

doi:10.3969/j.issn.1001-893x.2014.06.007

引用格式:陈天堂. 文本信息中的目标动向感知技术[J]. 电讯技术,2014,54(6):730-734. [CHEN Tian-ying. Target Activity Awareness on Text Information[J]. Telecommunication Engineering,2014,54(6):730-734.]

文本信息中的目标动向感知技术*

陈天堂**

(中国西南电子技术研究所,成都 610036)

摘 要:通过建立目标动向的本体模型以及采用基于规则的方式来对文本信息中的目标动向描述语句进行语义分析,完成目标动向的识别和关联语义信息抽取;在提取大量的文本信息中的目标动向信息上建立目标与时空的关联关系,有利于文字信息分析人员感知目标在时空范围内的态势。通过对目标的动向信息抽取和结构化的关联关系转化及分析,可以为情报分析人员提供锁定目标的态势数据,并通过目标动向情况预测锁定目标的活动趋势、预警分析等。

关键词:文本信息;信息抽取;目标动向感知;动向识别;语义信息关联;动向预测

中图分类号:TP391.1 **文献标志码:**A **文章编号:**1001-893X(2014)06-0730-05

Target Activity Awareness on Text Information

CHEN Tian-ying

(Southwest China Institute of Electronic Technology, Chengdu 610036, China)

Abstract:The dissertation mainly discusses semantic analysis of target activity description sentences in text information by establishing ontology model of target activity in accordance with rule-based technologies to realize recognition of target activity and relational semantic information extraction. On the basis of successful extraction of target activity information from a huge amount of data for establishing relations between target and time-space, it is advantageous for text information analyzers to sense the situation of the target in the time-space domain. By extraction of target activity information, conversion and analysis of structural relationships, it is possible for intelligence analyzers to obtain situation data sufficient to lock down their target, and predict or early warn target activity trends by correspondent historical target activities.

Key words:text information;information extraction;target activity awareness; recognition of activity; related semantic information;activity prediction

1 引 言

大数据时代的到来使得人们获取的信息越来越多,在浩如烟海的信息中提取有用的信息是当前信息处理领域研究的热点,同时也是大数据环境下知识获取的重点。对于文本信息而言,如何在非结构化的文本中提取有用的信息、在对信息进提取后的凌乱的信息中找到信息之间的关联和关系、对提取后的信息完成预测和分析都是文本信息处理和分析

领域急需研究的技术。目前,国内外的信息抽取技术主要集中在命名实体识别和事件检测与识别两大方向,针对文字情报处理领域而言这两个技术远远不能满足需求。在文字信息中只是识别命名实体而不能智能化地提取实体之间的关系对于文本信息分析人员而言实用价值有限;针对文本信息进行事件主题识别只能对已经发生的事件或主题进行统计整理,不能为文字信息分析人员提供“目标-时空”之

* 收稿日期:2014-01-28;修回日期:2014-04-26 Received date:2014-01-28;Revised date:2014-04-26

** 通讯作者:ctiany@163.com Corresponding author:ctiany@163.com

• 730 •

间的现有关系来进行目标的态势信息及预测和预警。因此,本技术主要应用于文本信息处理和分析中,完成对文字信息中目标活动的识别和提取,并将非结构化的目标活动文字描述转化为三元组的语义关系进行存储。本技术中目标活动的识别及提取、目标活动的语义表示和存储是关键技术。

信息抽取技术在欧美等国迅速发展起来,在美国已经有不少公司和大学致力于信息抽取的研究和产品的开发,比较著名的公司和大学如 Cymfony 公司、Bhasha 公司、Linguamatics 公司、Revsolutions 公司,纽约大学、加利福尼亚大学、尤他州大学、华盛顿大学等。在英国、德国、意大利、爱尔兰等国家也有机构在进行相关研究,并且已经有很多成型的系统出现^[1]。国内的研究大多面向特定的领域,如金融领域、灾难性事件等,且面向完整的中文信息抽取系统的资源建设和评测目前在国内还没有大规模开展。国内的研究热点主要集中在事件抽取领域,成果主要有:上海交大和德国人工智能研究中心(DF-KI)合作,面向股票和投资领域研究了“基于信息抽取和文本生成的多语种信息检索模型”技术;北京大学的孙斌、袁毓林等进行了事件抽取的研究;清华大学的周剑辉对金融领域的事件抽取进行了研究。

