编号_____

南京航空航天大学

毕业设计

题　目　网络舆情话题本体半自动构建及进化方案

学生姓名　赵澈

学　　号　091070112

学　　院　经济与管理学院

专　　业　电子商务

班　　级　0910701

指导教师　马静　教授

二〇一四年六月

南京航空航天大学

本科毕业设计（论文）诚信承诺书

本人郑重声明：所呈交的毕业设计（论文）（题目：_____）是本人在导师的指导下独立进行研究所取得的成果。尽本人所知，除了毕业设计（论文）中特别加以标注引用的内容外，本毕业设计（论文）不包含任何其他个人或集体已经发表或撰写的成果作品。

作者签名：　　年　月　日

（学号）：

网络舆情话题本体半自动构建及进化方案

摘　要

随着互联网的发展与普及，对于网络舆情的话题跟踪与分析成为当前的一个热点问题。目前来看，通过舆情信息的不间断积累，引入舆情知识支持，对舆情信息进行深度语义分析，是对突发舆情进行快速合理预控和引导的基础。

本文结合文本挖掘技术、自适应话题跟踪以及本体与本体进化的相关内容，在对网络舆情话题的特征进行分析的基础上，提出了一种网络舆情话题本体（the Internet Public Opinion Topic Ontology, IPOTO）的 OWL 顶层本体结构。在顶层本体的基础上进行了话题初始本体的半自动构建，以及基于本体学习的话题本体进化的方案设计，其中包括话题模型和事件模型的建立、语料的选择及其文本的预处理、语料事件文本信息的抽取、事件-话题相似度与实例-话题相关度的算法以及人工修正方法等。最后基于方案进行了实验系统的开发，并选用“马航 MH370”话题的真实语料对方案进行了实验。

关键词：网络舆情，本体进化，文本挖掘，话题跟踪，OWL

Structure and Semiautomatic Evolution of the Internet Public Opinion Topic Ontology

Abstract

The tracking and analysis of the Internet public opinion topic has become a hot topic. This paper combines the related contents of text mining technology, adaptive topic tracking and ontology development and evolution, after the research of the Internet public opinion topic's characteristics, propose a Internet Public Opinion Topic Ontology(IPOTO)'s top-level structure with OWL. On the basis of IPOTO's top-level structure, design a set of program of semiautomatic initial ontology construction and ontology evolution for IPOTO, which contains the establishment of the topic model and the event model, selection and pre-processing of the text corpus, event text information's extracting, the event-topic similarity and the individual-topic relevance algorithm, and artificial correction method. Finally, develop an experimental system and conduct an experiment using real corpus of the topic "Malaysia Airlines Flight 370".

Key Words：Internet Public Opinion; Ontology Evolution; Text Mining; Topic Tracking; OWL

目 录

第一章 绪 论

1.1 选题背景及意义

随着互联网的发展与普及，对于网络舆情的话题跟踪与分析成为当前的一个热点问题。由于网络舆情具有直接性、突发性和偏差性，如何尽早关注可能产生重要影响的事件报道、消息发布，及时发现潜伏期的舆情话题，跟踪舆情的增长情况及话题内容的演化，成为舆情管理的核心挑战。近年围绕着网络舆情管理的相关研究日益增多，尤其是信息技术公司，迅速开发出系列网络舆情监控产品。如北大方正技术研究院推出的方正智思舆情预警辅助决策支持系统[1]；复旦大学媒体计算与 Web 智能实验室的“互联网舆情分析跟踪系统”[2]；北京拓尔思信息技术股份有限公司的 TRS 互联网舆情管理系统[3]；Autonomy 网络舆情聚成系统[4]；军犬网络舆情监控系统[5]、九瑞网络舆情监控系统[6]；东蓝科技舆情监测系统(ES-Focus)[7]；Goonie 互联网舆情监控系统[8]等等。然而从以上网络舆情监控产品相关技术介绍看，其舆情分析普遍停留在简单的网页词频统计，对网页主题没有语义层面上的把握，造成相关网页统计数据的严重失真。此外，网络舆情的汇集和分析机制的指标体系欠缺，语义内容遗漏和偏差现象严重，使得这些产品的可用性及实际功效遭到质疑[9]。网络舆情话题的深度内容处理，涉及网页结构技术、语义识别技术、内容与表现形式动态变化的跟进技术，这是一相当复杂的难题。因此，在网络舆情分析中切入语义要素，科学的多方融合一些学科领域在文本挖掘、知识发现、机器学习等相关方面的成果，探索深度舆情分析的实用方法，是舆情监管的唯一出路。

在此背景下，本文引入语义本体技术，尝试从管理视角下的舆情任务为框架指导，半自动地构建舆情话题本体，在舆情话题的不断动态演化中，实现舆情本体的进化，最终实现网络舆情的动态自适应话题跟踪，从语义高度实现舆情的动态监管。

1.2 国内外相关研究现状

1.2.1 国外相关研究现状

可用于舆情分析的核心技术当属“话题识别与跟踪技术”，其代表是美国的 TDT（Topic

Detection and Tracking，话题识别与跟踪）系统[10]。TDT 主要研究针对新闻文本的五个任务：报道切分(Story Segmentation)、新报道识别（New Event Detection)、关联识别(Story Link Detection)、话题识别(Topic Detection)、话题跟踪(Topic Tracking)，其中话题跟踪任务（Topic Tracking）的主要是指跟踪已知话题的后续报道[11][12][13]。话题跟踪任务（Topic Tracking）是 TDT 评测中比较重要的任务，评测中参加单位使用的方法在原有信息检索方法的基础上进行改进，从训练报道中抽取特征集作为话题特征，当新报道到来时，如果与话题特征匹配得好，则判定为话题相关，否则被判定为不相关。使用的方法大致包括：向量检索和概率检索[14]、最近邻分类、神经网、Boosting Bayes 分类器、决策树、动态聚类和支持向量机等[15]，在建立模型时使用了实体名词识别、词特征向量、TF-IDF 权重计算、打分规范化、文本扩展、无监督自适应、多种方法组合等方法。具备自学习能力的自适应话题跟踪已逐渐成为 TT 领域新的研究趋势[16][17]，而传统仅基于统计的策略不能真实地描述其语义空间，因此涉及内容表示的话题识别与跟踪研究开始受到关注[18][19]。自适应话题跟踪任务(Adaptive Topic Tracking，ATT)研究成果主要包括基于内容和基于统计的方法。在基于内容的 ATT 相关研究中，GER&D 尝试采用文摘技术跟踪话题的发展趋势，其核心思想是分别提取话题与报道的文摘代替全文描述，其缺陷是跟踪系统没有嵌入自学习机制，话题模型没有利用检测到的后续相关报道自适应地进行更新[20]。Dragon 和 Umass 分别尝试了无指导 ATT 研究，其跟踪系统每次检测到相关报道，都将它嵌入话题模型并改进特征的权重分布，后续报道的相关性则以新生成的话题模型为评估对象，从而实现跟踪系统的自学习功能[21][22]。但是这两种方法并没有很大程度地提高话题跟踪系统的性能，其主要原因在于自学习模块对于跟踪反馈不施加任何鉴别地全部用于话题模型的更新，而系统反馈本质上是一种伪反馈[23]，即同时包含相关报道和不相关报道，因此学习过程将大量不相关信息也嵌入话题模型，从而导致话题漂移[24][25]。基于这一现象，LIMSI 在原有自学习过程中嵌入二次阈值截取功能，通过设置一个比阈值更高的过滤指标，截取伪反馈中相关度较高的报道嵌入话题更新模块，从而削弱了话题漂移[26]。

1.2.2 国内相关研究现状

国内针对自适应话题跟踪方法的探索也取得了丰富的研究成果。东北大学的王会珍等提出了基于反馈学习的自适应方法，其基本思想是采用增量式方法对话题追踪模型进行修

正[27]。哈尔滨工业大学的郑伟等针对话题本身的漂移现象，基于改进的相关性模型，将话题核心与新颖部分分离，对跟踪中伪相关反馈包含的新颖信息进行检测和建模，并在此基础上动态调整话题空间，改进了话题跟踪的效果[28]。清华大学的贾自艳等借鉴 Single-Pass 聚类思想，结合新闻要素提出一种基于动态进化模型的话题检测与跟踪算法 [29]。金珠引入了知网这个语义资源，利用从知网管理系统导出后的知识词典，在语篇情感计算角度实现了对话题相关信息的组织[30][31]。焦健提出并实现了一种基于知网的报道特征规范化的话题跟踪算法，采用知网知识库求得两个词语之间的相似度，并根据相似度对话题特征进行规范[32]。宋丹将语义引入话题跟踪中，提出了一个用语义框架对话题和报道进行表示的方法，通过对报道中的时间、地点、人物等信息进行识别和抽取以及建立语义框架槽，如面向地点类名实体建立地理树，匹配过程基于两名实体在地理树中路径的覆盖率进行计算，并设计对应的权重计算方法[33]。张辉等在文章中分析了传统文档向量空间模型的不足，结合新闻报道的特征，提出了一种三维文档模型，将每篇报道分解为新闻标题特征、内容特征、实体特征三个维度，用三个向量分别表示三个维度上的文档特征，同时采用增量学习的方法对话题模型进行动态修正，使之能够自动适应话题的演变[34]。朱恒民等关注到网络环境因素，提出了基于链接网络图的舆情话题跟踪方法[35][36]。刘炜等基于前人在 TDT 中对语义矢量的相似性计算研究，以及本体和语法结构在文本相似性研究方面的应用成果，提出了以词频分析作为辅助手段，将新闻中的关键要素归纳为时间、空间、参与事件的主客体、行为等几个语义类，同时借助 WordNet 与本体技术计算文档特征词的相似度，并且结合文本的语法结构特点，共同应用于文本的相似度计算，并以此作为新事件检测中相似度计算的基础[37]。

1.2.3 研究现状问题总结

以上研究中存在的比较突出的问题可以归结如下：

（1）对消息报道的针对性研究不足：消息报道和一般文本在书写风格、内容、结构等方面有较大差异。而且，话题形成之后仍然动态演化，使话题描述词汇在短时间内变化剧烈，这些都给话题表示带来了较大困难，有必要对通用文本表示方法进行改进或构建专门的新闻表示知识（本体）模型；

（2）消息报道来的快，去的快，提供给话题分析的已知信息较少：信息量的不足给话

题的表示、发现和追踪都带来了较大困难，使许多对训练数据有需求的方法变得不那么有效，由此造成的错误在后来的数据流处理中也很难得到修正，适当的知识支持不可或缺；

（3）消息报道流中话题个数不确定，未知话题的内容不可预测：传统的许多经典方法在这个动态环境中基本无法开展。

1.3 本文的研究内容及技术路线

1.3.1 研究内容

本文主要研究网络舆情话题的本体表达及话题本体的构建与进化方法，从而得到一个能够在一定程度的人工监督下自适应地跟踪某特定网络舆情话题的系统。

本文的研究内容可以大致分为五个部分：

（1）介绍网络舆情分析、话题跟踪及其相关关键技术的研究意义以及国内外研究现状，并提出其中存在的问题以及本体相关理论和技术在此领域可能的利用方式；

（2）对舆情分析、文本挖掘、话题跟踪、本体进化的相关基础理论进行研究；

（3）分析网络舆情话题跟踪中需要关注的要素信息类型；

（4）进行网络舆情话题本体的构建与进化算法的方案设计，包括模型建立、文本新信息抽取和相似度算法等；

（5）进行实验系统的开发，并使用真实的网络舆情话题语料进行话题本体的构建与进化实验，并对结果进行分析。

1.3.2 技术路线

本文的技术路线如图 1.1 所示：

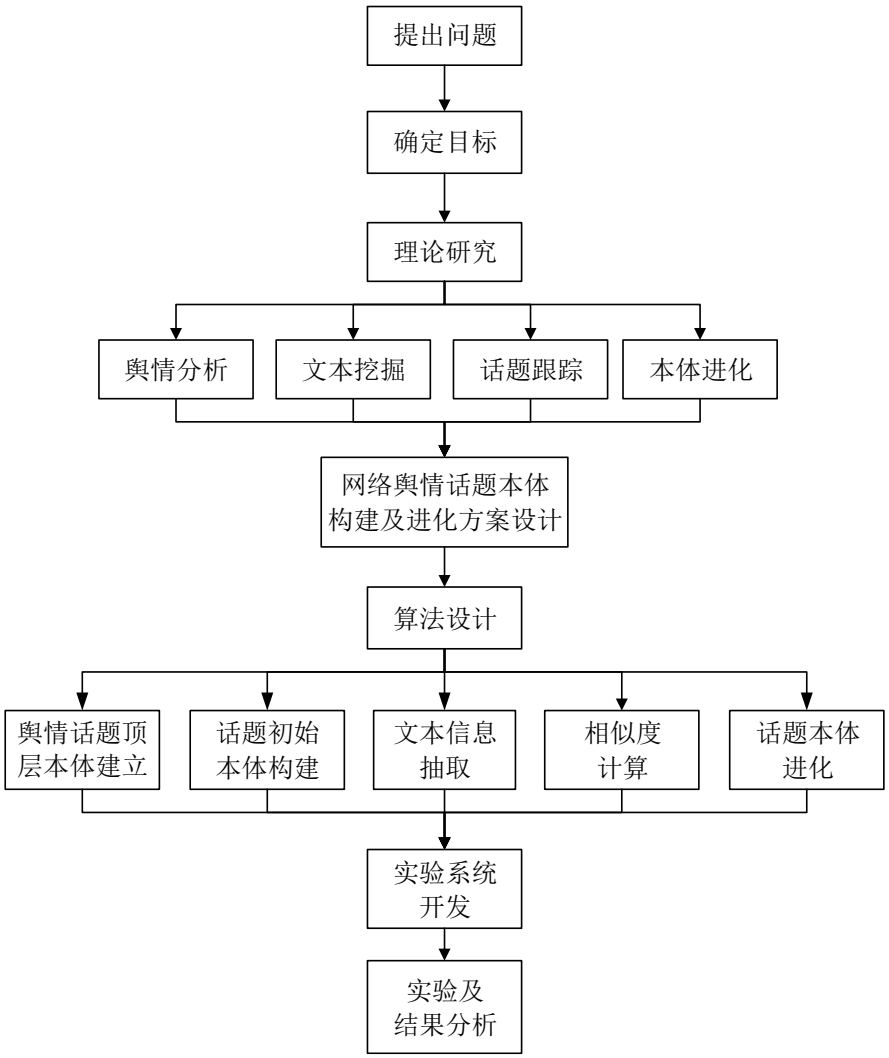

提出问题
确定目标
理论研究
舆情分析
文本挖掘
话题跟踪
本体进化
网络舆情话题本体
构建及进化方案设计
算法设计
舆情话题顶
层本体建立
话题初始
本体构建
文本信息
抽取
相似度
计算
话题本体
进化
实验系统
开发
实验及
结果分析

图 1.1 技术路线

第二章 相关基础理论研究

2.1 网络舆情的定义与性质

目前对于网络舆情的定义与性质，学界仍未达成一个有效的共识。刘毅认为网络舆情是通过互联网表达和传播的，公众对自己关心或与自身利益紧密相关的各种公共事务所持有的多种情绪、态度和意见交错的总和[38]。姜胜洪认为网络舆情热点是网民思想情绪和群众利益诉求在网络上的集中反映，是网民热切关注的聚焦点，是民众议论的集中点，反映出一个时期网民的所思所想[39]。曾润喜认为网络舆情是由于各种事件的刺激而产生的通过互联网传播的人们对于该事件的认知、态度、情感和行为倾向[40]。

2.2 文本挖掘技术

2.2.1 中文分词技术

词是最小的能够独立活动的有意义的语言成分。在汉语中，词与词之间不存在分隔符，词本身也缺乏明显的形态标记，因此，中文信息处理的特有问题就是如何将汉语的字串分割为合理的词语序列，即汉语分词。汉语分词是句法分析等深层处理的基础,也是机器翻译、信息检索和信息抽取等应用的重要环节。目前主流的中文分词系统包括 SCWS、FudanNLP、ICTCLAS 等，其中由中国科学院计算技术研究所研发的汉语词法分析系统 ICTCLAS 具有不错的应用效果，其主要功能包括中文分词、词性标注、命名实体识别、新词识别，同时支持用户词典、支持繁体中文，支持 gb2312、GBK、UTF8 等多种编码格式[41]。

2.2.2 词频排序分析

词频排序分析是一类对文本中的词出现的频率进行各种统计从而判断该词是否属于可代表文本主题的主题词或特征词的文本统计分析方法。最简单的词频排序算法就是统计同一篇文档中同一个词出现的次数占文档内所有词的比例，也就是所谓的 TF（Term Frequency）值，一般认为某个特定词的 TF 值越高，该词在文档中的语义重要程度越高，然而该方法的缺点是没有考虑一些通用词的情况，即在各种文档中出现的频率都较高的词，如作为连词“的”、“和”或者动词“是”、“做”。包括 TF-IDF 在内的更高级的算法则避免

