事务所、律所、公益类组织等。参照 SDOM 抽象模型中的主体，涉及股份公司、会计师事务所、律所、证券公司等。

（6）自然人：包括董事、监事、高管，及与公司发生关联的自然人。对于关联自然人不做进一步限定，可以包括白手套、潜在受益人等。

实体基础域，也将与事件的投影相关联。

表 1 实体域与事件域通过事件类型及对应的事件角色关联

事件类型	事件角色
资产重组	交易主体、交易对手、交易标的、监管机构、资产购买方、资产出售方、担保方、会计师事务所、律所、证券公司等
司法监管	处罚机构、当事人、监管对象
宏观经济指标	指标、发布单位
货币政策	政策部门、政策工具
……	……

2.地域

标准拟对实体所属地域进行定义。宏观指标会区别国别，区域经济或政策需要在实体域层面区分地域差异，而部分自然灾害类事件需要区分所发生的地区。

宏观、政策、产品-证券、自然人实体类事件，一般标定好国别/地区，若涉及省/市、县/区的，予以标定；

市场作为实体类事件，可不用设置地域字段；

公司、产品-实体产品作为实体类，若能具象到具体的生产基地、公司所在地是最优的选择；可考虑与静态的知识库进行关联，实现公司与公司生产基地之间关系的关联。若公司有海外生产基地的，也需明确到国别/地区、省/市或具体地点。

3.时域

时域的定义有利于实体域的完善，通过具体时间，如某年、某季、某月、某日，或时间区间来定义，如财报期、近月、近两周等。

例 5：2022 年 10 月 31 日，恒生电子发布 2022 年第三季度财务报告。此例中，“2022 年第三季度”即为时域概念。

在实体域层面进行时域的定义，有利于事件实体域定义的完备。

结构化数据相对便利，也即该指标的时域定义，如 2022 年 11 月 CPI 同比变化这类宏观表达，时域为 2022 年 11 月，用来与 CPI 同比数据联合构成实体域。

非结构化数据中，我们也应注意对时域进行拆分，不然，在面向指标和实体的时候，所提供信息不完备。如 2022 年中央经济工作会议，我们会将其中的政策主题形成事件，一般 2022 年 12 月某日为事件发生时间，而所讨论的是 2022 年的工作总结和 2023 年的经济工作规划，其中的“2022 年”“2023 年”都是时域概念。

4.指标域

指标域指：描述实体变化特性或状态特性的指标，如变化幅度、变化速度等

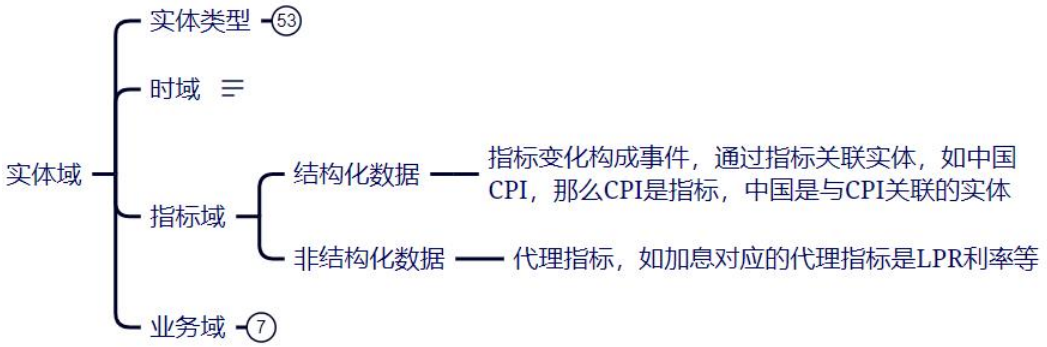
依据事件定义，结构化数据的边际变化也是事件。结构化数据，包括海外宏观经济指标、中国宏观经济指标、EDB