本技术不属于事件抽取,而是根据文本信息处理业务的目标态势智能提取业务需求而衍生的一种目标活动感知技术;运用命名实体识别技术、关系抽取等成熟技术并结合本体语义技术来完成业务需求的抽取任务,最终达到对文本信息中目标态势信息的识别和抽取。本技术首先采用语义本体技术对目标动向进行本体建模,通过本体模型来指导完成文本中相关命名实体的识别,命名实体识别采用传统的统计结合模型的方式来完成;再运用句法分析来通过命名实体在句子中的句法成分来建立命名实体之间的语义关系,即形成目标动向语义信息;最后将目标动向语义信息进行存储;目标动向语义信息采用三元组的方式存储到知识库中便于进行后续统计、推理等处理。

2 相关概念及系统框架

目标是指进行作战或者采取行动时需要考虑的一个实体或者一个物体,它可以是为支持指挥员作战目标与作战意图所采取行动而识别出的地域、集群、设施、部队、装备、能力、功能、个人、人群、系统、实体或者行为。本文研究的目标主要是个人,通过对个人的行为活动信息的抽取可以将此技术延伸并

覆盖所有的目标类别。

目标动向是指目标的行为活动。行为是指人们一切有目的的活动,它由一系列简单动作构成,在日常生活中所表现出来的一切动作的统称。本文主要研究人物目标的行为活动。为了在海量的文本信息中提取人物目标的活动情况,并为文本信息分析人员较好地掌握和预测人物目标的活动趋势,本文通过研究发现,人物目标活动中的言论、出行两大类别的活动是最具有价值的情报信息。人物目标发表的言论代表该人物目标的立场,对推测某重要人物在某事件上的下一步行为具有重要的预测作用。同样,人物目标的出行活动代表了该人物目标正在进行或即将预谋某事件的意图,对预测某人物目标的活跃度和行为意图提供线索。因此,本文中的人物目标动向的抽取主要集中与人物目标的言论和出行两种活动行为。

本文主要研究文本信息中人物目标的动向识别与关联信息的抽取技术。通过对文本信息中人物目标的动向信息识别后,建立人物目标“动向—时空”关系、“动向—目标”关系进行关联构建,从而形成人物目标在时空范围内的态势信息,有效地为文本信息处理和分析人员提供锁定目标的态势及预警分析、通过目标动向情况预测锁定目标的活动趋势等。

本文采用构建活动本体的方法来对文本信息中的目标动向信息进行识别和抽取。采用本体的方式一是便于很好地扩展抽取的体系结构,二是本体建模使得动向具有丰富的语义信息。本体模型可以更加有效地形成抽取规则,以及为目标动向的识别提供语义信息。目标动向抽取的系统流程如图 1 所示。

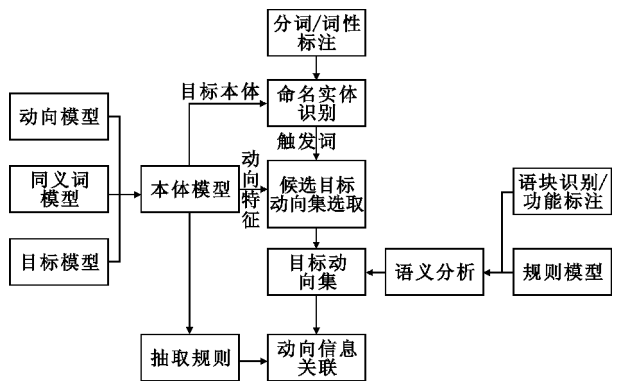


图 1 基于目标本体的动向抽取流程
Fig. 1 Flow chart of activity extraction based on target ontology

首先构建本体模型,本体模型包括不同目标的动向模型、目标模型以及同义词模型。动向模型主要是完成动向的定义,明确动向识别的范畴以及动

向的语义信息;目标模型主要是针对动向的主体(及对象)进行本体建模;同义词模型主要是动向模型的扩展。在本文中,如出行动向中的“访问”是一个动向本体,人物、组织、设施等属于目标本体,“访问”动向本体中的描述词“拜访”、“拜见”等属于同义本体。

在本体模型构建完成的基础上,对文本信息进行如下处理:

- (1)对文本信息进行词法分析,即分词和词性标注;
- (2)根据本体模型中的目标本体进行命名实体识别;
- (3)通过命名实体识别和动向本体的特征进行候选目标动向集的选取;
- (4)对候选动向集进行语义分析,语义分析主要包括句法分析和语义关联标注;
- (5)通过本体模型形成的抽取规则对目标动向集进行信息抽取,完成动向关联信息的抽取和存储。