了这个问题，但是由于单纯的词频排序分析并未引入文本内各词本身的语义特征，因此具有较大的局限性。

2.2.3 关键词共现分析

关键词共现分析是一种研究同一篇文档或多篇文档中一个或多个关键词共同出现的特征，从而挖掘关键词之间的相关性或关键词所在文档之间的相关性的统计分析方法。一般认为，若两个特定的关键词在多篇文档内都共现，说明两个关键词之间具有相关性。关键词共现分析常常与词频排序分析结合使用，比如由词频排序分析得到可代表某文档主题的多个特征词，再对这些特征词在其它文档中的共现频次进行统计，从而挖掘该文档的相关文档。

2.3 自适应话题跟踪

2.3.1 自适应话题跟踪

自适应话题跟踪 ATT 的一般跟踪过程如图 2.1 所示：

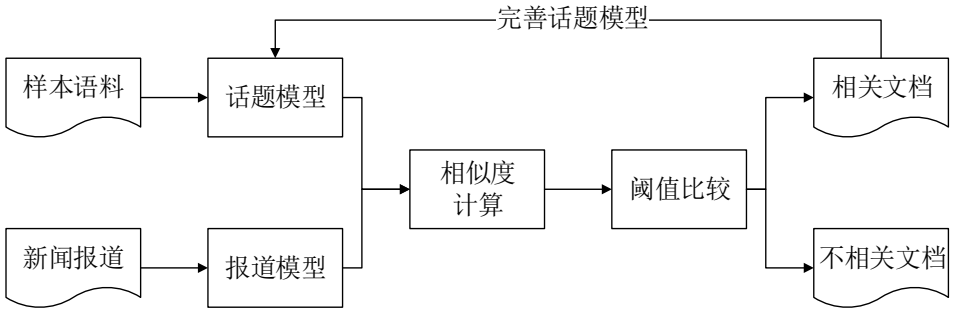


图 2.1 自适应话题跟踪过程

自适应跟踪可以划分为两种，即有指导的自适应跟踪和无指导的自适应跟踪。前者基于人工相关反馈进行学习过程和话题模型的优化和更新，后者则完全无人工参与。

2.3.2 话题/报道模型建立

话题跟踪的话题模型和报道模型的建立，最常使用的模型表示方法有向量空间模型和概率模型两种。以下详细介绍向量空间模型（Vector Space Model，VSM）。

向量空间模型是目前话题跟踪领域的主流表示模型，其原理是将对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。

该模型将文档 *d* 映射为一个特征向量：

$$V(d)=(w_1(d)t_1,w_2(d)t_2,\ldots,w_n(d)t_n)$$

其中 *ti(i=1,2,⋯,n)*为一列互不相同的词语（在信息检索研究领域被称为索引项），*wi(d)* 为 *ti* 在文档 *d* 中的权值，一般被定义为词语 *ti* 在 d 中出现频率 *tfi(d)*的函数值，常用的如词语 *ti* 的 *TF-IDF* 值，计算公式如下：

$$TF_i(d)=\frac{\text{词语}t_i\text{在文档}d\text{中出现的次数}}{\text{文档}d\text{的总词数}}$$

$$IDF_i=\log_{10}\left(\frac{\text{文档总数}}{\text{出现词语}t_i\ \text{的文档数}+1}\right)$$

$$TF-IDF_i(d)=TF_i(d)*IDF_i$$

根据 TF-IDF 公式，文档集中包含某一特定词语的文档越多，说明它区分文档类别属性的能力越低，其权值越小；另一方面，某一文档中某一特定词语出现的频率越高，说明它区分文档内容属性的能力越强，其权值越大。

通常需要对报道词集进行停用词过滤后，得到向量特征选取需要的候选集合，此外由于许多文档中的词数较多，一般还会先计算各词的 TF-IDF 值，选择值最高的前 n 个词语作为可代表该文档的特征词。对于新闻报道短小及话题漂移问题，向量空间模型多采用信息扩充的方法解决[42]。

2.3.3 计算模型相似度

相似度计算的方法和所使用的表示模型具有较大关系，需要结合模型本身特点并充分利用表示模型的内容。在相似度计算方面，TDT 使用的大多是一些已有的计算方法。下面介绍向量空间模型的余弦相似度计算方法：

两文档 *di*，*dj* 间的相似度可以看作其对应的特征向量之间的夹角的余弦值，公式如下：

$$sim(d_i,d_j)=\frac{\sum_{k=1}^{n}w_k(d_i)*w_k(d_j)}{\sqrt{(\sum_{k=1}^{n}w_k{}^2(d_i))(\sum_{k=1}^{n}w_k{}^2(d_j))}}$$

其中，*wk(di)*和 *wk(dj)*分别为关键词 *k* 在文档 *di*，*dj* 的向量中的对应权值。

2.3.4 阈值比较

判断某个新报道是否是给定话题的相关报道，通常的做法是把新报道和话题进行比较，如果相似度高于某个阈值，则把该报道标识为话题的相关报道。

而阈值的确定则没有统一的标准，通常情况下都是在实验过程中，根据多次实验的结果确定阈值的大小。

2.4 本体相关理论及技术

2.4.1 本体简介

本体(Ontology)最早是哲学领域的一个概念，后来被引用到计算机界的人工智能、数据库、自动推理、知识工程等领域中。目前被普遍接受的本体的定义是由 Studer 提出的，即“本体就是共享概念模型的明确的形式化规范说明”。该定义包含四层含义：概念化、明确化、形式化和共享性[43]。

本体的突出特点是它规范化描述特定领域的概念或术语，建立不同领域间的共享的概念体系，为这些领域的实际应用提供基础支持。目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇（术语）和词汇之间相互关系的明确定义。本体在面向机器的知识组织、揭示语义联系和知识共享等方面均可有所作用。

Perez 等人用分类法组织了本体，认为本体可以表示成由五个基本的建模元语组成的五元组 O=（C，R，F，A，I），其中 C、R、F、A 和 I 分别表示类或概念（Classes/Concepts）、关系（Relations）、函数（Functions）、公理公理（Axioms）以及实例或个体（Instances）[44]。

2.4.2 本体的构建

本体的构建采用最多的是手工的方法，手工方法主要在领域专家的帮助下，手工获取本体中的概念、属性和规则。手工构建本体的过程大致包括需求分析、计划制定、领域分析、本体设计、本体实现和本体完善等内容[45]。这种方法的由于有领域专家的参与，因此对概念、属性定义比较准确，但大量的人工降低了效率，并且对于一个复杂领域而言提取的信息较难做到没有遗漏。

目前大多数本体的构建一般仅基于各自领域和具体工程的考虑，因此方法上具有不小的差异。Gruber 于 1995 年提出的 5 条较为通用的构建本体的规则具有较大的影响：清晰

（Clarity）、一致（Coherence）、可扩展性（Extendibility）、编码偏好程度最小（Minimal Encoding Bias）以及最小约束性（Minimal ontological Commitment）[46]。

各国研究者已开发出许多本体编辑软件工具，不同的工具所支持的本体描述语言不尽相同，大致可以分为两类：第一类包括 Ontolingua、OntoSaurus、WebOnto 等，这三个工具都基于某种特定的语言，并在一定程度上支持多种基于 AI 的本体描述语言；第二类包括 Protégé、WebODE、OntoEdit、OliEd 等，这些工具独立于特定的语言，可以导入/导出多种基于 Web 的本体描述语言格式（如 XML、RDF/RDFS、OWL 等），具有良好的可扩展性[47][48]。

2.4.3 本体的进化

本体进化（Ontology Evolution）是指根据实际的应用需要，在现有本体的基础之上，依据相应的理论、方法及标准，对本体的概念、结构及关系不断进行完善和改进的过程与方法。需要注意以下两点：构建本体不是一劳永逸的，初始本体构建后需要不断添加新的概念，不断完善本体概念的关系，不断丰富本体的实例，以满足实际的应用需求；本体进化不仅包括本体概念及关系的数量上的增加，还包括本体概念（关系）的删除、修改。

总的来说，本体进化是本体的不断完善，根据实际应用需求，对本体的概念、关系进行增加、修改、删除。由于本体是一个综合性的整体，概念之间、关系之间、概念与关系之间联系紧密，一个地方的改动也会引起其他地方的变化，因此本体的进化是一个系统性的工作，需要考虑到本体各个部分的数据一致性[49]。

2.4.4 本体学习技术

本体学习（Ontology Learning）的研究目标是利用机器学习和统计等技术自动或半自动地从已有的数据资源中获取期望的本体，由于实现完全自动的知识获取技术还不现实，所以目前的本体学习过程主要是指在用户指导下进行的一个半自动的过程；纯文本以及 XML、HTML、DTD 等都可以作为本体学习的数据源，根据数据源的结构化程度，本体学习技术可被分为三大类：基于结构化数据的本体学习技术、基于非结构化数据的本体学习技术和基于半结构化数据的本体学习技术，目前已存在的一些本体构建工具（如 Hasti、OntoLearn、Text-To-Onto、OntoLIFT、OntoBuilder）各自使用了一些不同的本体学习技术[50]。

2.4.5 网络本体语言 OWL

网络本体语言 OWL（Web Ontology Language）是一种定义和实例化网络本体的语言，

旨在提供一种可用于描述网络文档和应用之中所固有的那些类及其之间关系的语言。其功能在于为网络文档和应用中固有的类以及其间的逻辑关系提供描述，使得基于此技术的网络应用更加人性化和智能化，节省用户自身资源搜索时间并将这些处理交给计算机系统内部处理[51]。OWL 网络本体语言家族大致分为两个系统共三种子语言：基于描述逻辑进而丰富表达和精准计算属性的 OWL Lite 和 OWL DL，以及以资源描述架构 RDF（Resource Description Framwork）提供兼容叙述的 OWL Full。在表达能力和推理能力上，每个子语言都是前面的语言的扩展。

OWL 中主要包含的基本元素有类（classes）、属性（properties）、类的实例（instances/individuals of classes）以及实例间的关系（relationships between instances）。在此需要一提的是，OWL Full 与前两种子语言的一个重要区别是，它可以定义类的类（classes of classes）而前两者不行。

类可以具有子类和父类，并可通过 unionOf 等构造符（construct）操作多个类从而产生新的类；属性是一种二元关系（binary relation），属性的定义中至少需包括属性名、定义域（domain）和值域（range），反映了两个域之中的概念间关系。属性分为数据类型属性（datatype properties）和对象属性（object properties）两种，其中前者是类的实例与数据类型间的二元关系，而后者是来自两个类的实例间的二元关系（属性的定义域和值域是类，但实际取值要落到每个实例）。属性可具有子属性作为比原属性更加具体的属性。

以下 OWL 语句定义了一个名为 madeFromGrape 的对象属性，其定义域和值域分别是葡萄酒 Wine 和酿酒葡萄 WineGrape（Wine 和 WineGrape 是两个已经定义的类，定义语句在此省略），这表示葡萄酒具有属性 madeFromGrape，且属性值的范围是酿酒葡萄的范围：

```
<owl:ObjectProperty rdf:ID="madeFromGrape">
  <rdfs:domain rdf:resource="#Wine"/>
  <rdfs:range rdf:resource="#WineGrape"/>
</owl:ObjectProperty>
```

OWL 中属性的值域可被用作推理具有该属性的概念所属的类，如在以上语句的基础上定义一个实例 LindemansBin65Chardonnay 如下，则可推知 LindemansBin65Chardonnay 是属于 wine，因为它具有 madeFromGrape 属性，而该属性的定义域是 Wine：

```
<owl:Thing rdf:ID="LindemansBin65Chardonnay">
```

```
    <madeFromGrape rdf:resource="#ChardonnayGrape" />
</owl:Thing>
```

在属性上还可定义属性特征（property characteristics）以及属性约束（property restrictions）。此外类和实例可分别通过属性 equivalentClass 和 equivalentProperty 定义相等的类和实例。

第三章　网络舆情话题跟踪的要素分析

3.1　网络舆情话题的特征

3.1.1 网络舆情话题的发起和传播

传播平台：网络舆情话题的传播需要依托于互联网平台，主要的平台包括各大新闻门户网站如新浪新闻、腾讯新闻，社交网络平台如新浪微博、QQ 空间、天涯论坛、虎扑论坛，以及一些个人或组织博客。

传播媒介：网络舆情话题承载于具体的文本、图像、音频、视频等媒介，其中占最大比例的媒介是文本。所以文本也是舆情话题跟踪最优先关注的媒介。

传播趋势：网络舆情话题的传播在一开始往往速度较慢，产生的相关信息很少，但是一旦话题成为热点事件，其相关新闻报道以及网络舆论会出现非常爆发性、病毒性的增长和快速传播，接下来话题将开始出现漂移，最终话题将沉寂下来。因此舆情话题常常呈现较强的时效性，对于话题文本的发布时间是需要重点关注的。社交网络平台和博客上的话题信息由于受到的管理和控制较弱，传播速度往往要高于新闻门户网站。

3.1.2 网络舆情话题文本的平台差异

网络舆情话题的文本由于来源的传播平台不同，具有一定的差异。

从语法的角度分析：

通过新闻门户网站发出和传播的话题相关文本，其用语较为官方化和书面化、语句格式较为整齐；

通过社交网络平台（如论坛、微博）、个人博客等发出和传播的话题相关文本，其用语较为个人化和口语化，语句格式较为零乱且常出现错别字。

从语义的角度分析：

（反映用语习惯、格式、语义话题集中度、平台可信度、二级词性等）

通过新闻门户网站发出和传播的话题相关文本，其内容一般紧扣话题本身、语义集中度高。此外该类文本的语义更为客观，即使是如社论性质的文本，主观性也较弱，在该类

文本中带有倾向性和感情色彩的词汇的使用数量较少。该类文本的语义往往较为完整，内容一般可以满足新闻报道的 6W 要素划分（时间、地点、人物、起因、经过、结果）。前方已经分析到，由于此类话题文本的发布和传播受到较强的管控，虽然传播的速度下降，但也使得文本内包含的信息具有更高的可信度。

通过社交网络平台、个人博客等发出和传播的话题相关文本，其内容则常常出现偏离话题的情况，即使文本中出现了话题的特征词，整个文本的语义重点也并非是话题本身，而是分散或延伸到话题以外的地方。此外，文本的内容往往只满足 6W 要素划分的一部分，部分要素会有缺失，但在这之外增加了较强的观点性和主观倾向性（情感倾向也是网络舆情话题跟踪中需要关注的与一般新闻话题跟踪不相同的内容）。由于受到的管控较弱，此类文本内包含的信息的可信度相对较低，常常带有许多错误和偏差。

两类话题文本的总体对比如表 3.1 所示：

表 3.1 两类话题文本的对比

话题文本来源平台	文本语法特征	文本语义特征
新闻门户网站	官方化 书面化 语句格式整齐	语义集中、客观、完整、可信度较高
社交网络平台、博客	个人化 口语化 语句格式零乱	语义分散、主观、残缺、可信度较低

3.2 网络舆情话题跟踪的相关概念

基于话题跟踪和本体等相关理论，在此提出一些本文将使用到的网络舆情话题跟踪的相关概念。

事件：指由某些原因引起的、在特定的时间和地点发生的、并伴随某些结果的一个特例，描述相关话题的某个或某些方面。

网络舆情话题（简称舆情话题或话题）：由一个种子事件或活动以及与其直接相关的事件或活动组成。包括一个核心事件或活动，以及所有与之相关的事件或活动。

语料事件：指通过某种规则选择的用以跟踪某特定网络舆情话题的线上事件，一个语料事件包含的信息主要由其语料事件文本所承载，比如一篇网络新闻报道便可认为是一个语料事件，语料事件的信息主要承载于该报道所在网页内的原始文本中。

语料事件文本：指语料事件的原始文本经过预处理去除与话题跟踪无关的干扰信息、进行一定的内容和编码格式规范化后的文本。

3.3 网络舆情话题跟踪的要素信息类型

在网络舆情话题的特征分析基础上，借鉴新闻报道的 6W 要素分类方法，结合专家经验，得到网络舆情话题跟踪所需要关注的六个要素类型：事件要素、时间要素、地点要素、实体要素、描述要素、谓语要素。以下对其分别进行说明：

（1）事件要素

指承载一个事件的原始文本本身，以及该文本的发布平台、发布时间、发布人。

（2）时间要素

指事件中包含的具体时间点。

（3）地点要素

指事件中包含的具体地点。

（4）实体要素

指事件中包含的人物、组织机构和其他客观物体，它们是一个话题中最值得关注的对象，是事件内所有行为的发出者或接受者，一般在文本中作为名词出现。

（5）描述要素

指事件中具有反映和补充实体特征以及文本发布者的观点、态度、情感倾向的作用的要素信息，一般在文本中作为形容词、副词等出现。

（6）谓语要素

指事件中的实际行为和变化，一般在文本中作为动词出现。谓语要素能够反映事件中发生的各种动态变化，以及实体要素之间的关系[51]。

第四章　网络舆情话题本体半自动构建及进化方案设计

4.1　网络舆情事件/话题信息模型构建

4.1.1 事件信息模型

在网络舆情话题的跟踪要素分析基础上，参考自适应话题跟踪中的 VSM 报道模型，构建出语义化的网络舆情事件信息模型（简称事件模型），用以表示一个事件的事件信息。该模型表示为七个特征向量组成的特征向量：

$$V\left(EventInfo\right)=\begin{pmatrix} w(Event)*V\left(Event\right),w(Time)*V(Time),w(Location)*V(Location), \\ w(Entity)*V(Entity),w(Description)*V(Description), \\ w(Predicate)*V(Predicate),w(Extra)*V(Extra) \end{pmatrix}$$