数据、公司财务数据等。

此外，代理指标也归纳为指标域数据。将非结构化文本抽取的事件，与其主体相关联的结构化数据，称之为其代理指标，如“通胀预期走高”对应的代理指标为“CPI”，“原油价格大幅上涨”其代理指标为“布伦特近月连续合约收盘价”。一般来说，代理指标往往在事件生产时会进行预定义，若很难发现对应的代理指标，则可缺失。

指标域指标，可与资讯数据库相应指标保持一致。

图 4 指标域



资料来源：恒生电子 技术平台总部

实体域中，针对指标的变化，存在对统计方式描述的需求，包括累计、TTM、当期、平均、余额、年化等的口径差异。

表 2 统计方式

当期(当月/当季/当周/当日)	平均	同比
累计	最高	环比
TTM	最低	
余额		

年化

5.路径

上文定义，实体本体域是实体、路径、指标域、业务域、时域的模块化组合，其中，可以允许有多个实体在一个事件中：

事件实体域 ＝{“基础域”:entity，
“地域”：location，
“指标域”：indicator
“时域”：time ，
“路径”：path}

其中，路径可简单理解为通过节点、边连接的多跳的情况，也即存在多个节点和边的情况。

例 6：2021 年半年报显示，恒生电子营业收入 20.5 亿元，同比增长 26.5%。我们倾向于以“实体（恒生电子）-属性（营业收入）-值（20.5 亿元）”作为事件定义中第二象限的本体部分，营业收入作为属性，其值必须对应一个具体的时域，在例 6 中，属性及其值所对应的时域为“2021 年半年报”，具体可表达为“2021-06-30”。这里的时域，与事件发生时间/结束时间、信息源披露时间是不同的定义。

例 7：2021 年 9 月 24 日，恒生电子股价上涨 0.51%。此例中，第二象限的本体部分为“实体（恒生电子）-属性（收盘价）-值（54.66 元/股）”，本体部分所涉及的收盘价是必须对应有一个时间的，也即“2021-09-24”。注意，这里的时间

是指收盘价所对应的时间，与事件的时间是不同的，若我们有一个事件“股价变动”，那么此事件若反映的也是收盘价的变动，那么事件发生时间和本体部分属性及其值域对应的时间重合，但从定义上仍然是两个时间。

例 8：2021 年 8 月 26 日，恒生电子提名丁玮为公司独立董事。此例中，第二象限的本体部分为“实体（恒生电子）-关系（任职关系）-实体（丁玮)”，本体部分所涉及的任职关系对应有一个时间，该时间反应的是任职关系的静态知识。而任职关系发生变化，如确立独董和独董任职终止所对应的时间应确立为事件所对应的时间。