3 本体模型的构建

本体建模方法有 TOVE 方法、IDEF-5 方法、骨架法、METHONTTLOGY 法、SENSUS 法、循环获取法以及七步法等。不管是采用哪种方法,其最重要的主体部分包括:一是确定本体建模的目标和范围;二是本体识别和分类,即识别领域中概念和关系;三是本体元建模,即概念模型定义同类概念的层次关系,关系模型定义概念之间的关联关系及关系的约束等,实体模型定义概念实例化的实例之间关系和约束;四是本体集成,尽可能复用和合成来自其他领域的概念和术语;五是本体评价,主要是对本体正确性和有效性进行评估。

本文主要研究人物目标的动向,并将人物动向认为最有价值的两种动向定义为 人物目标的言论和出行活动行为。因此,本文确定动向的范畴为 人物目标的言论、人物目标的出行。人物目标的动向需包含的要素如下:

目标动向:<发生地点、发生时间、涉及地点、行为、主动者、参与者>

概念模型如下:

目标动向:={ 人物目标动向、机构目标动向……}

人物目标动向:={ 言论、出行}

言论:={ 演说、申明}

出行:={ 访问、旅游、参观、参加}

针对不同的动向构建本体如图 2 所示。

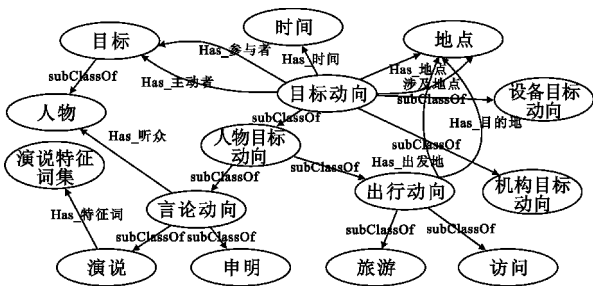


图 2 动向本体模型语义网络图
Fig.2 Semantic network graph of activity ontology model

4 目标动向识别

人物目标动向识别是指在从文本信息中识别动向集并提取动向集的过程。为了完成在文本信息中进行动向集的提取,给出两个定义。

定义 1:文本 Document:={S₁, S₂, ..., S_n} ,其中 S_i表示文本 Document 中的一个句子。句子的区分是由“。”、“?”、“!”等分割的文本。

定义 2:动向集 Activity_Sentence:={a₋S₁, a₋S₂, ..., a₋S_n} ,a₋S_i表示动向集中的一个动向句。每个 a₋S_i ∈ Document,且 a₋S_i中具有动向特征词,符合动向本体的语义描述。

a₋S_i表示一个描述动向的句子,其结构用句法树来进行表示。对于每个 a₋S_i都可以用句法树标识出其句子成分、短语结构以及词性等。句子成分如主语、谓语、宾语等,本文根据语块标注规范(V3.0)的语法来描述句子成分,其中 S 为主语语块、P 为谓语语块、O 为宾语语块;短语结构如名词短语、动词短语等,用 NP、VP 等表示;词性及单词的词性标注,如名词、动词标注为 N、V 等^[2]。如一个动向句为“温家宝近期访问柬埔寨”,其词性标注和功能标注如下:

词性标注:

温家宝/nr 近期/t 访问/v 柬埔寨/ns

命名实体标注:

温家宝/nr 近期/t 访问/v 柬埔寨/ns

语块功能标注:

[S 温家宝/nP] [D 近期/t] [P 访问/v] [O 柬埔寨/nP]

本文的人物目标动向识别处理流程如图 3 所示。

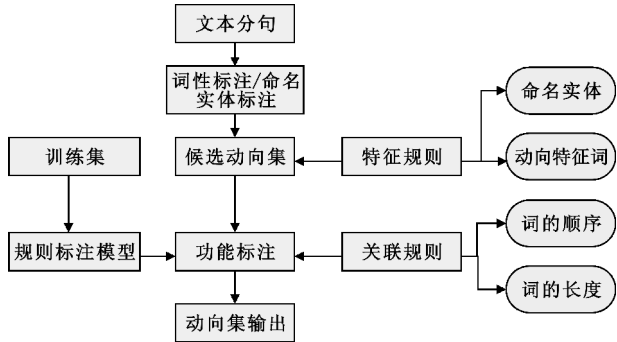


图 3 动向集识别流程
Fig.3 Flow chart of activity set recognition

动向集 的识别其中一个重要环节是语块功能的自动标注或识别,常用的方式是选取一个中文树库作为训练语料库,树库的主要目的是表示句子成分的结构属性^[3],用以分析中文各种语法构成的句法结构,并采用条件随机域模型(CRF)对其进行训练后进行功能块识别和标注。本文采用规则的方式来对句子的语块进行标注^[4],其方法是从新闻文档中选取包含目标动向集的文章集作为训练样本,将这些训练文本进行分词和词性标注,并在此基础上进行特征规则的提取,特征规则包括语块功能边界规则划分、语块关系规则,最后通过规则来对测试样本进行功能的划分和标注。