其中的七个特征向量分别代表某事件文本中的一类要素信息：事件要素信息、时间要素信息、地点要素信息、实体要素信息、描述要素信息、谓语要素信息和扩展要素信息，前六类要素信息在本文前一章已加以说明，扩展要素信息是指在事件文本信息抽取时无法被标准化到前六类要素信息中，但对于事件来说仍具有一定特征语义的要素信息。由于一个话题事件信息模型对应的是一个事件，因此特征向量 *V(Event)*内部的向量空间如下式：

$$V\left(Event\right)=\mathrm{w}t$$

w 取 *1*，*t* 为某事件的事件要素信息。

时间要素信息与话题事件信息模型是 *n:1* 的对应关系：

$$V\left(Time\right)=(w_1\left(Time\right)t_1,w_2\left(Time\right)t_2,\ldots,w_n\left(Time\right)t_n\ )$$

其中 *ti(i=1,2,...,n)*为一列互不相同的时间要素信息，权重 *wi(i=1,2,...,n)*在本文中设为对应的时间要素信息在事件中的 *TF* 值（*EventTF*）。

EventTF 的计算发生在事件文本信息抽取过程中，详细计算方法见§4.2.1。

其他要素信息的向量空间表示与时间要素信息类似。

事件模型的各要素信息内部包含一个或多个数据项，具体说明见§4.2.2 至§4.2.9。

4.1.2 话题信息模型

网络舆情话题信息模型（简称话题模型）用以表示一个特定话题的信息。该模型参考自适应话题跟踪中的 VSM 话题模型，表示为七个特征向量组成的特征向量：

$$V(TopicInfo)=\begin{pmatrix} w(Event)*V(Event), w(Time)*V(Time), w(Location)*V(Location), \\ w(Entity)*V(Entity), w(Description)*V(Description), \\ w(Predicate)*V(Predicate), w(Extra)*V(Extra) \end{pmatrix}$$

其中的七个特征向量分别代表话题的一类要素信息：事件要素信息、时间要素信息、地点要素信息、实体要素信息、描述要素信息、谓语要素信息和扩展要素信息，要素信息实例均来源于话题的相关事件。由于一个话题事件信息模型对应的多个相关事件，因此特征向量 *V(Event)*内部的向量空间如下式：

$$V(Event)=(w_1(Event)t_1, w_2(Event)t_2, \ldots, w_n(Event)t_n)$$

其中 *ti(i=1,2,...,n)*为一列从属于不同相关事件的事件要素信息，在本文中权重 *wi(i=1,2,...,n)*暂设为 *1/n* ，即认为话题各相关事件的事件要素信息权重相等。

特征向量 V(Time)内部的向量空间如下式：

$$V(Time)=(w_1(Time)t_1, w_2(Time)t_2, \ldots, w_n(Time)t_n)$$

其中 *ti(i=1,2,...,n)*为一列互不相同的时间要素信息，权重 *wi(i=1,2,...,n)*在本文中设为对应的时间要素信息在话题中的 *TF* 值（*TopicTF*）。

TopicTF 的计算发生在初始话题本体构建和话题本体进化过程中，详细计算方法见 §4.4.3 和 §4.5.6。

其他要素信息的向量空间表示与时间要素信息类似。

话题模型的各要素信息内部包含一个或多个数据项，与事件模型相同。

4.1.3 事件/话题信息模型的区别

事件信息模型用以储存某一语料事件对应的要素信息，一个事件信息只与一个语料事件相关；话题模型用以储存某一话题对应的要素信息，一个话题信息与多个话题相关事件相关。

4.2 事件文本信息半自动抽取方法

4.2.1 总体流程

事件文本信息半自动抽取是指由自动方法主导、人工调整辅助，对某个话题跟踪语料事件对应的语料事件文本（获得方法见§4.3 和§4.5）进行处理，并得到事件、时间、地点、实体、描述、谓语、扩展等七种类型的事件的要素信息，其本质是一种文本挖掘过程。总体流程是，先对语料事件文本进行事件信息抽取和时间信息抽取，得到的待分词文本经过 ICTCLAS 系统分词与词性标注（计算所二级词性标注），得到分词结果文本。接下来对分词结果文本进行其他五种要素信息的抽取。所有抽取完毕后得到的要素信息，由人工进行合并词、简写词、同义词的补充等修正，导入事件信息模型，得到该事件的事件信息。最后将人工补充的信息词添加到 ICTCLAS 系统的用户词典，从而优化分词效果。

总体流程如图 4.1 所示：

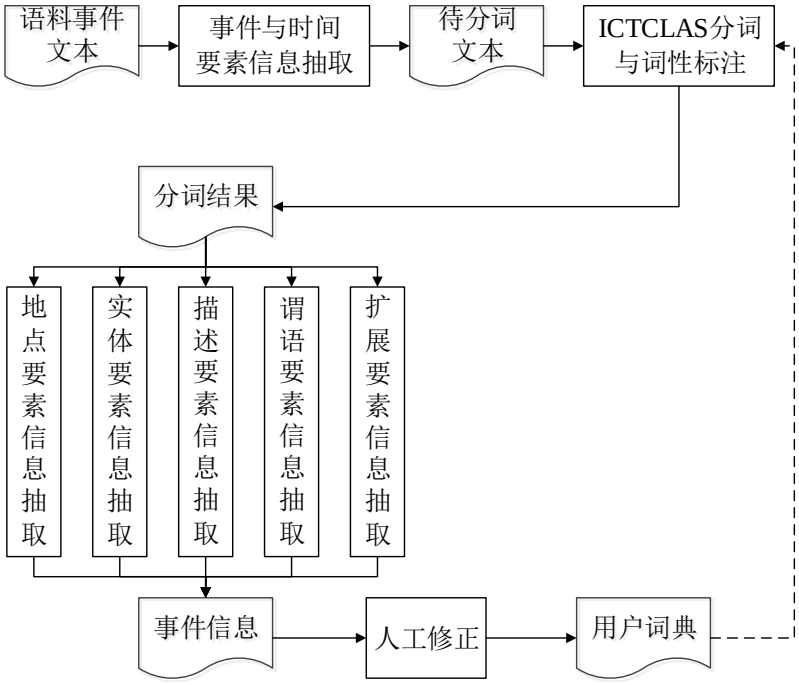


图 4.1 事件文本信息半自动抽取总体流程

在进行对一个事件的各类要素信息（事件要素信息除外）抽取时，均需计算要素信息实例的权重值 *w*，在本文中，一个属于某特定类型的要素信息实例 *i* 的权重值（即 EventTF）计算方法如下：

$$EventTF_i = \frac{Count}{TotalCount}$$

其中 *TotalCount* 是该类型的所有要素信息实例的在该事件语料文本中被匹配的总频次，*Count* 为实例 *i* 在该事件语料文本中被匹配的频次。这里的匹配频次对于本次抽取流程中的事件语料文本内出现的多个相同要素信息实例进行累计。比如“飞机”一词在文本中被匹配共 3 次，那么其对应的 Count 值为 3 而不是 1。

4.2.2 事件（Event）要素信息抽取

完整的事件要素信息包括：事件序号（eventID）、发布平台（fromPlatform）、发布人（fromPublisher）、发布时间（fromTime）、语料事件文本（originalText）。采用直接从语料事件文本中获取的方法进行抽取。

4.2.3 时间（Time）要素信息抽取

完整的时间要素信息包括：年（year）、月（month）、日（day）、时（hour）、分（minute）、秒（second）、时间具体程度（accurateLevel），以及权重 w。

本文采用多条正则表达式逐个匹配的方法提取时间要素信息，从最精确的表达式开始匹配，每匹配一个表达式后，将匹配项中的数字以字符串格式分别放入年、月、日、时、分、秒之中，并计算得到该时间信息的时间具体程度和 EventTF 值（权重值），之后删除语料文本中的所有匹配项，接着进行下一个表达式的匹配。时间要素信息抽取结束将得到待分词文本。具体使用的正则表达式如表 4.1 所示（考虑到秒的价值不大，因此并未进行正则匹配）：

表 4.1 时间要素信息匹配正则表达式

序号	正则表达式	时间具体程度
1	[0-9]{4}年[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点][0-9]{1,2}分	5
2	[0-9]{2}年[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点][0-9]{1,2}分	5
3	[0-9]{4}年[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点]	4
4	[0-9]{2}年[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点]	4
5	[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点][0-9]{1,2}分	4
6	[0-9]{4}年[0-9]{1,2}月[0-9]{1,2}[日,号]	3
7	[0-9]{2}年[0-9]{1,2}月[0-9]{1,2}[日,号]	3

表 4.1(续)

序号	正则表达式	时间具体程度
8	[0-9]{1,2}月[0-9]{1,2}[日,号][0-9]{1,2}[时,点]	3
9	[0-9]{1,2}[日,号][0-9]{1,2}[时,点][0-9]{1,2}分	3
10	[0-9]{4}年[0-9]{1,2}月	2
11	[0-9]{2}年[0-9]{1,2}月	2
12	[0-9]{1,2}月[0-9]{1,2}[日,号]	2
13	[0-9]{1,2}[日,号][0-9]{1,2}[时,点]	2
14	[0-9]{1,2}[时,点][0-9]{1,2}分	2
15	[0-9]{4}年	1
16	[0-9]{2}年	1
17	[0-9]{1,2}月	1
18	[0-9]{1,2}[日,号]	1
19	[0-9]{1,2}[时,点]	1
20	[0-9]{1,2}分	1

匹配举例：“2014 年 03 月 08 日 00 点 42 分”被正则表达式 1 所匹配，得到的时间要素信息为{“2014”,“03”,“08”,“00”,“42”,5}；“3 月 9 日”被正则表达式 9 所匹配，得到的时间要素信息为{“”,”3”,”9”,“”,“”,“”,2}。

该匹配方法可匹配使用不同时间量词（如“3 月 8 日”和“3 月 8 号”），不同时间位数（如“3 月 8 日”和“03 月 08 日”），不同具体程度（如”2014 年 3 月”和”2014 年 3 月 8 日”）的文本，但存在以下一些问题：

（1）年份位数标准化

“2014 年”和“14 年”的实际语义相同，但没有经由位数的标准化进行统一；

（2）中文数字匹配

类似“三月八日”的文本无法被匹配为时间要素信息，其对应的分词结果文本是“三月/t 八日/t”，按照§4.2.9 的扩展信息抽取规则，“三月”和“八日”将分别被存入扩展要素信息；

（3）“分隔符+数字”格式匹配

类似“03:42”的时间格式无法被匹配为时间要素信息，其对应的分词结果文本是“03:42/m”，按照§4.2.9 的扩展信息抽取规则，该信息将不被考虑；

（4）模糊时间匹配

类似“今晚”、“凌晨”、“昨天下午”的模糊时间无法被匹配为时间要素信息，对应的分词结果文本是“今晚/t”和“昨天/t 下午/t”，按照§4.2.9 的扩展信息抽取规则，“今晚”、“凌晨”、“昨天”和“下午”将分别被存入扩展要素信息。

对于前 3 类无法正则匹配的时间信息，可在必要时（特别是在进行话题初始本体构建时）由人工进行识别、标准化并补充进事件的时间要素信息中。对于第 3 类模糊时间文本，视为扩展要素信息即可。

4.2.4 ICTCLAS 分词与词性标注

时间匹配结束后，待分词文本由 ICTCLAS 系统进行分词与词性标注。

以下是腾讯一则新闻的原始文本：

据新华视点消息，记者从中国民航局空管局了解到，马来西亚航空公司 B777（9MMRO）型飞机，执行 MAS370 吉隆坡至北京航班任务，起飞时间 8 日 00:42（北京时）。该机 01:20 在胡志明管制区同管制部门失去通讯联络，同时失去雷达信号，经向相关管制部门联络证实，该机一直未与我国管制部门建立联络或进入我国空管情报区。

该原始文本经过 ICTCLAS 分词和计算所一级词性标注的结果文本如下：

据/p 新华/n 视点/n 消息/n ，/w 记者/n 从/p 中国/n 民航/n 局/q 空管/v 局/n 了解/v 到/v ，/w 马来西亚/n 航空公司/n B777/x （/w 9MMRO/x ）/w 型/k 飞机/n ，/w 执行/v MAS370/x 吉隆坡/n 至/p 北京/n 航班/n 任务/n ，/w 起飞/v 时间/n 8 日/t 00:42/m （/w 北京/n 时/n ）/w 。/w 该机/r 01:20/m 在/p 胡志明/n 管制区/n 同/p 管制/v 部门/n 失去/v 通讯/n 联络/v ，/w 同时/c 失去/v 雷达/n 信号/n ，/w 经/p 向/p 相关/v 管制/v 部门/n 联络/v 证实/v ，/w 该机/r 一直/d 未/d 与/p 我国/ns 管制/v 部门/n 建立/v 联络/v 或/c 进入/v 我国/ns 空管/v 情报/n 区/n 。/w

该原始文本经过 ICTCLAS 分词和计算所二级词性标注的结果文本如下：

据/p 新华/nz 视点/n 消息/n ，/wd 记者/n 从/p 中国/ns 民航/n 局/qv 空管/vn

局/n 了解/v 到/v ，/wd 马来西亚/nsf 航空公司/n B777/x （/wkz 9MMRO/x ）/wky 型/k 飞机/n ，/wd 执行/v MAS370/x 吉隆坡/nsf 至/p 北京/ns 航班/n 任务/n ，/wd 起飞/vi 时间/n 8 日/t 00:42/m （/wkz 北京/ns 时/ng ）/wky 。/wj 该机/r 01:20/m 在/p 胡志明/nrf 管制区/n 同/p 管制/vn 部门/n 失去/v 通讯/n 联络/vn ，/wd 同时/c 失去/v 雷达/n 信号/n ,/wd 经/p 向/p 相关/vn 管制/vn 部门/n 联络/vn 证实/v ,/wd 该机/r 一直/d 未/d 与/p 我国/ns 管制/vn 部门/n 建立/v 联络/v 或/c 进入/v 我国/ns 空管/vn 情报/n 区/n 。/wj

可以看出，一级词性标注中，“中国”被直接标注为“/n”，在 ICTCLAS 官方文档中的解释为名词，而在二级词性标注中，“中国”被标注为“/ns”，解释为地名。要将标注的词性和本文中划分的各要素信息（事件与时间除外）相对应，一级词性标注的粒度是远远不够的，因此在本文中采用计算所二级词性标注。

一些二级词性的对应的词语义作用较小，比如代表标点符号的“/w*”（“。”）、代表介词的“/p”（“至”），代表动词“是”的“/shi”等，或者语义较为混淆如代表数词的“/m”（“00:42”）和代表字符串的“x”（“9MMRO”），因此在以下的要素信息抽取中暂时未加以考虑，被当做停用词对待。

分词的效果可以通过用户词典进行优化，如“马航”最初的分词结果是“马/n 航/v”，不符合预期结果，因此可在用户词典中添加“马航”一词，对应词性为 nt（机构团体名）。

4.2.5 地点（Location）要素信息抽取

待分词文本在经过分词与词性标注之后，开始进行地点、实体、描述、谓语和扩展要素信息的抽取。

地点要素信息包括地点词（location）、父级地点（parentLocation）、地点具体程度（accurateLevel）以及权重 w。父级地点能够使得后期的信息匹配和相似度计算得到优化，比如事件中出现”南京”，那么可以认定匹配到本体信息中的“江苏”，此外由地点父子层级关系可以得到地点具体程度值，并将此值用于相似度加权。但此处存在两个难点，其一是父级地点的确定（可能方法是人工处理或参考已有的地点信息数据包），另一个是地点总层级的不确定性，因此在本文的抽取过程中父地点不作考虑。抽取方法为：寻找分词结果中对应词性标注的词作为地点要素信息的地点词，并计算其 EventTF 值（权重值）。具体的

对应关系如表 4.2 所示：

表 4.2 地点要素信息对应词性

ICTCLAS二级词性	词性解释
ns	地名
nsf	音译地名

4.2.6 实体（Entity）要素信息抽取

实体要素信息包括：实体词（entity），实体类型（entityType），以及权重 w。抽取方法为：寻找分词结果中对应词性标注的词作为实体要素信息，并计算其 EventTF 值（权重值）。此外由于本体实体具有三个子类：物体、人物、机构，因此二级词性也要分别与这三个子类一一对应。具体的对应关系如表 4.3 所示：