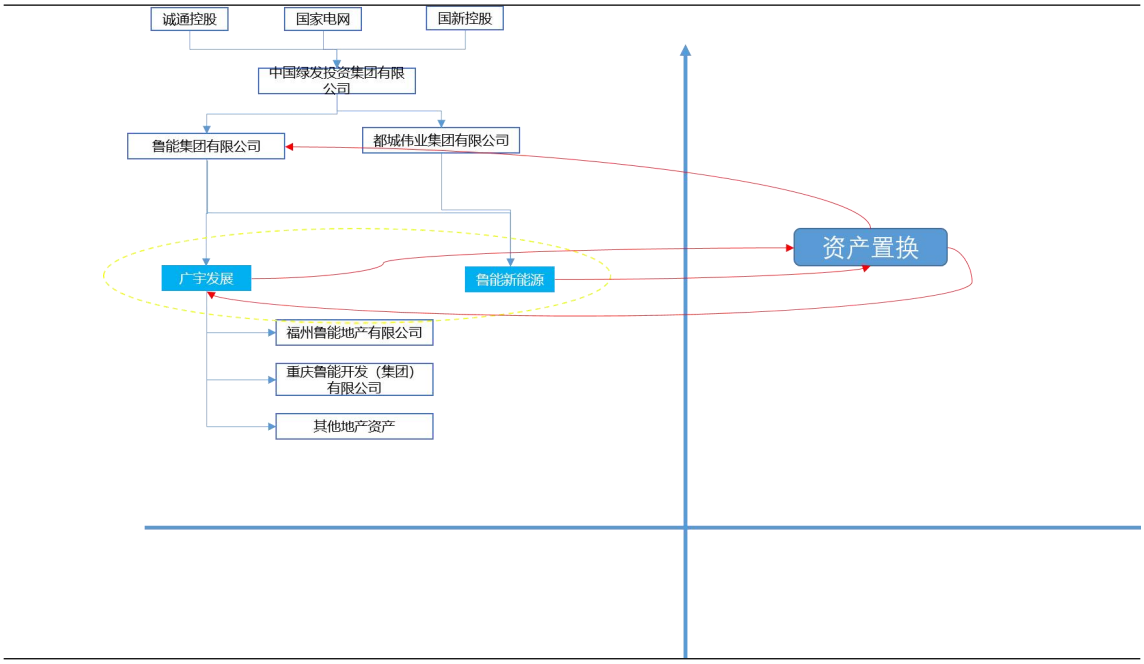
实体及其关系/属性及其值域，在构建成为第二象限的本体域中，可以有多跳情况的存在，多条情况可以通过路径进行描述。路径位于左侧第二象限部分，通过实体、关系的方式，可长可短，可以有多跳，也可以终止于实体，终止于属性值，如例 9 所示的，当中可能会涉及实体本体域的复杂路径。

例 9：2021 年 9 月 6 日，天津广宇发展股份有限公司（以下简称“公司”或“广宇发展”）拟与控股股东鲁能集团有限公司（以下简称“鲁能集团”）、关联方都城伟业集团有限公司（以下简称“都城伟业”）进行资产置换。经初步筹划，拟置入公司的资产为鲁能集团与都城伟业合计持有的鲁能新能源（集团）有限公司 100%股权，拟置出的资产为广宇

发展所持所属房地产公司及物业公司股权等资产负债，差额部分补足方式由各方另行协商确定，但不涉及发行股份。

此案例中，第二象限的本体部分至少涉及如下论元：实体 1（鲁能集团）-关系 1（控股股东）-实体 2（广宇发展）-关系 2（子公司）-实体 3（福州鲁能地产有限公司）……第二象限的本体域当中存在多跳的关系，无论其中结构多复杂，其中发生变化的部分，我们抽象为“资产置换”事件。

图 5 例 9 中涉及的实体域的多跳情况说明



资料来源：恒生电子 技术平台总部

实体为事件的必选论元。依据不同的事件，实体可以层次化、角色化。

前例资产置换事件中，实体包含多个，其中，对实体可以进行层次化、角色化：目标资产 1 广宇发展旗下地产资产，

卖方为广宇发展，买方 1 和 2 分别为鲁能集团和都城伟业；目标资产 2 鲁能新能源，买方 3 为广宇发展，卖方 1 和 2 为鲁能集团和都城伟业。其中，所涉及的多种角色划分，应由静态知识库来承担，参考知识图谱、产业链等成熟的行业标准。事件聚焦于变化/边际变化。