- 动向集识别流程如图 3 所示,处理步骤如下:
- (1) 首先对文本信息进行分句,将文本分为 n 个待处理的句子集 $\{S_i | i = (1, 2, \cdots, n)\}$;
 - (2) 对每个句子 S_i 进行特征规则的匹配。建立候选特征集 Activity_S,特征规则为: IF S_i 中有命名实体 And S_i 中有动向特征词;将 S_i 存放到 Activity_S 中;
 - (3) 将候选 Activity_S 中的每个候选动向句 a_{S_i} 进行功能块识别。将候选动向句 a_{S_i} 的词性标注结果作为功能标注的输入,得到每个候选动向句 a_{S_i} 的功能块标注。功能块标注特征规则如下:
 - 1) 语块边界规则:即通过词-词性-句子成分的关系来简单判断句子成分。如:主语语块是名词性成分,一般放在谓词前,因此可以把放在谓词前的名词、代词和修饰名词的形容词等划为一个主语语块;
 - 2) 语块关系规则:通过训练样本来提炼语块之间的关系,如主语省略的情况下,是否在该句中关联上主语;
 - (4) 对每个候选动向句 a_{S_i} 的功能块进行分析。运用关联规则来进行识别动向句的正确性。关联规则模式如下:

- 1) IF 谓语中包含动向特征词 And 谓语的主语中包含命名实体, a_{S_i} 为动向句;
- 2) IF 谓语中包含动向特征词 And 谓语的宾语中包含命名实体, a_{S_i} 为动向句。

5 目标动向关联信息抽取

目标动向句的关联信息抽取主要是将动向句中的地点、时间、主动者、参与者等信息提取出来,形成有结构的数据进行存储。动向句中的时间、地点、主动者、参与者属于命名实体,对动向句信息的抽取可以看成是建立实体与动向之间的关联的过程。研究

发现,通过动向句中的句法成分可以给出成分中相关实体之间的关联关系。构建动向与实体之间的关联,使得每个动向之间通过相同的实体具有关联性,从而形成达到通过目标动向来对锁定目标进行态势感知和预测的目的。

如动向句:温家宝总理于 2 月 28 日,在北京大学发表了演说。

句子成分:
[S 温家宝/nr 总理/n][D 于/p 2 月 28 日/nt],[D 在/p 北京大学/nt][P 发表/v 了/u 演说/v]。
句法树和关联关系如图 4 所示。

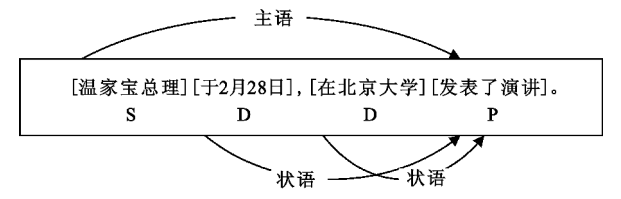


图 4 示例句法关系图
Fig. 4 A syntax relation graph example

谓语:演说
语义关系:
演说-主语-温家宝;演说-补语-于 2 月 28 日,演说-补语-在北京大学
2 月 28 日为时间实体,北京大学为地点实体。
语义抽取:
(演说-subject-温家宝;演说-时间-2 月 28 日;演说-地点-北京大学)
关联关系生成:

=>(动向句 1,has_主动者,温家宝)(动向句 1,has_时间,2 月 28 日)(动向句 1,has_发生地点,北京大学)

根据动向本体模型的描述,动向关联信息抽取的过程就是动向实例化的过程,即将动向本体模型中描述的动向属性进行值的填充。实例化示例如图 5 所示,文本用三元组来描述动向模型及实例化结果。

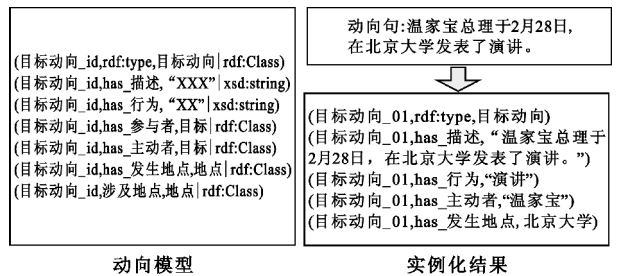


图 5 根据动向模型实例化动向关联信息
Fig. 5 Instantiate activity relationship information according to activity model