表 4.3 实体要素信息对应词性

ICTCLAS二级词性	词性解释	实体子类
n	名词	物体
nr	人名	人物
nr2	汉语名字	人物
nrj	日语人名	人物
nrf	音译人名	人物
nt	机构团体名	机构
nz	其他专名	物体
vn	动名词	物体

4.2.7 描述（Description）要素信息抽取

描述要素信息包括描述词（description），倾向类型（tendencyType），倾向程度（tendencyLevel），以及权重 w，其中倾向类型和倾向程度是指描述词带有的主观情感语义。默认的倾向类型为空，倾向程度为 0，即为中性词。倾向类型和倾向程度的确定，可能方法是人工处理或参考已有的情感信息数据包。在本文中暂不考虑，所以将使用默认值。

描述要素信息的抽取方法为：寻找分词结果中对应词性标注的词作为描述要素信息，并计算其 EventTF 值（权重值）。具体的对应关系如表 4.4 所示：

表 4.4 描述要素信息对应词性

ICTCLAS二级词性	词性解释
a	形容词
ad	副形词
an	名形词
ag	形容词性语素
al	形容词性惯用语

4.2.8 谓语（Predicate）要素信息抽取

谓语要素信息包括谓语词（Predicate）和权重 w，抽取方法为：寻找分词结果中对应词性标注的词作为谓语要素信息，并计算其 EventTF 值（权重值）。具体的对应关系如表 4.5 所示：

表 4.5 谓语要素信息对应词性

ICTCLAS二级词性	词性解释
v	动词
vi	不及物动词
vd	副动词
vn	名动词
vl	动词性惯用语

4.2.9 扩展（Extra）要素信息抽取

扩展要素信息的格式：{扩展词}。扩展要素信息包括较为模糊但语义作用不能忽略的二级词性词，如代表方位词的“/f”（“东”）代表处所的“/s”(“途中”)和代表状态词的“/z”（“正好”）。

抽取方法为：寻找分词结果中对应词性标注的词作为扩展要素信息，并计算其 EventTF 值。具体的对应关系如表 4.6 所示：

表 4.6 扩展要素信息对应词性

ICTCLAS二级词性	词性解释
t	时间词
tg	时间词性语素
f	方位词
s	处所
b	区别词
z	状态词

4.3 网络舆情话题顶层本体构建

4.3.1 网络舆情话题顶层本体的构建要求

网络舆情话题顶层本体的构建需要满足一些基本要求：

（1）完整：顶层本体的结构应足以容纳网络舆情话题跟踪中所关注的语义要素；

（2）独立于话题：顶层本体应不依赖于任何特定话题的内容，即可被用于构建任何舆情话题的话题本体；

（3）独立于算法：顶层本体应尽量不依赖于某种特定的相似度算法模型，即本体结构可支持本文涉及的相似度算法在内的多种相似度算法。

4.3.2 网络舆情话题顶层本体的结构

网络舆情话题顶层本体指满足网络舆情话题跟踪需要，可作为任意一个特定话题本体构建的基础的本体，特点是只含有类（Class）级别的概念及概念关系，不含有实例（Individuals），且顶层本体内的所有概念都不依赖于某个特定的网络舆情话题。

在第三章网络舆情话题的跟踪要素分析的基础上，本文提出一种使用 OWL 语言构建的网络舆情话题本体（the Internet Public Opinion Topic Ontology, IPOTO）的顶层本体结构。该本体的顶层结构如下：

信息实例权重（InfoWeight）、事件（Event）、时间（Time）、地点（Location）、实体（Entity）、描述（Description）、谓语（Predicate）、扩展（Extra）、具体程度（AccurateLevel）、倾向级别（TendencyLevel）、倾向类型（TendencyLevel）、年（Year）、月（Month）、日（Day）、

时（Hour）、分（Minute）、秒（Second），其中实体包含三个子类：人物（Human）、组织机构（Organization）、物体（Object）。

信息实例权重是指单个要素信息实例的统计指标值，用以衡量该实例与话题的相关性，在本文中，话题的事件要素信息实例的 InfoWeight 为 *1/本体内事件要素信息实例个数*，话题的时间、地点、实体、描述、谓语、扩展要素信息实例的 InfoWeight 即为实例的权重 *w*（本文中为 *TopicTF*）。

话题顶层本体内含有的对象属性及对应的定义域与值域如表 4.7 所示：

表 4.7 话题顶层本体的对象属性

序号	对象属性	定义域	值域
1	hasInfoWeight	Event/Time/Location/Entity/Description/Predicate/Extra	InfoWeight
2	hasTime	Event	Time
3	hasLocation	Event	Location
4	hasEntity	Event	Entity
5	hasDescription	Event	Description
6	hasPredicate	Event	Predicate
7	hasExtra	Event	Extra
8	fromPlatform	Event	Entity
9	fromPublisher	Event	Entity
10	fromTime	Event	Time
11	hasAccurateLevel	Time/Location	AccurateLevel
12	locatedIn	Location	Location
13	doByPredicate	Entity	Predicate
14	doneByPredicate	Entity	Predicate
15	hasTendencyLevel	Tendency	TendencyLevel
16	hasTendencyType	Tendency	TendencyType
17	hasYear	Time	Year
18	hasMonth	Time	Month

表 4.7(续)

序号	对象属性	定义域	值域
19	hasDay	Time	Day
20	hasHour	Time	Hour
21	hasMinute	Time	Minute
22	hasSecond	Time	Second

对于各对象属性的说明如下：

（1）对象属性 1 表示要素信息具有信息实例权重；

（2）对象属性 2 至 7 事件具有对应的各类要素信息；

（3）对象属性 8 至 10 表示事件具有发布平台、发布人、发布时间；

（4）对象属性 11 表示时间或地点要素信息具有信息具体程度；

（5）对象属性 12 表示地点间的从属关系，更小范围的地点（子地点）位于更大范围的地点（父地点）之中；

（6）对象 13 和 14 表示实体发出或接受的动态变化，而动态变化承载于谓语要素信息中。这两个属性进行配对使用，则可反映多个实体间的动态关系和关系中实体的主客位置；

（7）对象 15 和 16 表示描述要素信息具有情感倾向类型和情感倾向程度；

（8）对象 17 至 22 表示时间要素信息具有年、月、日、时、分、秒；

（9）本文中所有对象属性的定义域和值域的集合大小约束（即可包含多少个取值）均使用 OWL 的 some 关键字，表示从 0 到 n 均可。

顶层本体自身不具有类的实例，但话题本体的类来自于顶层本体的类且具有实例，因此在这里设定各个类的实例格式及举例如表 4.8 所示：

表 4.8 话题本体类的实例格式

序号	类	实例格式	举例
1	InfoWeight	0 到 1 的浮点数字	0.55
2	Event	事件名称	(某条新闻的标题)
3	Time	多组数字+“时间量词”相联结的字符串	14 年 03 月 08 日

表 4.8(续)

序号	类	实例格式	举例
4	Location	字符串	北京
5	Entity	字符串	飞机
6	Description	字符串	愤怒
7	AccurateLevel	正整形数字	5
8	TendencyType	字符串	喜悦/悲伤
9	TendencyLevel	0 到 1 的浮点数字	0.8
10	Predicate	字符串	发布
11	Extra	字符串	上空
12	Year	2 位或 4 位数字	12/2012
13	Month	1 位或 2 位数字	3/03
14	Day	1 位或 2 位数字	8/08
15	Hour	1 位或 2 位数字	0/00
16	Minute	1 位或 2 位数字	6/06
17	Second	1 位或 2 位数字	1/01

4.3.3 网络舆情话题顶层本体的 OWL 表达

使用 Protégé对话题顶层本体结构进行 OWL 表达，图 4.2 是使用 Protégé内置的可视化工具 OntoGraph 制作的话题顶层本体结构图，其中块状与圆圈表示的是本体的类，实线箭头表示类的层级关系（箭头从父类指向子类），虚线箭头表示类间的对象属性关系（箭头从定义域指向值域）：

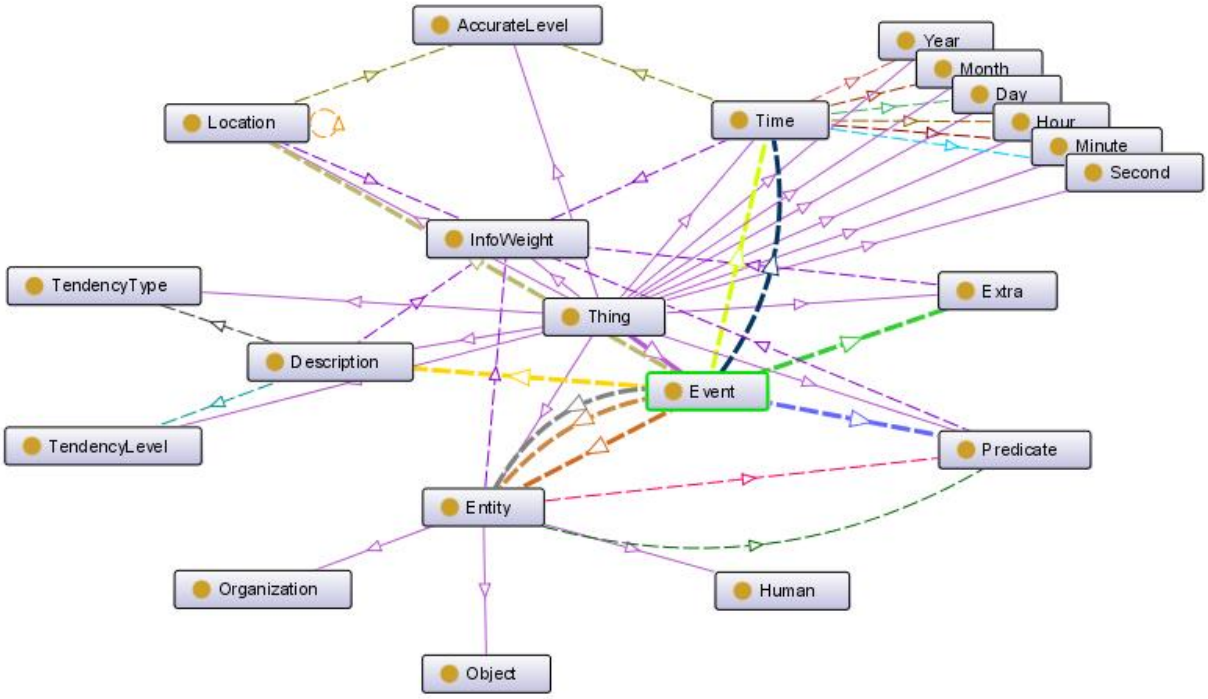


图 4.2 话题顶层本体在 Protégé 中的可视化

4.4 网络舆情话题初始本体半自动构建

4.4.1 总体流程

网络舆情话题初始本体构建是指寻找某一特定话题的相关语料事件，使用 §4.2 中所述的事件文本信息抽取方案，配合人工的调整和补充，得到该话题的本体要素信息，并将其导入话题顶层本体，从而得到同时具有话题顶层本体框架和对应该话题的本体实例的话题初始本体。本质上话题初始本体的构建也是一种话题顶层本体的进化。

每次构建流程只处理一个语料事件文本，当所有用以构建话题初始本体的语料事件文本处理完毕，得到的本体即为话题初始本体。总体流程如图 4.2 所示：

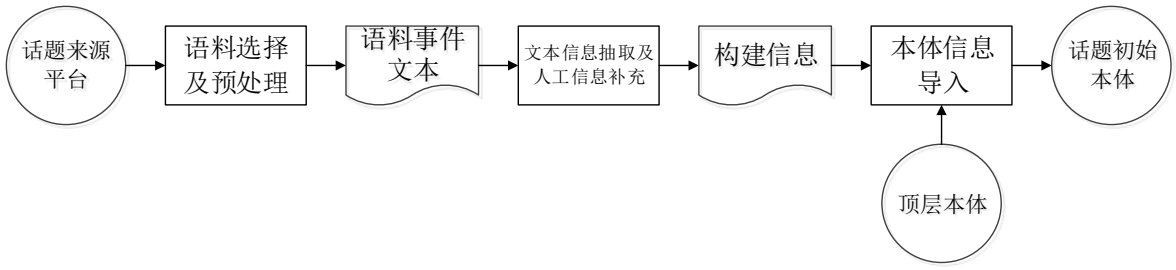


图 4.2 话题初始本体构建流程

4.4.2 语料选择及预处理

在本文中，话题初始本体构建的前提是已识别和确定要进行跟踪的具体话题，即已经成为或较可能成为热点话题的网络舆情话题。用以构建的语料需为较规范、可信的语料事件（比如来源为各大新闻门户网站，或是社交网络中官方认证账号的信息），语料事件主要由人工进行选择和采集，数量应在保证语料规范、可信的前提下尽可能多。语料事件的原始文本经过预处理切除干扰文本并统一编码为 UTF-8 后，由人工进行事件要素信息的识别与添加，得到语料事件文本。

4.4.3 构建方法

对一个语料事件文本实行 §4.2 中所述的文本信息抽取方案，得到的要素信息予以人工筛选，只留下与话题相关的要素信息实例；此外由于所选择的语料事件大多来源于规范、可信的新闻报道，语料文本中可供抽取的主观性、情感性的描述要素信息偏少（如 §4.2.4 所引的新闻文本中完全不含描述要素信息），因此还需要人为在社交网络平台、博客上获取少量的与话题相关的带有倾向性的描述要素信息。

以上过程得到的构建信息，剔除其中的权重 *EventTF*，计算实例的 *TopicTF* 为 InfoWeight 属性的值，得到要导入的话题本体类的实例及对应的对象属性值，将这些信息导入话题顶层本体的话题信息模型中，得到此次构建完成后的本体的话题信息。每次构建流程得到的话题信息均导入前一次构建时得到的话题信息之中，当所有构建流程结束，将最终得到的话题信息内的要素信息实例转换为话题本体类的实例格式，使用 Protégé 生成 OWL 表达的话题初始本体。

一个对应于某特定本体类的要素信息实例 *i* 的 *TopicTF* 计算方法如下：

$$TopicTF_i = \frac{Count}{TotalCount}$$

其中 *TotalCount* 是该类的实例的在该话题本体中被匹配的总频次，*Count* 为实例 *i* 在该话题本体中被匹配的频次。这里的匹配频次对于本次构建流程中的事件语料文本内出现的多个相同要素信息实例不进行累计。比如“飞机”一词在文本中被匹配共 3 次，其对应的 *Count* 值为 1 而不是 3。

由于本文提出的话题模型和文本信息抽取方案不支持部分本体对象属性的信息提取（如反映实体间动态关系的 doByPredicate 和 doneByPredicate、地点的层级关系 locatedIn

等），所以在必要时可由人工直接对 OWL 表达的本体进行补充。

4.5 基于本体半自动学习的网络舆情话题本体进化

4.5.1 总体流程

基于本体半自动学习的网络舆情话题本体进化是指采用半自动的本体学习技术，从网络舆情事件来源平台（如腾讯新闻、新浪微博）上获取与话题相关的事件文本，识别其中的相关话题信息，并将其导入话题本体，从而使话题本体能够将舆情话题表达得更为准确和完善。

其总体流程是：从来源平台获取可能相关的语料事件并处理得到语料事件文本，对语料事件文本进行文本信息抽取，得到的事件模型与（未进化过的话题初始本体或已进化过的）话题本体所映射的话题模型进行话题相似度计算，若超过预设的话题相似度进化阈值（阈值 1），则认为该事件属于话题相关事件，并将对应的事件要素信息中的事件、时间、地点、描述、扩展信息导入本体的话题模型。由于实体信息和谓语信息是舆情话题本体中最为关键的两类要素信息，因此对其进化进行更严格的控制，对二者分别设定一个话题相关度进化阈值（阈值 2 和阈值 3），逐个计算其中每个信息词的话题相关度，若超过对应进化阈值，则认为该信息词与话题相关并将其导入本体的话题模型。最终将新的本体话题模型转化成进化后的话题本体，该次进化结束。