$$Identity = < ibuyer, itarget, iseller, ..., in >$$

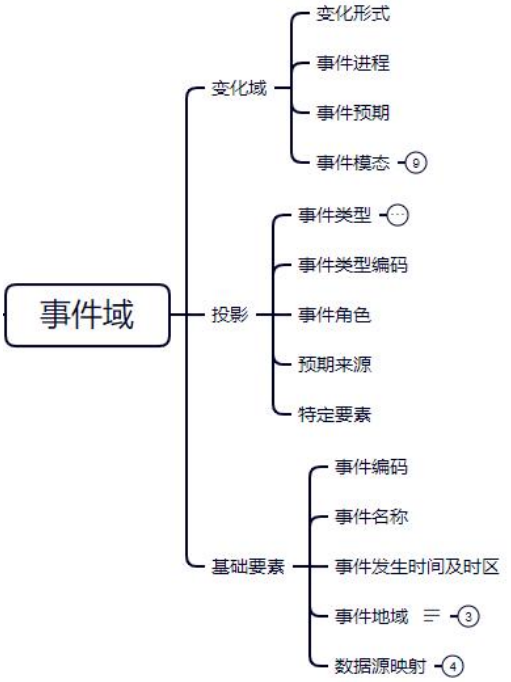
另外，如下几种情况，也分别予以说明：

宏观事件、政策事件的实体可以包括政府组织及相关部门，也可以包括相关金融工具，如利率、汇率、政策等。若能实现与公司关联的，可以与公司进行关联。

（三）事件域

我们将事件相关的数据，分为事件 ID 及时间、地域、数据源、变化域、模态域 5 部分。

图 6 事件域



资料来源：恒生电子 技术平台总部

1.变化域

事件域变化域，包括变化形式、进程、预期、模态四大部分。

表 3 变化域构成

事件域构成	事件域要素	集合属性	必填约束	数据类型
变化形式	变化形式	Y	非必填	VARCHAR
事件进程	事件进程	N	非必填	VARCHAR
事件预期	事件预期	N	非必填	VARCHAR
事件模态	事件模态	Y	非必填	VARCHAR

变化形式，事件的分析方法上，围绕自身历史比较法、横向比较法，所衍生的变化趋势、力度、横向比较趋势，可参考形成正、负、不变的三级定量。

事件进程，定义事件的进程属性，也即事件开始、进展/持续影响、结束三种状态。依据业务域不同，进程的设定也有所差异，司法监管有司法监管的进程类型、定向增发有定向增发的进程。

事件预期，用于区分事件是否是预期要发生的事情。在事件生产时，可着重将预期属性予以标注，也即，存在将、或、拟、可能会等表达时，正常蕴含有预期成份。

事件模态，用于区分事件的主观性判断。此字段为参考性字段，便利于对事件进行定性或定量的分析。模态（Modality）在不同领域有不同的含义。在知识处理领域，模态是指特定的主体针对事件的一种主观的状态。可参考从以下几个类型中新增模态域要素：

（1）是否证实：提供事件发生是否证实的模态说明，用于事件的核实、反馈、删除。

（2）预期模态：与计划有关的模态：比如由动词“计划、准备、打算、想要……”等所刻画的计划状态。

（3）模糊度模态：与认知有关的模态：比如由动词“知道、相信、怀疑、否认、能够、也许、必然、可能……”甚至于引入量化概率等所刻画的认知状态。

（4）情感模态：与情感/评价有关的模态：比如由动词“喜欢、偏爱、讨厌、赞赏、鄙视……”等所刻画的情感/评价状态。

（5）责任模态：与道义/责任有关的模态：比如由动词“应该、必须、得（děi）、不得、允许、禁止……”等所刻画的道义/责任状态。

2. 基础要素

此节，我们定义事件名称、事件编码、时间、进程属性。

（1）事件名称：考虑到大量工业化事件抽取的需求，在命名上，标准保持开放性。建议以“实体+变化域”“实体+动词”的形式进行事件命名。事件命名应遵循通俗易懂、词组规范、符合常识的一般性要求：