通过实体-目标动向之间的关联,使得可以将抽取出的信息在时空范围内进行统计,并对目标进行简单的预测和态势评估。通过对文本信息的目标动向的抽取和关联关系的建立可以做如下一些态势感知:

(1)某目标对象在时空范围上的路线信息。快速分析并展示某人物目标的路线信息,掌握其活动的行径路线对其动向意图提供基础依据;

(2)某目标对象的活跃度评估。通过文本信息分析和检测某目标对象的活动频率,通过其活动频率来推断其活动意图。如某人物目标最近一段时间的活跃度较前几个月的活跃度明显提高,那么可以预示某人物目标最近可能会参加某些事件,建议相关部门对其进行监控;

(3)通过对某区域上所有目标的活动分析目标之间的潜在关联关系和活动意图。如某区域内出现了较多的人物目标,这些人物目标之间如果相互认识或有关系,那么可以预判该区域内将会出现某组织的群集事件。

6 结 论

本文主要研究了文本信息中目标动向的识别以及目标动向的语义表示等技术,完成对文本信息中目标的动向感知,将目标动向通过时间、地点、目标实体进行关联后实现动向之间的关联。本文重点描述了动向本体模型的构建、目标动向的识别以及动向关联信息的抽取和建立技术。通过在文本信息中识别目标动向的基础上建立目标与时空的关联关系,将有利于文字信息分析人员感知某目标在时空范围内的态势。本技术运用语义技术对目标动向的语义信息进行了保留,既可以为分析人员提供目标关系数据的自动化处理,又可以采用智能相关技术完成语义数据的推理及信息挖掘、知识发现,如对抽取的目标动向信息通过“目标-时间;目标-地点”等关联关系完成路径推理和规则推理^[5],发觉目标之间的潜在关联关系;还可以在抽取后的语义数据上采用统计分析等技术完成预测、预警等,如某目标历史的运动轨迹与最新轨迹是否具有较大差异,如果具有较大差异则可能属于异常行为,系统可辅助分析人员进行预警提示。

本技术中主要采用句法树及句子成分来确定实体之间的关系,因此对于文本信息中复杂的句式以

及不符合句法规范的句子将不能准确进行关系的识别,并且复杂的句式在工程应用中完成句法分析所耗费的时间代价较大,这是本技术在后续研究中应该重点研究的方向。此外,需进一步研究语义关系数据对预测、预警分析技术的支撑力度及相关技术。

参考文献:

- [1] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨:哈尔滨工业大学工学,2008.
TAN Hong-ye. Research on Chinese Event Extraction [D]. Harbin: Harbin Institute of Technology, 2008. (in Chinese)
- [2] 周强,赵颖泽. 汉语功能块自动分析[J]. 中文信息学报,2007,21(5):18-24.
ZHOU Qiang, ZHAO Ying-ze. Automatic Parsing of Chinese Functional Chunks[J]. Journal of Chinese Information Processing, 2007,21(5):18-24. (in Chinese)
- [3] 周强. 汉语句法树库标注体系[J]. 中文信息学报,2004,18(4):1-8.
ZHOU Qiang. Annotation Scheme for Chinese Treebank [J]. Journal of Chinese Information Processing, 2004, 18 (4):1-8. (in Chinese)
- [4] 王呈艳. 基于规则抽取的汉语语块识别[D]. 保定:河北大学,2011.
WANG Cheng-yan. Chinese Chunk Identification Based on Rule Extraction [D]. Baoding: Hebei University, 2011. (in Chinese)
- [5] 陈天堂,苏智慧. 基于语义推理的文本信息关联关系分析技术[J]. 电讯技术,2014,54(1):68-73.
CHEN Tian-ying, SU Zhi-hui. Text Information Relationship Analysis Based on Semantic Reasoning[J]. Telecommunication Engineering, 2014, 54 (1) : 68 - 73. (in Chinese)

作者简介:



陈天堂(1982—),女,四川彭州人,2010年于中国科学院成都计算机应用研究所获工学博士学位,现为工程师,主要研究方向为情报智能处理与知识工程技术。

CHEN Tian-ying was born in Pengzhou, Sichuan Province, in 1982. She received the Ph. D. degree from Chengdu Institution of Computer Applications, Chinese Academy of Sciences, in 2010. She is now an engineer. Her research concerns intelligent information analysis and knowledge engineering.

Email: ctiany@163. com