每次进化流程只处理一个语料事件文本。总体流程如图 4.3 所示：

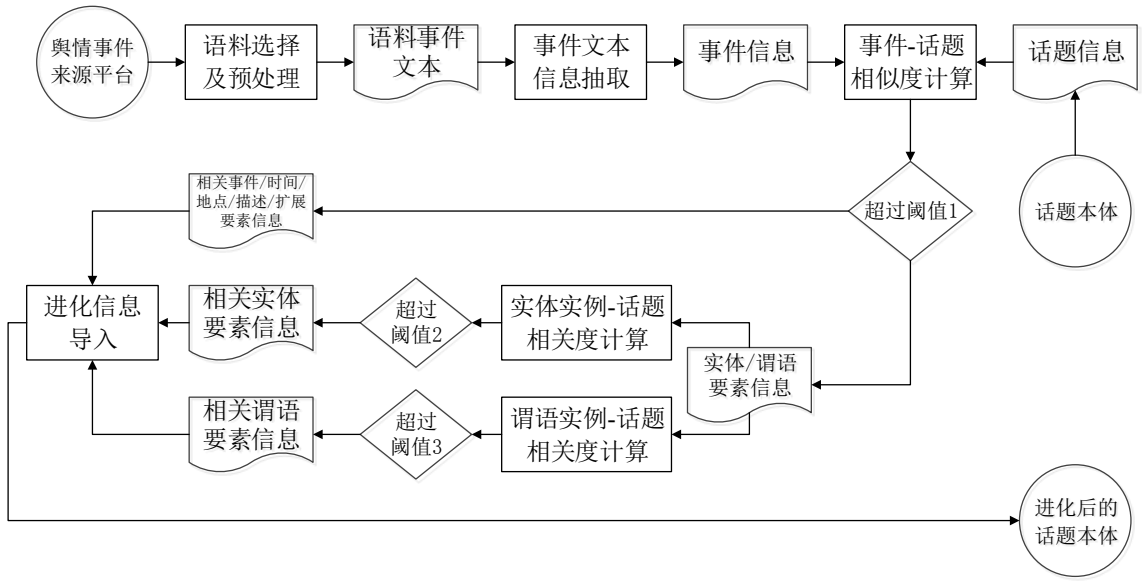


图 4.3 话题本体进化流程

4.5.2 语料选择及预处理

话题本体进化的语料规范性要求相对较低，新闻门户网站、社交网络平台、博客等都可作为语料事件的来源。针对新浪微博等社交网络消息文本的短文本特点，本文认为消息的发布时间在某个长度区间内的一定数量的消息可视为同一个语料事件的原始文本并进行合并，比如发布时间在 2014 年 4 月 6 日 16 时内的微博，每 100 条合并为一个语料事件的原始文本。

语料事件的原始文本经过预处理，切除干扰文本、统一编码为 UTF-8。经由人工进行事件要素信息的识别与添加，最后得到语料事件文本。

4.5.3 事件/话题信息生成

语料事件文本经过事件文本信息抽取生成事件信息，话题信息则通过将话题本体内的类实例及对象属性转换为话题模型内的要素信息格式得到。

4.5.4 事件-话题相似度算法

事件-话题相似度算法用于计算一个特定事件语料文本所含事件信息与话题本体所含话题信息的相似度并判断是否超过预设的事件-话题相似度阈值（*threshold_TopicSim*，阈值 1），从而判断该事件是否为话题相关事件。

对于一个特定的语料事件 *e*，具体的计算过程如下：

输入：语料事件 *e* 的事件信息，待进化话题本体 *t* 的话题信息，*threshold_TopicSim*

输出：*e* 是否为 *t* 的相关事件

算法：

（1）计算事件 *e* 的时间要素信息相似度：

$$Sim(Time)_e = \sum_{i}^{i} EventTF_i * w(TimeAccurateLevel)_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的时间要素信息实例，

$$w\left(TimeAccurateLevel\right)_i = AccurateLevel_i / 6$$

（6 为时间要素信息的具体程度 *AccurateLevel* 的最大值）

（2）计算事件 *e* 的地点要素信息相似度：

$$Sim(Location)_e = \sum_i^i EventTF_i * w(LocationAccurateLevel)_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的地点要素信息实例，*w(LocationAccurateLevel)* 与地点要素信息的具体程度 AccurateLevel 有关，在本文中暂设为 *1*。

（3）计算事件 *e* 的实体要素相似度：

$$Sim(Entity)_e = \sum_i^i EventTF_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的实体要素信息实例。

（4）计算事件 *e* 的描述要素相似度：

$$Sim(Description)_e = \sum_i^i EventTF_i * TendencyLevel_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的实体要素信息实例，*TendencyLevel* 暂设为 *1*。

（5）计算事件 *e* 的谓语要素相似度：

$$Sim(Predicate)_e = \sum_i^i EventTF_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的谓语要素信息实例。

（6）计算事件 *e* 的扩展要素相似度：

$$Sim(Extra)_e = \sum_i^i EventTF_i$$

其中，*i* 为事件 *e* 文本中一个被匹配到的扩展要素信息实例。

（7）计算事件 *e* 的可信度：

$$Reliability\left(Event\right)_e = w(FromPlatform) * Reliability\left(FromPlatform\right)_e$$

$$+w(FromTime) * Reliability\left(FromTime\right)_e$$

$$+w(FromPublisher) * Reliability\left(FromPublisher\right)_e$$

其中，

$$Reliability\left(FromPlatform\right)_e = \left\{1/3,\ 2/3\right\}$$

（新浪微博：1/3 ，其他平台：2/3）

$$Reliability\left(FromTime\right)_e = 1 - \frac{NowTime - FromTime}{NowTime - StartTime}$$

(*NowTime* 为当前计算进行时的时间，*StartTime* 为所有事件中最早的 *fromTime*)

$$Reliability\left(FromPublisher\right)_e = \left\{0,\ 1\right\}$$

（有发布人：*1*，无发布人：*0*）

在本文中 *w(FromPlatform)*、*w(FromTime)*和 *w(FromPublisher)*暂时分别设为 *1/3*。

（8）计算事件 *e* 的事件-话题相似度：

$$Sim\left(Event,\ Topic\right)_e = \begin{pmatrix} w(Time)*Sim\left(Time\right)_e + w(Location)*Sim\left(Location\right)_e \\ +w(Entity)*Sim\left(Entity\right)_e + w(Description)*Sim\left(Description\right)_e \\ +w(Predicate)*Sim\left(Predicate\right)_e + w(Extra)*Sim\left(Extra\right)_e \end{pmatrix} * Reliability\left(Event\right)_e$$

其中，*w(Time)*、*w(Location)*、*w(Entity)*、*w(Description)*、*w(Predicate)*、*w(Extra)*在本文中暂时分别设为 *2/10, 1/10, 3/10, 1/10, 2/10, 1/10*。

（9）判断 *Sim(Event,Topic)*是否超过（大于等于）*threshold_TopicSim*。

从算法公式可知，最终算得的相似度 *Sim(Event, Topic)*值域为*[0, 1]*，因此进化阈值 *threshold_TopicSim* 的取值也应在此范围内，具体取值可根据之前本体进化的结果每次进行灵活调整。

（10）若 *Sim(Event,Topic)*小于 *threshold_TopicSim*，事件 *e* 被判为不相关事件，本体进化失败，本次进化终止；若大于等于进化阈值，事件 *e* 被判为相关事件，事件-话题相似度算法结束，开始对事件 *e* 进行实例-话题相关度算法。

4.5.5 实例-话题相关度算法

实例-话题相关度算法是在本次进化涉及的语料事件已被判定为话题相关事件的基础上，对该事件所含有的某一类或某几类要素信息（不包括事件要素信息）的各个实例进行进一步的逐个筛选，从而得到可导入话题本体的与话题相关的要素信息个体的集合。在事件-话题相似度算法中并未针对每一个要素信息实例进行筛选，粒度上只考虑事件整体的话题相关度，因此无法过滤掉对应于某一类要素信息的与话题不相关的单个实例。

本文中提出的实例-话题相关度算法参考关键词的共现分析方法，将存在于话题本体内的要素信息实例全部视为话题关键词。单个要素信息实例与话题的相关度，可用该要素信

息实例与相同要素类型的各话题关键词在各相关事件文本中的总共现频率来进行衡量。

此外本文认为，对要素信息实例的筛选重点类型是实体要素信息和谓语要素信息两类，因为这两类信息是对一个事件最基本的描述，因此算法主要考虑这两种要素信息类型。对于不同类型的要素信息，实例-话题相关度计算没有区别，都是对要素信息实例进行共现分析，区别主要在于设定了不同的相关度阈值，以对应不同要素信息类型的实例间可能表现出的不同共现特点。

对于一个特定的话题相关事件的实体/谓语要素信息 *i*，具体的计算过程如下：

输入：实体要素信息 *i*，待进化话题本体 *t* 所有相关事件的事件信息，待进化话题本体 *t* 的话题信息中的所有实体关键词，*threshold_TopicRelevance(Entity)*/*threshold_TopicRelevance(Predicate)*（阈值 2/阈值 3）

输出：*i* 是否为 *t* 的相关实体/谓语要素信息实例

算法：

此次本体进化涉及的话题相关事件中的一个特定实体/谓语要素信息 *i*，具体算法如下：

（1）计算实例-话题相关度：

$$Relevance = \sum^{e} \frac{CoOccurFrequency_e}{EventCount}$$

其中，

e 指话题相关事件；

EventCount 指话题相关事件的总数；

$CoOccurFrequency_e$ 指在一个话题相关事件的语料文本中 *i* 的关键词共现频率，对于特定的 *i* 和相关事件 *e*，其计算公式如下：

$$CoOccurFrequency = \sum^{k} \frac{CoOccurCount_k}{KeywordCount}$$

其中，

k 为 *e* 中出现的话题的实体/谓语关键词；

KeywordCount 指 *e* 中出现的话题关键词总数（重复出现不计）；

$CoOccurCount_k$ 指 *e* 中 *i* 与 *k* 的共现次数，在本文中若 *i* 与 *k* 在 *e* 中同时出现，则该值

等于 *1*，否则等于 *0*；

KeywordCount 若为 *0*，则 *CoOccurFrequency* 等于 *0*。

（2）判断 *Relevance* 是否超过（大于等于）*threshold_TopicRelevance(Entity)/ threshold_TopicRelevance(Predicate)*。

从算法公式可知，最终算得的实体要素信息实例或谓语要素信息实例的 *Relevance* 值域均为*[0, 1]*，因此两个阈值的取值也应都在此范围内，具体取值可根据之前本体进化的结果每次进行灵活调整。

（3）若实体要素信息实例 *i* 的 *Relevance* 小于 *threshold_TopicRelevance(Entity)*，*i* 被判为话题相关的实体要素信息实例；同理，若谓语要素信息实例 *i* 的 *Relevance* 小于 *threshold_TopicRelevance(Predicate)*，*i* 被判为话题相关的谓语要素信息实例。

该算法每次处理一个实体/谓语要素信息实例 *i*，对此次进化涉及的相关事件 *e* 内所有 *i* 逐个计算完成，即可得到用于本体进化的话题相关实体/谓语要素信息。

4.5.6 进化信息导入

经过事件-话题相似度算法和实例-话题相关度算法计算，即可得到此次本体进化要导入的进化信息。在本文算法中进化信息包括：超过阈值 1 的事件内的除实体和谓语要素信息以外的要素信息，超过阈值 2 的该事件实体要素信息以及超过阈值 3 的该事件谓语要素信息。

剔除进化信息中包含的权重 *EventTF*，计算实例的 *TopicTF* 为 InfoWeight 属性的值，得到要导入的话题本体类的实例及对应的对象属性值。将这些信息导入待进化话题本体的话题模型中，得到此次进化后话题本体的话题信息。最后将话题信息内的要素信息实例转换为话题本体类的实例格式，使用 Protégé 生成 OWL 表达的进化后话题本体。*TopicTF* 的计算规则与§4.4.3 话题初始本体构建流程中的 *TopicTF* 的计算方法相同。

4.5.7 人工修正和补充

对最终得到的 OWL 话题本体进行人工修正和补充，具体内容有：

（1）每个类型的要素信息均保留 *TopicTF* 值排序靠前的一定数量的实例；

（2）删除自动算法导入的无关要素信息实例；

（3）补充 doByPredicate、doneByPredicate、locatedIn 等对象属性信息。

第五章　网络舆情话题本体构建及进化实验

5.1　实验系统设计

5.1.1 系统模块设计

实验系统主要包括三大模块：事件文本信息抽取模块、本体进化模块、信息查询模块。

5.1.2 数据库设计

下面是数据库设计过程中使用到的关系模式：

1. 本体（本体 id，本体话题名称，进化次数，创建时间）

2. 事件（事件 id，发布日期，发布 URL（微博为空），发布平台，发布人，素材文本，创建时间）

3. 本体-事件（本体 id，事件 id，是否相关，事件相似度，时间相似度，地点相似度，实体相似度，描述相似度，谓语相似度，扩展相似度，话题相似度）

4. 事件-事件时间（事件 id，时间 id，事件中时间累计频数，占事件时间总频数比）

5. 事件-事件地点（事件 id，地点 id，事件中地点累计频数，占事件时间总频数比）

6. 事件-事件实体（事件 id，实体 id，事件中实体累计频数，占事件时间总频数比）

7. 事件-事件描述（事件 id，描述 id，事件中描述累计频数，占事件时间总频数比）

8. 事件-事件谓语（事件 id，谓语 id，事件中谓词累计频数，占事件谓语总频数比）

9. 事件-事件扩展（事件 id，扩展 id，事件中扩展累计频数，占事件时间总频数比）

10. 实体-本体相关度（实体 id，本体 id，实体与本体的相关度）

11. 谓语-本体相关度（谓语 id，本体 id，谓语与本体的相关度）

12. 本体-本体时间（本体 id，时间 id，本体中时间累计频数，占本体时间总频数比）

13. 本体-本体地点（本体 id，地点 id，本体中地点累计频数，占本体地点总频数比）

14. 本体-本体实体（本体 id，实体 id，本体中实体累计频数，占本体实体总频数比）

15. 本体-本体描述（本体 id，描述 id，本体中描述累计频数，占本体描述总频数比）

16. 本体-本体谓语（本体 id，谓语 id，本体中谓语累计频数，占本体谓语总频数比）
17. 本体-本体扩展（本体 id，扩展 id，本体中扩展累计频数，占本体扩展总频数比）
18. 时间（时间 id，年，月，日，时，分，秒，时间具体度）
19. 地点（地点 id，父地点 id，地点名，地点具体程度）
20. 实体（实体 id，实体名，实体类型）
21. 描述（描述 id，描述，情感类别，倾向值）
22. 谓语（谓语 id，谓语）
23. 扩展（扩展 id，扩展）
26. 谓语关系（主体实体 id，客体实体 id，谓词 id）

5.2 系统实现

5.2.1 平台环境

本文搭建的实验系统的平台环境为 Windows 8、C#.NET Framework 4.5、关系数据库 MySQL。此外使用了用以进行分词和词性标注的 ICTCLAS50 中科天玑组件演示程序，以及用于构建和可视化本体 OWL 表达的 Protégé。