a) 通俗易懂：不建议过于文艺的表达，如“产品如火燎原”“大储户储两翼齐飞”等等表达；不建议过于模糊的表达，如“飞龙在天”“或跃在渊”等。

b) 词组规范：如股权变动、通胀预期走强等均可以作为相对规范的命名，不建议出现及、和、等、或等字词；不建议出现字母、数字；不建议出现标点符号。

c) 符合常识：事件命名应符合金融业常识。

（2）事件编码：应考虑采用有意义编码与无意义编码相结合的方式，并保持一定开放性。建议采用“事件时间+编码”的方法。

（3）事件时间：需明确所发生的时区、日期、时间。

a) 时区：涉及事件发生和媒体披露的时区。

b) 日期+时间：对事件及媒体披露时间分别进行识别，日期包含基本的年/月/日，时间包括时/分/秒，并形成类似于“YYYY-MM-DD HH:MM:SS”的表达。

（4）进程属性：我们定义事件的进程属性，也即事件开始、进展/持续影响、结束三种状态。

（5）事件发生地域：事件所发生的地点，涉及国别/经济体、地区、地点，具体地点包括具体等的生产地或对应的具体地点。

a) 宏观、政策、产品-证券、自然人实体类事件，一般标定好国别/地区，若涉及省/市、县/区的，予以标定；

b) 市场作为实体类事件，可不用设置地域字段；

c) 公司、产品-实体产品作为实体类，若能具象到具体的生产基地、公司所在地是最优的选择；可考虑与静态的知识库进行关联，实现公司与公司生产基地之间关系的关联。若公司有海外生产基地的，也需明确到国别/地区、省/市或具体地点。

（6）数据源映射：我们对事件数据源进行定义，采用宏观经济指标、EDB 数据、财务指标等结构化数据感知形成事件的，我们需具备如下数据支持：

a) 用于对事件的源进行映射，保留对所依赖的数据的映射关系，其特点如下：

i. 结构化数据映射：在指标域层面保留映射关系即

可，在此则可为空。

ii. 文本中抽取：会涉及两类来源，若为爬取的文本数据，可考虑保留 URL；若为采购的 text 文本，则保留该文本的 ID。

b) 爬取：事件生产者通过爬取新闻、公告等，然后进行事件生产的。

c) 采购：事件生产者通过所采购的标题、正文等数据进行事件生产的。

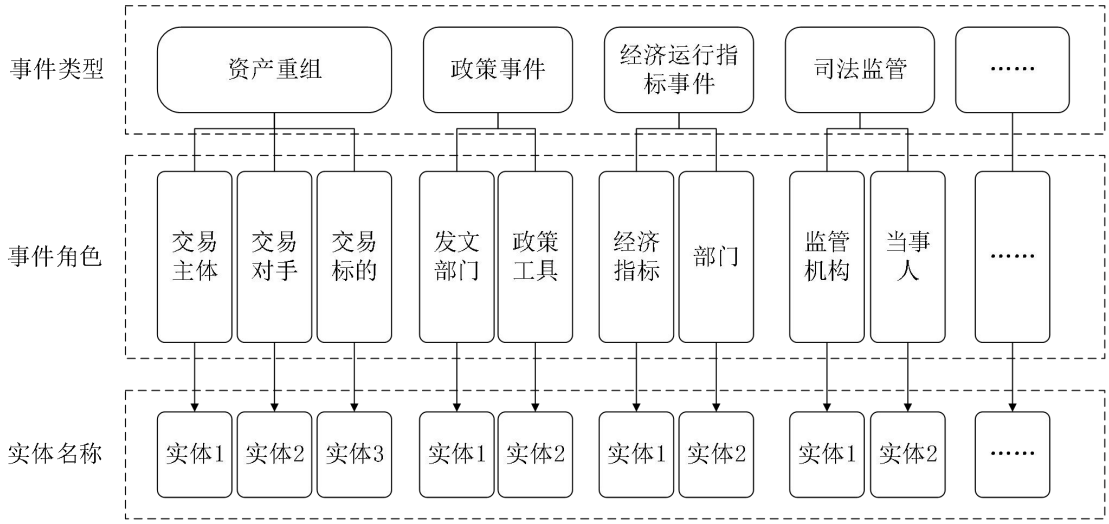
3. 投影

投影，是用来解决事件与实体角色之间关联的问题。

事件类型，是使用资讯数据对实体进行分析的共性事务。同一事件类型，具有共同的模型要素。事件类型可作为事件的分类，也可实体类型构建笛卡尔积形成基于实体的分类。如，针对公司类实体，其并购重组、融资、司法监管、分红派息、财务、信息披露等是跨场景使用者的共性事务。宏观运行指标事件、货币政策事件等宏观类事件，也是使用者的共性分析事务。