5.2.2 数据库实现

数据库的数据表及存储过程实现如图 5.1 所示：

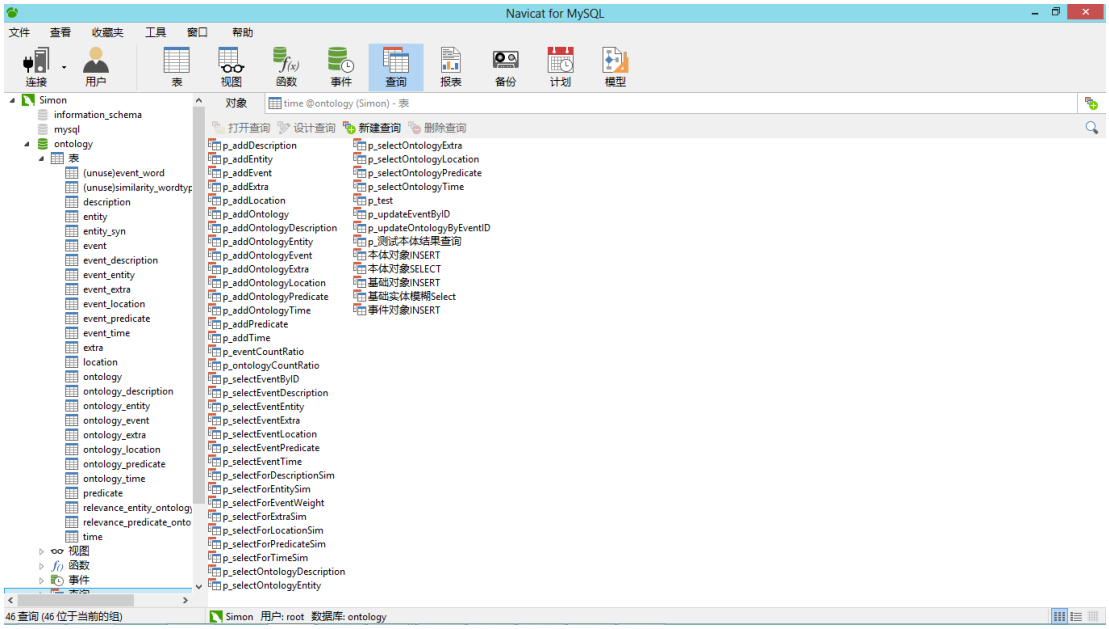


图 5.1 数据库实现

5.2.3 界面实现

实验系统的部分界面实现如图 5.2 至 5.5 所示：

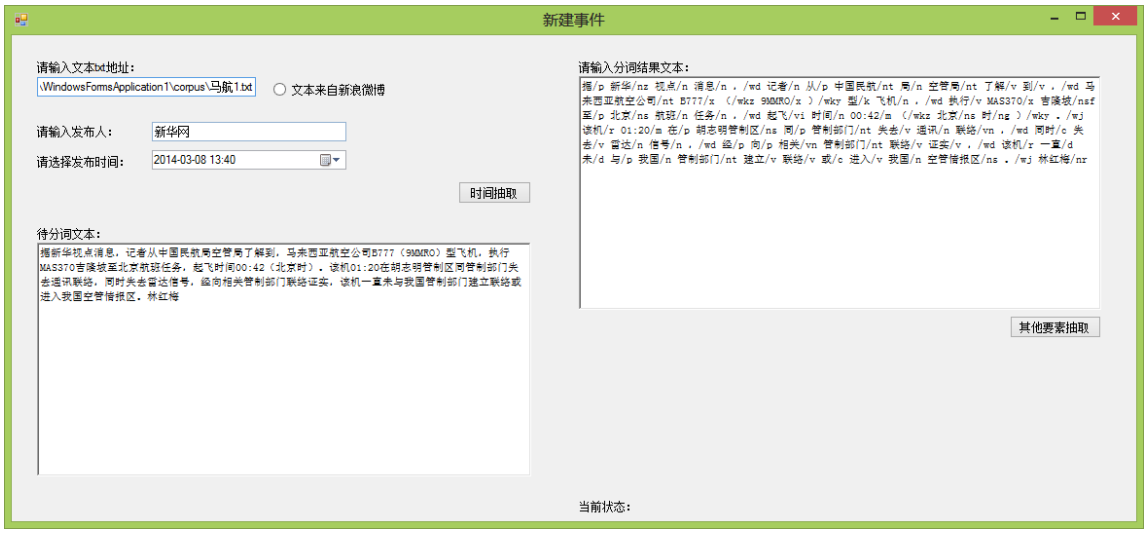


图 5.2 事件文本信息抽取界面

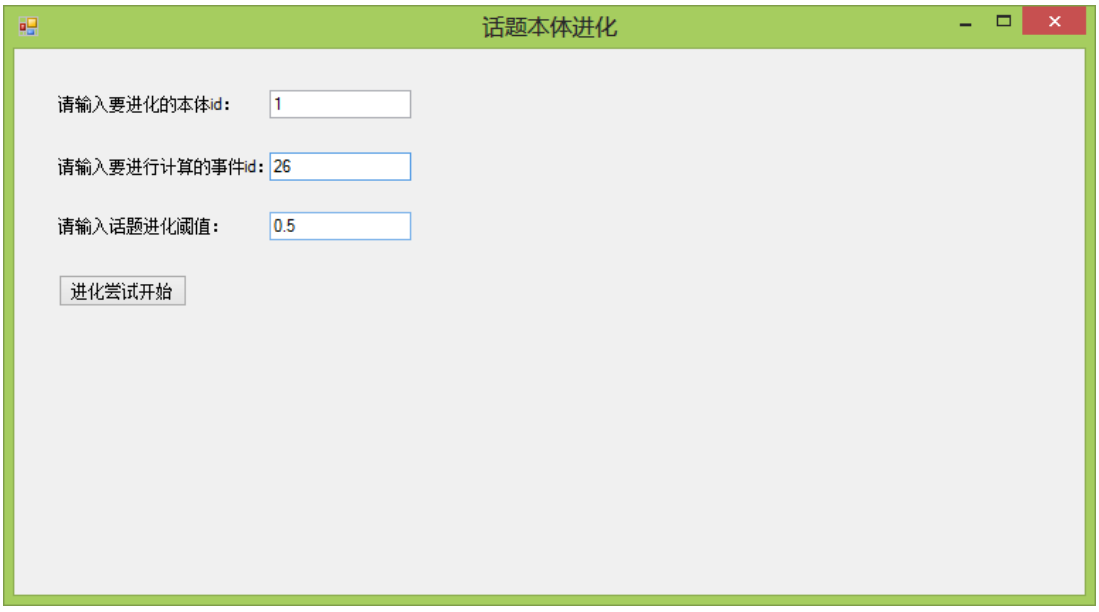


图 5.3 话题本体进化界面

图 5.4 事件信息查询界面

图 5.5 本体信息查询界面

5.2.4 方案简化

实验系统使用的话题本体进化方案在前文提出的方案上进行了一定的简化：

（1）对于事件-话题相似度算法，§4.5.4（7）中的 *Reliability(fromTime)*计算方法简化为：

$$Reliability\left(fromTime\right)=\{0,1\}$$

（若存在发布时间则值为 1，否则为 0）

（2）未使用实例-话题相关度算法，而将事件-话题相似度算法得到的话题相关事件的所有要素信息实例都作为话题相关实例进行进化；

（3）进化后得到的本体未进行人工修正和补充，即整个进化过程除进化阈值设定外全部自动进行。

5.3 实验语料选择

文本进行实验所使用的网络舆情话题为：马航 MH370 事件，实验语料全部是与该话题相关的真实语料文本，来源是腾讯新闻（新闻门户网站）和新浪微博（社交网络平台）。

来自腾讯新闻的一个话题语料事件的原始页面如图 5.6 所示：

吉隆坡飞往北京客机在胡志明管制区失去联系

马来西亚航班失联 | 新华网 2014-03-08 09:09 | 我要分享 | 707975

据新华视点消息，记者从中国民航局空管局了解到，马来西亚航空公司B777（9MMRO）型飞机，执行MAS370吉隆坡至北京航班任务，起飞时间8日00:42（北京时）。该机01:20在胡志明管制区同管制部门失去通讯联络，同时失去雷达信号，经向相关管制部门联络证实，该机一直未与我国管制部门建立联络或进入我国空管情报区。林红梅 返回腾讯网首页>>

图 5.6 腾讯新闻“马航 MH370”语料事件

来自新浪微博的一些语料原始文本如图 5.7 所示：

1 你家只有展昭昭**03月09日 23:59?来自微博 weibo.com**:#310李宇春生日快乐#我最亲爱的亲爱的姑娘 生日快乐 ，我可是看着你长大的[偷乐]，所以在我心中，你永远san岁 同时#祈福马航#只盼平安

2 -毛开心**03月09日 23:59?来自三星GALAXY S4**:夜晚到了,总带有一些凄凉和悲伤。我们都要坚强。为国人祈祷,为马航祈祷,为社会祈祷,为世界祈祷。我们都可以很好。* [心] 我在: http://t.cn/8FDjwCW

3 王振CUPL**03月09日 23:59?来自微博 weibo. com**:MH370一个又一个的消息出来，而马航却一个又一个的否认，徘徊于肯定与否定中，何时能休？相对于失踪，宁愿更希望他们穿越了时空。

4 深圳晚报**03月09日 23:59?来自微博 weibo.com**:#祈福马航mh370 #失联客机mh370同样让马来西亚人民揪心；民众自发聚集跪在街边祈祷平安；机长的女儿写出让人流泪的留言：爸爸，你出现在所有新闻和报纸上。快回家吧来亲自看看吧！您不激动么？[泪] 但愿明天起来，一切是虚惊一场。370快回家吧！（图：@ThatsexactlywhatIneed @王左中右）

5 保钓梁智**03月09日 23:59?来自优酷网连接分享**:河南.长葛 为马航 MH370失联客机祈福 （分享自 @优酷） http://t.cn/8sP2cOJ

6 沐昔Joice**03月09日 23:59?来自Weico. iPhone**:睡前最后一件事和醒来第一件事就是关注一下马航失联飞机最新进展，真的好希望明天一早醒来他们都回来了。真的不要再发生这种让人恐慌的事了，愿世界和平！

7 木墩_**03月09日 23:59?来自Android客户端**:其实我还是希望马航失联飞机只是暂时失去联系，会突然出现一个新闻告诉我们，他们只是停靠在一个神奇的地方，失去通信，随后所有乘客都平安返航。[蜡烛][蜡烛][蜡烛][蜡烛][蜡烛] 我在:http://t.cn/8sP2VMm

8 当香烟爱上火柴**03月09日 23:59?来自iPhone客户端**:每个人的生命里，当有这么一段岁月，像这肆意开放的木棉花，奔放、开朗、灿烂，热情洋溢，鲜红的花火犹如燃烧的火焰，在这拥堵的世界划下属于自己的哪怕刹那光辉----有感昆明恐怖事件到马航失踪事件，生命的脆弱，世事的无常， 我在:http://t.cn/8F5php9

9 幽蓝鬣语**03月09日 23:59?来自微博 weibo.com**:珍惜自己现有的生活，认真过好每一天，310不止有生日祝福，还有对马航事件的真挚祈愿！@刘诗诗 生日快乐！我们共同等待同胞回家！

10 宝宝龙可**03月09日 23:59?来自微博 weibo.com**:#马航航班失踪#【飞机上的那些鲜活面孔】目前离马航MH370航班失联已过去16个小时。飞机上，有才华横溢的少数民族画家，有虔诚的佛教徒，有热爱旅行的老年驴友；有幸福一家五口；有事业刚起步的80后年轻职员，尚在襁褓之中的婴儿……他们的人生，绝不该这样画上句点。今夜，请为他们祈祷：平安，归航！

11 陳小珂Yumi**03月09日 23:59?来自Android客户端**:#310李宇春生日快乐#春春，生日快乐.我爱你[心][鲜花][爱你]同时也#祈福马航#[蜡烛][蜡烛][蜡烛]希望奇迹发生，[飞机] 有爱就有希望，有希望就不要放弃。MH370你快回来[飞机]#祈福马航#

12 花爷宇哥主唱大人**03月09日 23:59?来自Android客户端**:#祈福马航#这个三月发生了太多的意外，很多事情我们都无法阻止，我们都只是在平凡不过的人类，没有超人力没有瞬间转移我们有的只是那颗热血，不放弃就一定能看见希望，我是火星人，我相信他们都还活着[心] http://t.cn/8Fs5KTN

13 虫-爬爬**03月09日 23:59?来自三星GALAXY S4**:刚在看马航飞机上中国乘客起飞前在互联网留下的只言片语 瞬间觉得心情无比沉重 生命真的太脆弱 你永远不知道下一秒会发生什么 也许现在我们应该做的就是学会珍惜 珍惜身边的亲人和朋友 无论是相伴一生还是擦肩而过 学会珍惜命运带给我们的每个经历 从中体会世态悲喜一致3月9~

14 拒绝SAY**03月09日 23:59?来自皮皮时光机**:【简介：赶赴马航客机失事海域救援的井冈山舰】9日晨3时许从湛江出发的井冈山舰是一艘登陆舰，舰长210米，宽28米，排水量19000吨，速度22节，该舰能携带4艘大型气垫登陆艇，同时起降2架中型直升机，飞行甲板宽大，使该舰在海上用途广泛，此次搭载了30名医护人员、10名潜水员，带有救生和水下探测装备。

15 大城小爱的2012**03月09日 23:59?来自微博 weibo.com**:[蜡烛]花花偶尔也很萎的，记得以前晚上你在我屋子里趴着就喜欢在我电脑椅附近睡觉，有几次没注意你在后边挪动椅子的时候就把你给吓到了，你个笨蛋吓得就赶紧起来结果脑袋就撞到了调整座椅高度的调节杆上了，哎！看着人心疼，你这小笨蛋！667天了，希望在天堂的花花幸福，希望马航失联的航班能平安返航

16 gnidding**03月09日 23:59?来自微博 weibo.com**:得知两个同事在马航上，虽然不认识，但还是心情沉重。如果可以我愿意相信他们穿越了时空。Bless

17 Leney小雅**03月09日 23:59?来自小米手机3**:马航飞机[飞机][飞机]失误到现在40多小时了还没有找到，真心着急[泪][泪]祈祷机上的人员平安无事[祈祷][祈祷][祈祷] 早日取得联系、好让大家心安 我在:http://t.cn/8sP2V7F

18 -杨妞-**03月09日 23:59?来自iPhone 5s**:Tata的鞋子，Teenie Weenie的针织衫，就剩下食草堂的包包和S. DEER的棉布裙子了。春天来了。祈福马航，祈福素不相识的善良的你们，愿所有人安好。

19 妮妮baby716**03月09日 23:59?来自三星Galaxy NOTE III**:@央视新闻：#马航飞机失联#【走了这么久，你们在哪里？】40个小时过去了，天又要黑了。失联，还是失联，多少人的心被撕裂！一次次徒劳拨打手机，一声声焦急呼唤，“平安”变得如此温暖和令人期盼！多希望收到你们的消息，多害怕爱沉入海底。坚持，努力，说好不放弃。转发祈祷，愿愿诚产生奇迹！[心]

20 Wendy姣姣DT**03月09日 23:59?来自Android客户端**:久久睡不着。想的事情很多。看了好多图片又把我看哭了。马航现在是全国人民都在呼唤等待的名字。如果这世上没有那么多的意外该有多好。如果一切都能化险为夷多好。真的希望你们都平安。好像此刻心里能感觉到你们面临危险时的恐惧。真的希望你们可以满心欢喜的回来，开心的告诉我们你们经历的那些奇迹。

21 赵子健Students**03月09日 23:59?来自Android客户端**:#马航飞机失联# 今年流行啥不好！非得流行什么”爸爸去哪了？时间去哪了？”这下好了吧，飞机TMD去哪了！/抠鼻 http://t.cn/8Fsb0YV

图 5.7 新浪微博“马航 MH370”语料事件

5.4 实验过程

5.4.1 “马航 MH370”话题初始本体半自动构建

话题初始本体的构建选取了“马航 MH370”事件发生早期的 4 则腾讯新闻——

（1）新华网于 2014-03-08 09:09 发布的《吉隆坡飞往北京客机在胡志明管制区失去联系》

（2）新华社于 2014-03-08 09:31 发布的《国家民航总局核实失联飞机上有 158 名中国人》

（3）新华网于 2014-03-08 16:37 发布的《马来西亚联系中国等多国参与搜救失联马航客机》

（4）国际在线于 2014-03-09 08:52 发布的《越南飞机发现两片油污带 或与失联客机有关(图)》作为语料事件，预处理过程得到的事件语料文本类似以下格式：

据新华视点消息，记者从中国民航局空管局了解到，马来西亚航空公司 B777（9MMRO）型飞机，执行 MAS370 吉隆坡至北京航班任务，起飞时间 8 日 00:42（北京时）。该机 01:20 在胡志明管制区同管制部门失去通讯联络，同时失去雷达信号，经向相关管制部门联络证实，该机一直未与我国管制部门建立联络或进入我国空管情报区。林红梅

对语料事件文本进行事件信息抽取时，由人工在 ICTCLAS50 分词系统的用户词典补充的词汇如表 5.1 所示：

表 5.1 初始本体构建中对用户词典的人工补充

序号	词汇	词性
1	马航	nt
2	民航局	nt
3	中国航空	nt
4	越南航空	nt
5	空管局	nt
6	马来西亚航空公司	nt
7	胡志明管制区	ns
8	空管情报区	ns

表 5.1(续)

序号	词汇	词性
9	管制部门	nt
10	民航总局	nt
11	马来西亚政府	nt
12	吉隆坡国际机场	ns
13	新闻发布会	n
14	苏邦空中交通控制中心	nt
15	搜救飞机	n
16	油污带	n
17	失联海域	ns
18	飞行管理总公司	nt
19	救难中心	nt
20	巡逻机	n
21	MH370客机	n

得到的各语料事件的事件信息，对其进行人工筛选，剔除与话题无关的各类型要素信息实例若干，剩下的信息导入话题信息中。

在得到的话题信息中由人工补充的要素信息实例如表 5.2 所示：

表 5.2 初始本体构建中要素信息实例的人工补充

序号	要素信息实例	所属类型
1	2014	Year
2	平安	Description
3	悲伤	Description
4	坚强	Description
5	揪心	Description
6	恐慌	Description

表 5.2(续)

序号	要素信息实例	所属类型
7	恐怖	Description
8	脆弱	Description
9	无常	Description
10	幸福	Description
11	意外	Description
12	沉重	Description
13	着急	Description
14	不舒服	Description
15	难过	Description
16	焦急	Description

事件信息和话题信息在 MySQL 数据库中的储存（以时间要素信息为例）如图 5.8-5.10 所示（其中 countRatio 指的是 EventTF 和 TopicTF）：

timeID	year	month	day	hour	minute	second	accurateLevel	updateTime
124			8				1	2014-06-03 21:24:17.000000
125			8	0	41		3	2014-06-03 21:25:30.000000
126		3	8				2	2014-06-03 21:25:31.000000
127				6	30		2	2014-06-03 21:25:31.000000
128				2	40		2	2014-06-03 21:25:32.000000
129			8	0	21		3	2014-06-03 21:25:57.000000
130				9	25		2	2014-06-03 21:25:57.000000
131				10	25		2	2014-06-03 21:25:57.000000
132				1	21		2	2014-06-03 21:25:57.000000
133	2014					(Null)	1	2014-06-03 21:43:59.000000

图 5.8 时间要素信息的数据表

eventID	timeID	count	countRatio	updateTime
27	124	1	1	2014-06-03 21:59:
29	125	1	0.2	2014-06-03 22:01:
29	126	1	0.2	2014-06-03 22:01:
29	127	1	0.2	2014-06-03 22:01:
29	128	1	0.2	2014-06-03 22:01:
29	124	1	0.2	2014-06-03 22:01:
30	129	1	0.142857	2014-06-03 22:02:
30	130	1	0.142857	2014-06-03 22:02:
30	131	1	0.142857	2014-06-03 22:02:
30	132	1	0.142857	2014-06-03 22:02:
30	124	3	0.428571	2014-06-03 22:02:

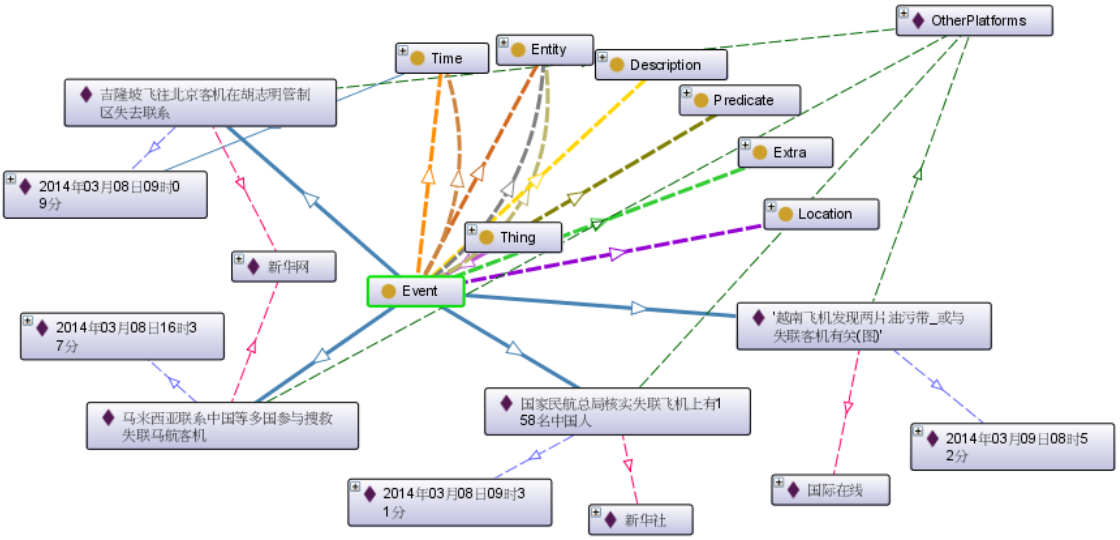
图 5.9 事件信息中的时间要素信息

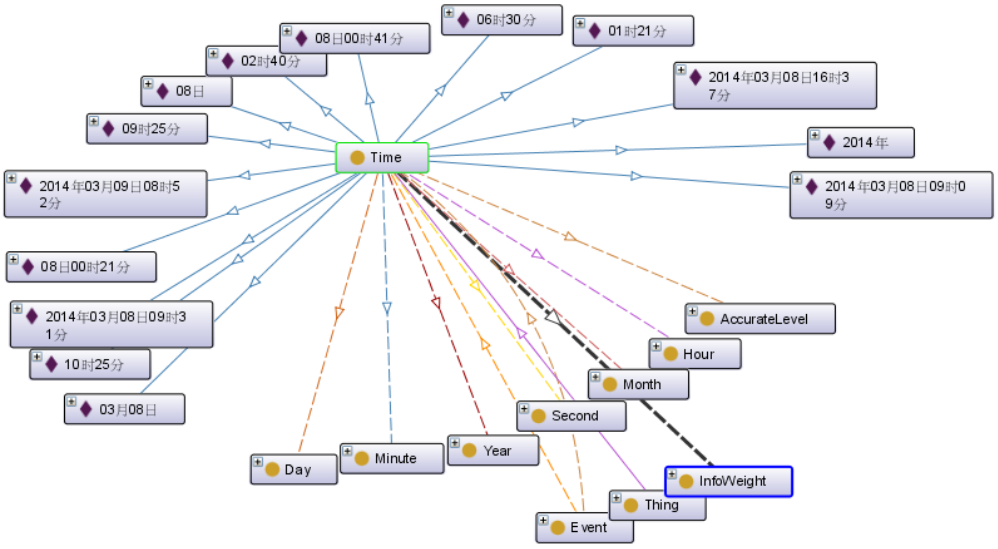
ontologyID	timeID	count	countRatio	updateTime
1	124	3	0.272727	2014-06-03 22:05:
1	125	1	0.0909091	2014-06-03 22:05:
1	126	1	0.0909091	2014-06-03 22:05:
1	127	1	0.0909091	2014-06-03 22:05:
1	128	1	0.0909091	2014-06-03 22:05:
1	129	1	0.0909091	2014-06-03 22:05:
1	130	1	0.0909091	2014-06-03 22:05:
1	131	1	0.0909091	2014-06-03 22:05:
1	132	1	0.0909091	2014-06-03 22:05:
1	133	1	-1	2014-06-03 22:20:

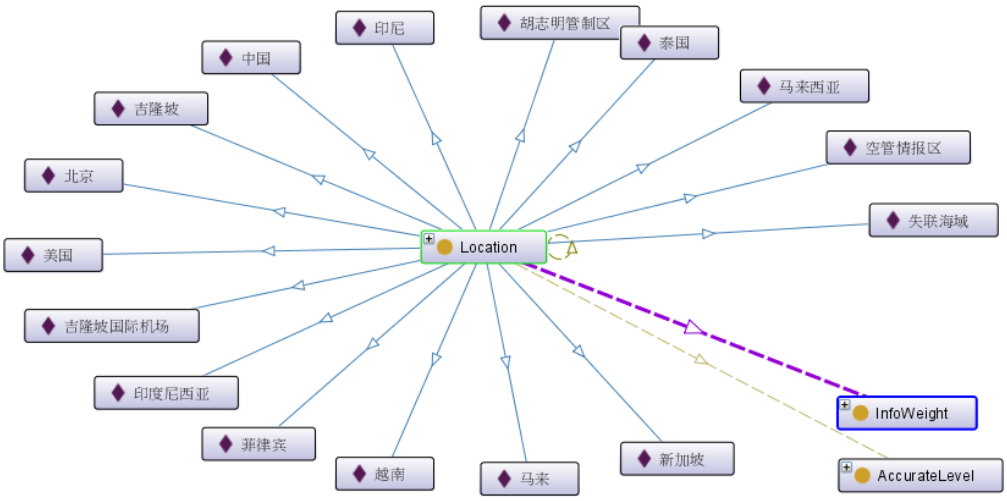
图 5.10 话题信息中的时间要素信息

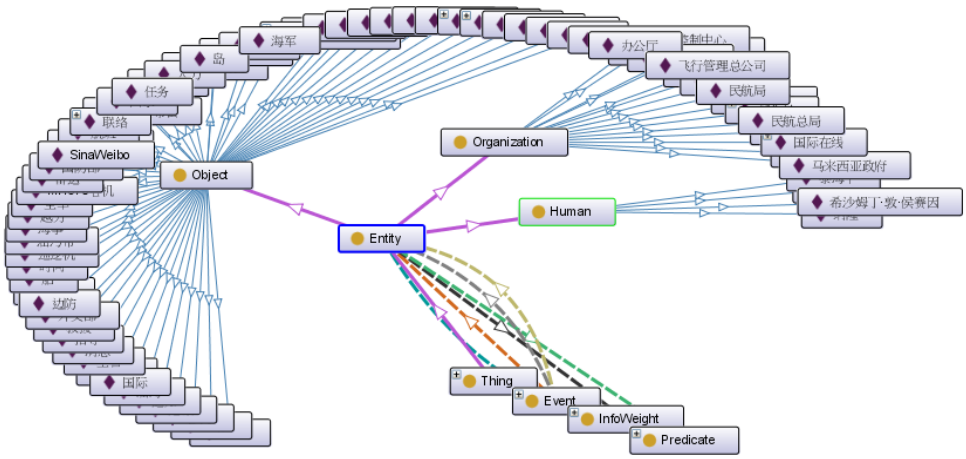
timeID 为 133 的时间要素信息即为人工补充的实例 2014 年。

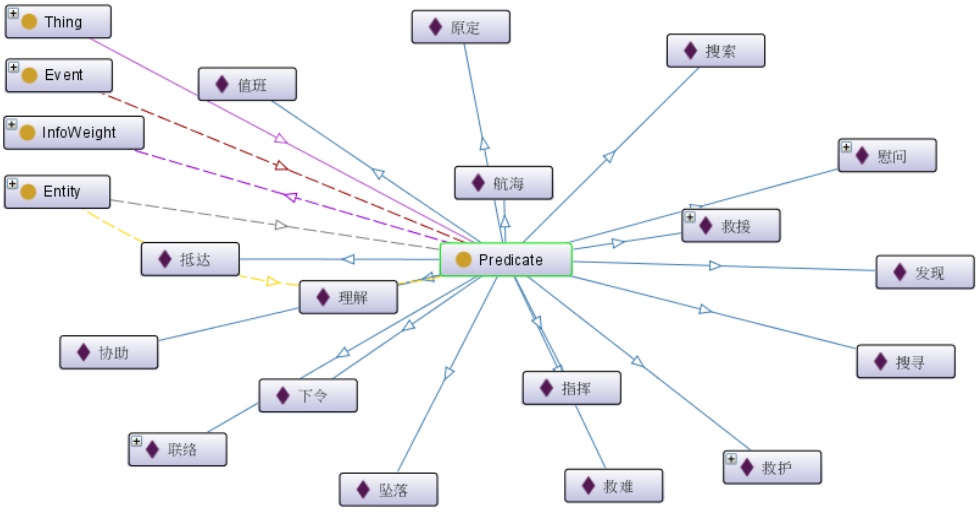
将最终得到的话题信息导入顶层本体，得到“马航 MH370”话题初始本体的 OWL 表达在 Protégé 中的可视化如图 5.11 所示：


OtherPlatforms
Time
Entity
Description
吉隆坡飞往北京客机在胡志明管制区失去联系
Predicate
Extra
2014年03月08日09时09分
Location
Thing
新华网
Event
2014年03月08日16时37分
'越南飞机发现两片油污带_或与失联客机有关(图)'
马来西亚联系中国等多国参与搜救失联马航客机
国家民航总局核实失联飞机上有158名中国人
2014年03月09日08时52分
2014年03月08日09时31分
国际在线
新华社


06时30分
08日00时41分
01时21分
02时40分
2014年03月08日16时37分
08日
09时25分
2014年
Time
2014年03月09日08时52分
2014年03月08日09时09分
08日00时21分
AccurateLevel
2014年03月08日09时31分
Hour
10时25分
Month
03月08日
Second
Year
Minute
Day
InfoWeight
Thing
Event


印尼
胡志明管制区
泰国
中国
马来西亚
吉隆坡
空管情报区
北京
失联海域
Location
美国
吉隆坡国际机场
印度尼西亚
InfoWeight
菲律宾
越南
马来
新加坡
AccurateLevel


海军
办公厅
飞行管理总公司
任务
民航局
民航总局
联络
Organization
SinaWeibo
国际在线
Object
Human
Entity
Thing
Event
InfoWeight
Predicate


Thing
原定
搜索
Event
值班
InfoWeight
慰问
航海
Entity
救援
抵达
Predicate
发现
理解
协助
搜寻
下令
指挥
联络
救护
坠落
救难

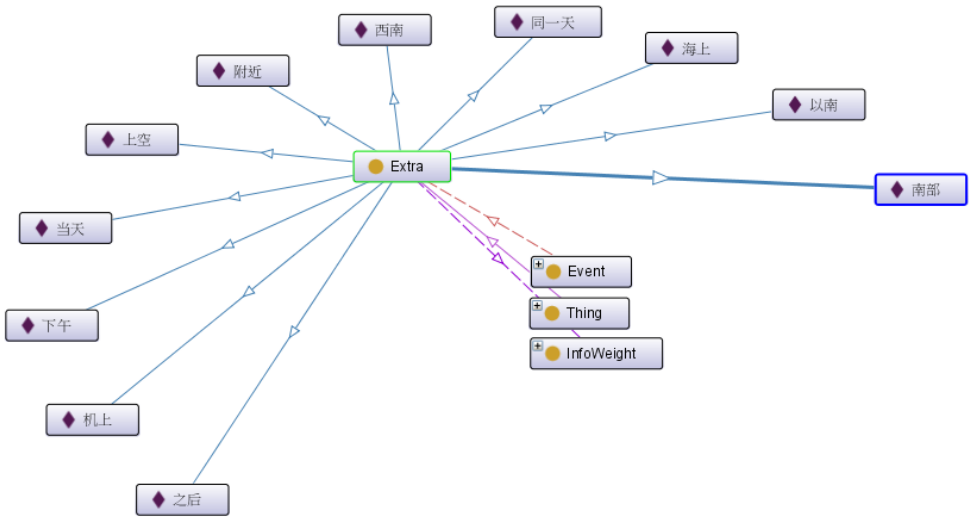


图 5.11 半自动构建的“马航 MH370 事件”话题初始本体

该话题初始本体包含的要素信息统计如表 5.3 所示：

表 5.3 “马航 MH370”话题初始本体的要素信息统计

要素信息类型	实例数量
Event	4
Time	14
Location	16
Entity	72
Description	15
Predicate	17
Extra	11

5.4.2 “马航 MH370”话题本体进化

在已产生的“马航 MH370”话题初始本体上，进行话题本体的 10 次进化。语料事件含腾讯新闻 2 个以及新浪微博 8 个（共 800 条微博）

其中，腾讯新闻分别是：中国新闻网于 2014-03-08 11:42 发布的《马航公布失联客机驾驶员信息 正机师 53 岁经验丰富》（腾讯新闻 1）以及新华网于 2014-03-09 09:23 发布的《我国海军派 2 艘舰艇参加马航失联航班救援》（腾讯新闻 2），通过预处理得到事件语料文本

共 2 则；

新浪微博的语料通过对 2014 年 03 月 09 日和 2014 年 03 月 10 日发布的含有关键字“马航”的新浪微博爬取，两天发布的微博各采用了 400 条，每 100 条合并为一个事件，并对其文本进行预处理，得到事件语料文本各 4 则（新浪微博 1 至新浪微博 4，新浪微博 5 至新浪微博 8），总共 8 则。

每次进化的进化阈值都根据之前的进化情况予以人工调整，当事件-话题相似度大于进化阈值，则认为语料事件为话题相关事件，并导入其事件信息使本体进化，否则该次本体进化失败。各次进化的具体情况如表 5.4 所示：

表 5.4 “马航 MH370”话题本体的各次进化情况

序号	事件语料文本	设定的进化阈值	事件-话题相似度	进化状态
1	腾讯新闻1	0.30	0.452	成功
2	腾讯新闻2	0.30	0.327	成功
3	新浪微博1	0.30	0.142	失败
4	新浪微博2	0.20	0.139	失败
5	新浪微博3	0.15	0.138	失败
6	新浪微博4	0.14	0.161	成功
7	新浪微博5	0.17	0.262	成功
8	新浪微博6	0.30	0.354	成功
9	新浪微博7	0.35	0.334	失败
10	新浪微博8	0.35	0.356	成功

10 次进化过程后（成功 6 次）得到的“马航 MH370”话题信息内包含的事件要素相关信息（按事件-话题相似度降序排列）如图 5.12 所示：

topic	title	topicSim	fromPlatform	fromPublisher	fromTime	fromURL	originalText
马航MH370事件	国家民航总局核实失联飞机	0.622222	OtherPlatforms	新华社	2014-03-08 09:31:00	http://news.qq.com/a/2014(	新华社e哥有话说
马航MH370事件	吉隆坡飞往北京客机在胡志	0.533334	OtherPlatforms	新华网	2014-03-08 09:09:00	http://news.qq.com/a/2014(	据新华视点消息
马航MH370事件	马航公布失联客机驾驶员信	0.451665	OtherPlatforms	中国新闻网	2014-03-08 11:42:42	http://news.qq.com/a/2014(	当地时间3月8日
马航MH370事件	(Null)	0.355494	SinaWeibo		2014-03-10 23:50:06		:马航飞机还没找
马航MH370事件	(Null)	0.353913	SinaWeibo		2014-03-10 23:50:17		:【社交媒体在马
马航MH370事件	我国海军派2艘舰艇参加马	0.32698	OtherPlatforms	新华网	2014-03-09 09:23:31	http://news.qq.com/a/2014(	新华网北京3月9
马航MH370事件	越南飞机发现两片油污带	0.312615	OtherPlatforms	国际在线	2014-03-09 08:52:00	http://news.qq.com/a/2014(	一架越南搜救飞
马航MH370事件	(Null)	0.262775	SinaWeibo		2014-03-10 23:50:17		:也许失联马航说
马航MH370事件	马来西亚联系中国等多国参	0.204959	OtherPlatforms	新华网	2014-03-08 16:37:00	http://news.qq.com/a/2014(	新华社快讯：民
马航MH370事件	(Null)	0.161213	SinaWeibo		2014-03-09 23:50:37		:失踪马航飞机的