事件类型上，设置实体的投影要素。如，并购重组类事件，参与方包括主体、标的、交易对手等实体角色，该实体角色，在事件类型层面予以确定。

图 7 投影



资料来源：恒生电子 技术平台总部

（1）事件类型：标准提供较少争议的事件类型列表，若所生产的事件属于已有类型列表，可参考形成事件的分类。可参考事件类型编码方式进行顺序编码。

（2）事件类型编码：编码方式将按照事件类型名称的中文拼音首字母+数字的形式，保持 4 位字母。若出现重叠，则在数字层面进行顺序编码。

（3）实体角色：基于事件类型，定义事件类型的角色。如经济运行指标事件中，实体角色往往只需要“指标名”；而在货币政策类事件中，则需要有“部门”“政策工具”的实体角色。

（4）预期来源：用于事件域中事件预期与实体的映射。如某证券公司研究所出具的研究报告，对未来要发生的事情

所做的预测，形成事件预期，则保留这一要素与实体做投影。预期来源本质上是一种事件指向实体的关系。

（5）特定要素：用于定义特定事件类型下对要素的需求，如事件“资产重组”对“是否构成重大资产重组”这一要素是必须，用于区分该事件的复杂度、影响度。

4. 案例

表 4 司法监管要素模型

要素名称	要素类型	数据类型	通用或特定	集合属性	必填约束
数据源映射-url	事件属性	VARCHAR	通用	否	必填
文号	关系	VARCHAR	特定	否	必填
事件发生事件/日期	事件属性	DATETIME	通用	否	必填
事件类型	事件属性	STRING	通用	否	必填
事件类型编码	事件属性	STRING	通用	否	必填
事件名称	事件属性	VARCHAR	通用	否	必填
事件编码	事件属性	VARCHAR	通用	否	必填
事件角色-监管机构	关系	VARCHAR	通用	是	必填
事件角色-当事人	关系	VARCHAR	通用	是	必填
实体类型-当事人类型	实体属性	STRING	通用	是	必填
违法事实	事件属性	TEXT	特定	是	非必填
违反条例	事件属性	TEXT	特定	是	非必填
判定结果	事件属性	TEXT	特定	是	非必填

表 5 资产重组要素模型

要素名称	要素类型	数据类型	通用	集合	必填约

			或特定	属性	束
事件编码	事件属性	VARCHAR	通用	否	必填
事件发生时间/首次信息发布日期	事件属性	DATETIME	通用	否	必填
事件类型	事件属性	VARCHAR	通用	否	必填
事件类型编码	事件属性	VARCHAR	通用	否	必填
事件角色-交易主体	关系	VARCHAR	通用	是	必填
事件角色-交易对手	关系	VARCHAR	通用	是	必填
事件角色-交易标的	关系	VARCHAR	通用	是	必填
事件对公司的影响	关系	VARCHAR	特定	否	非必填
是否重大资产重组	事件属性	VARCHAR	特定	否	必填
是否关联交易	事件属性	VARCHAR	特定	否	必填

课题负责人：	白硕	恒生电子	首席科学家
课题成员：	朱星	证通公司	首席技术官
	谢晨	申万宏源	首席信息官
	戴轶	证通公司	数据模型总监
	石宏飞	申万宏源	信息技术开发总部经理
	林金曙	恒生电子	技术专家
	吴斌泉	恒生电子	业务研究员

陈娅	恒生聚源	业务研究员
朱勤勤	丹渥智能	业务研究员
陈佼	恒生电子	技术专家
石洵	恒生聚源	业务研究员
卢长松	恒生电子	开发工程师
褚丽恒	申万宏源	技术专家
徐逸卿	恒生电子	运营专家