图 5.12 “马航 MH370”进化后话题信息的事件要素相关信息

10 次进化过程后（成功 6 次）得到的“马航 MH370”话题信息内包含的其他类型的要素信息（以时间和实体为例，按 topicTF 降序排列）如图 5.13 所示：

topic	year	month	day	minute	second	timeAccurateLevel	countRatio
马航MH370事件			8			1	0.125
马航MH370事件		3	8			2	0.0714286
马航MH370事件		3	9			2	0.0535714
马航MH370事件		3	10			2	0.0535714
马航MH370事件			10			1	0.0357143
马航MH370事件			47			1	0.0357143
马航MH370事件				40		2	0.0357143
马航MH370事件			11			1	0.0357143
马航MH370事件	2009					1	0.0357143
马航MH370事件				45		2	0.0357143
马航MH370事件	97					1	0.0357143
马航MH370事件			9			1	0.0357143
马航MH370事件			70			1	0.0178571
马航MH370事件			4			1	0.0178571
马航MH370事件	2007					1	0.0178571
马航MH370事件				50		2	0.0178571
马航MH370事件			8	41		3	0.0178571
马航MH370事件				20		2	0.0178571
马航MH370事件				30		2	0.0178571
马航MH370事件				30		2	0.0178571
马航MH370事件			8	21		3	0.0178571
马航MH370事件						1	0.0178571

topic	evolutionTimes	entity	entityType	countRatio
马航MH370事件	(Null)	航班	Object	0.00546448
马航MH370事件	(Null)	时间	Object	0.00546448
马航MH370事件	(Null)	相关	Object	0.00491803
马航MH370事件	(Null)	国家	Object	0.00491803
马航MH370事件	(Null)	马航	Organization	0.00491803
马航MH370事件	(Null)	消息	Object	0.00437158
马航MH370事件	(Null)	记者	Object	0.00437158
马航MH370事件	(Null)	飞机	Object	0.00437158
马航MH370事件	(Null)	工作	Object	0.00382514
马航MH370事件	(Null)	客机	Object	0.00382514
马航MH370事件	(Null)	人	Object	0.00382514
马航MH370事件	(Null)	马来西亚航空公司	Organization	0.00382514
马航MH370事件	(Null)	乘客	Object	0.00382514
马航MH370事件	(Null)	信号	Object	0.00327869
马航MH370事件	(Null)	我国	Object	0.00327869
马航MH370事件	(Null)	人员	Object	0.00327869
马航MH370事件	(Null)	话	Object	0.00327869
马航MH370事件	(Null)	海域	Object	0.00327869

图 5.13 “马航 MH370”进化后话题信息的时间/实体要素相关信息

进化最终得到的“马航 MH370”话题本体内包含的要素信息统计如表 5.5 所示：

表 5.5 “马航 MH370”进化后话题本体的要素信息统计

要素信息类型	实例数量
Event	4
Time	35
Location	68
Entity	1105
Description	276
Predicate	1039
Extra	193

由于得到的话题本体内包含的要素信息数量庞大，而使用 Protégé进行本体 OWL 生成和可视化时的要素信息导入目前限于系统实现的程度需采取手动形式（存在 Jena、Protégé OWL API 等 OWL 本体数据操作 API），工作量过于巨大，因此在此未进行 10 次进化后的话题本体的 OWL 可视化。

5.5 实验结果分析

（1）话题本体概括话题的效果

最后进化所得到的“马航 MH370”话题本体中虽然存在一些与话题相关度不够高的要素信息实例，但按词频排在较前位置的实例大部分都能反映话题“马航 MH370”话题的语义特征。

（2）事件-话题相似度算法的效果

从话题本体的 10 次进化情况可看出，无论是来自新闻门户网站的语料事件，还是来自新浪微博的语料事件，10 个事件算得的事件-话题相似度没有一个超过 0.5。这说明本文设计的事件-话题相似度算法是较为严格的。

（3）不同平台的语料差异

来自新浪微博的语料事件，其事件-话题相似度普遍小于来自腾讯新闻的语料事件。即使通过降低进化阈值的方法使“新浪微博 4”得以成为相关事件，从而通过其进化信息使后续新浪微博的语料事件相似度得以较大幅度的提高，最终“新浪微博 8”的相似度仍然没有超越“腾讯新闻 1”。这可以说明，作为社交网络平台的新浪微博，其语料文本与作为

新闻门户网站的腾讯新闻相比，文本中存在更多与话题无关的信息，这与§3.1.2 中所作的分析结论是一致的。

（4）实例-话题相关度算法的必要性

可以看到，进化后的话题本体在未经过人工修正的情况下，含有最多要素信息实例的要素类型便是实体要素信息和谓语要素信息，其中有不少实例均是与话题相关性较弱的，无法反映话题的特征，比如实体要素的“消息”、“工作”，谓语要素的“说”、“到”。因此通过实例-话题相关度算法对各类型要素信息（特别是实体和谓语）实例的进一步筛选，是很有必要的。

（5）标注词性与要素类型的对应效果

从进化后的话题本体可以看出，目前 ICTCLAS 的二级词性标注与要素信息类型的对应不够完善。比如实体要素信息中出现了“相关”、“3”等实例。

第六章　总结与展望

6.1　总结

6.1.1 主要工作

本文主要研究网络舆情话题的本体表达及话题本体的构建与进化方法，从而得到一个能够在一定程度的人工监督下自适应地跟踪某特定网络舆情话题的系统。

本文的研究内容可以大致分为五个部分：

（1)介绍网络舆情分析、话题跟踪及其相关关键技术的研究意义以及国内外研究现状，并提出其中存在的问题以及本体相关理论和技术在此领域可能的利用方式；

（2）对舆情分析、文本挖掘、话题跟踪、本体进化的相关基础理论进行研究；

（3）分析网络舆情话题跟踪中需要关注的要素信息类型；

（4）进行网络舆情话题本体的构建与进化算法的方案设计，包括模型建立、文本新信息抽取和相似度算法等；

（5）进行实验系统的开发，使用真实的网络舆情话题“马航 MH370”的语料进行话题本体的构建与进化实验，并对结果进行分析。

6.1.2 存在的不足

本文的研究存在许多不足之处，按前文的章节内容顺序总结如下：

（1）事件文本信息抽取方案

前文已提到文本信息抽取方案中存在两个缺陷：时间要素信息的匹配不够完善；地点、实体、描述、谓语、扩展要素信息匹配时，使用的词性与要素类型的对应规则不够完善。此外，对于同义词（简写词或合并词）的处理、地点要素信息的父子关系、以及描述要素信息的倾向类型和倾向程度等信息也没有设计出相应的自动方法进行抽取。

（2）话题顶层本体的 OWL 建模

顶层本体的 OWL 建模中对于数据属性、对象属性及其属性特征和属性约束的使用不够完善，比如：Year、Month、Day 等类的严格实例取值范围未使用数据属性进行约束；对

于 Event 的 hasFromTime 的属性值个数限制为 0 到 n，而非更加严格的 0 到 1。

（3）话题本体进化的相似度算法设计

本体进化算法基于事件与话题模型，然而目前的相似度算法设计并未非常充分地利用这种模型，比如常用的向量空间余弦相似度，以及 TF-IDF 算法中 IDF 的使用也未考虑进算法中。

事件-话题相似度算法中，对于某些要素信息类型特有的信息利用不到位，比如时间要素信息具有年、月、日、时、分、秒六个更基本的组成，可在这种粒度上进行更细致的相似度算法考虑。此外地点要素信息的父子关系与具体程度也未纳入算法设计考虑之中。此外在算法中，如何将话题信息权重 topicTF 与事件信息权重 eventTF 的结合，给出综合性的信息权重，这一问题也未得到解决。

（4）系统实现

当前系统未能实现的重要功能主要有三个：实例-话题相关度算法，ICTCLAS 分词系统与系统的整合，本体数据的自动化处理。此外对整个系统的算法性能的优化目前未进行考虑。

6.2 展望

在当前研究工作总结的基础上，对后续可研究的内容进行一定的展望。

6.2.1 对已存在不足的完善

后续研究可对事件文本信息抽取方案、话题顶层本体的 OWL 建模、话题本体进化的相似度算法设计以及系统实现上的不足进行针对性的完善。

6.2.2 语料事件的自动采集

后续研究可考虑本文未涉及到的对语料事件的自动采集。目前一个基本的想法是：在本体中选择 TopicTF 值排在前几位的时间、地点、实体、谓语要素信息实例形成关键词集合，通过关键词集合抓取互联网上数个来源平台的网页，处理后留下正文及发布平台等信息，转码成 UTF-8，形成作为语料事件文本。

6.2.3 话题本体的应用

后续研究还包括如何对网络舆情话题本体进行较好的应用：

（1）使用话题本体对网络舆情话题的发展趋势进行分析和监控；

（2）利用本体的推理能力对话题本体中的信息进行深层次的语义关系挖掘，从而得到有价值的信息。

参考文献

[1] 北大方正技术研究院.以科技手段辅助网络舆情突发事件的监测分析—方正智思舆情辅助决策支 持系统[J].信息化建设,2005(10):50-52.

[2] http://www.cs.fudan.edu.cn/mcwil/

[3] http://www.trs.com.cn/product/product-om.html

[4] http://www.Autonomy.com.cn

[5] http://www.54yuqing.com/

[6] http://www.ninemax.com/cp_yuqingjiankong.html

[7] http://www.hisys.com.cn/Modules/Hisys/Import_Operation.aspx

[8] http://www.goonie.cn/products/2010/08/content287.html

[9] 许鑫,章成志,李雯静.国内网络舆情研究的回顾与展望[J].情报理论与实践,2009(03):115-120.

[10] http://wiki.mbalib.com/wiki/TDT

[11] JamesAllan, Jaime Carbonell, George Doddington et al. Topic Detection and Tracking Pilot Study: Final Report, In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, San Francisco, CA, Morgan Kaufmann Publishers,Inc,1998:194-218P

[12] Yiming Yang, Jaime Carbonell, Ralf Brownetal. Learning Approaches for Detecting and Tracking New Events, IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval ,1999.

[13] The 2002 Topic Detection and Tracking (TDT2002) Task Definition and Evaluation Plan, ftp://jaguar.ncsl.nist.gov//tdt/tdt2002/evalplans/TDT02.Eval.Plan.v1.1.ps

[14] Walls F, Jin H, Sista Setal. Probabilistic models for topic detection and tracking [A]. The Aeousties, Speech, and Signal Processing (ICASSP) [C]. 1999, Phoenix, 1999: 521-524.

[15] Allan J, Carbonell J, Doddington Getal. Topic detection and tracking pilot study: final report[R]. The DARPA Broadcast News Transcription and Understanding Workshop, San Francisco,

1998:194-218.

[16] Seokkyung Chung, Dennis McLeod. Dynamic topic mining from news stream data[J]. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, 2003, 2888: 653-670.

[17] Juha Makkonen, Helena Ahonen-Myka, Marko Salmenkivi. Topic Detection and Tracking with Spatio-Temporal Evidence[J]. Advances in Information Retrieval, 2003, 2633:549-563.

[18] Florian Holz, Sven Teresniak. Towards Automatic Detection and Tracking of Topic Change. In Proceedings of CICLing' 2010. pp.327~33.

[19] Claudiu Cristian Musat, Julien Velcin, Marian-Andrei Rizoiu, Stefan Trausan-Matu. Concept-Based Topic Model Improvement. In Proceedings of ISMIS Industrial Session' 2011. pp.133~142.

[20] T Strzalkowski, G C Stein and G B Wise. GE. Tracker: A Robust, Lightweight Topic Tracking System[A]. In Proceedings of the DARPA Broadcast News Workshop [C]. San Francisco: Morgan Kaufmann, 1999.

[21] J P Yamron, S Knecht, and P V Mulbregt. Dragon’s Tracking and Detection Systems for the TDT2000 Evaluation [A]. In Topic Detection and Tracking Workshop [C]. USA: National Institute of Standard and Technology, 2000, 75–79.

[22] J Allan, V Lavrenko, D Frey, V Khandelwal. UMass at TDT 2000 [A]. Proceedings of Topic Detection and Tracking Workshop [C]. USA: National Institute of Standar and Technology, 2000, 109-115.

[23] N Lester, HE Williams. TDT2001 Topic Tracking at RMIT University[J]. The Topic Detection and Tracking (TDT) Workshop, 2001.

[24] W Lam, S Mukhopadhyay, J Mostafa, and M Palakal. Detection of Shifts in User Interests for Personalized Information Filtering [A]. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Konstanz: Hartung-Gorre Verlag,1996, 317-325.

[25] J Carbonell, Y Yang, J Lafferty, R D. Brown, T. Pierce, and X. Liu. CMU Report on TDT-2: Segmentation, Detection and Tracking [A]. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop [C]. San Francisco: Morgan Kauffman, 1999, 117-120.

[26] Y Lo, J L Gauvain. The LIMSI Topic Tracking System For TDT 2002 [A]. Topic Detection and Tracking Workshop [C]. Gaithersburg, USA, 2002.

[27] 王会珍，朱靖波，季铎. 基于反馈学习自适应的中文话题跟踪[J]. 中文信息学报，2006, 20(3):94-100.

[28] 郑伟，张宇，邹博伟等. 基于相关性模型的中文话题跟踪研究[A]. 第九届全国计算机语言学学术会议论文集[C]. 大连, 2007: 558-563.

[29] 贾自艳，何清，张海俊等. 一种基于动态进化模型的事件探测和跟踪算法[J]. 计算机研究与发展,2004, 41(7): 1273-1280.

[30] 金珠. 基于知网的话题跟踪和倾向性跟踪研究[D]. 大连:大连理工大学,2005.

[31] 金珠. 基于 HowNet 的话题跟踪及倾向性分类研究[J]. 情报学报, 2005, 24(5): 555-561.

[32] 焦健，瞿有利. 知网的话题更新与跟踪算法研究[J]. 北京交通大学学报，2009，33(5): 132-136.

[33] 宋丹. 基于语义和链接的话题跟踪方法[D]. 大连:大连理工大学, 2007.

[34] 张辉，周敬民，王亮，赵莉萍. 基于三维文档向量的自适应话题追踪器[J]. 中文信息学报, 2010, 24(5):70-76.

[35] 朱恒民，张相斌. 基于链接网络图的互联网舆情话题跟踪方法[J]. 情报学报，2011, 30(12):1235-1241.

[36] 朱恒民，李青. 面向话题衍生性的微博网络舆情传播模型研究[J].现代图书情报技术，2012(5)：60-64

[37] 刘炜，李明，杨合立. 基于本体的话题检测与跟踪技术[J]. 甘肃科技, 2011, 27(22): 42-45.

[38] 刘毅. 略论网络舆情的概念，特点，表达与传播[J]. 理论界, 2007 (1): 11-12.

[39] 姜胜洪. 网络舆情热点的形成与发展，现状及舆论引导[J]. 理论月刊, 2008 (4): 34-36.

[40] 曾润喜. 网络舆情管控工作机制研究[J]. 图书情报工作, 2009, 53(18): 79-82.

[41] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429.

[42] Zhang Xiaoyan, Wang Ting, Chen Huowang. Story link detection based on dynamic information extending[C]//IJCNLP,Hyderabad,2008: 40-47.

[43] Studer, Rudi, Riehard Benjamins and Dieter Fensel. Knowledge Engineering: Prinei Plesand

Methods. Data and Knowledge Engineering. Vol.25, 1998(l~2): 161~197.

[44] Gómez-Pérez A, Corcho O. Ontology languages for the semantic web[J]. Intelligent Systems, IEEE, 2002, 17(1): 54-60.

[45] 赵昭. 本体自动构建技术研究及其在教学中的应用[D]. 西安: 西南交通大学, 2011.

[46] Gruber T R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human Computer Studies, 1995.

[47] 郭嘉琦. 领域本体的构建及其在信息检索中的应用研究[D]. 北京: 北京邮电大学, 2007.

[48] 乔卫. 基于领域本体的 XML 语义信息抽取的研究与实现[D]. 武汉: 武汉理工大学, 2009.

[49] 王宇阳. 基于本体进化的自适应中文话题跟踪算法研究[D]. 南京: 南京航空航天大学, 1996.

[50] 杜小勇, 李曼, 王珊. 本体学习研究综述[J]. 软件学报, 2006, 17(9): 1837-1847.

[51] McGuinness D L, Van Harmelen F. OWL web ontology language overview[J]. W3C recommendation, 2004, 10(2004-03): 10.

[52] Kavalec M, Svátek V. A study on automated relation labelling in ontology learning. In: Buitelaar P, Cimiano P, Magnini B, eds. Ontology Learning from Text: Methods, Evaluation and Applications. Amsterdam: IOS Press, 2005. http://nb.vse.cz/~svatek/ olp05.pdf

致谢

本论文的顺利完成，离不开导师马静教授对我的悉心指导与鼓励。感谢我的父母，一直默默支持着我做自己想做的事。感谢母校对我的培育，带给我难忘的四年大学时光。最后也感谢这个不否定过去、不放弃未来，将“给岁月以生命，而非给生命以岁月”这一人生信条贯彻到底的自己。