doi:10.3969/j.issn.1001-893x.2014.01.013
引用格式:陈天莹,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):67-73. [CHEN Tian-ying,SU Zhi-hui. Text Information Relationship Analysis Based on Semantic Reasoning [J]. Telecommunication Engineering,2014,54(1):67-73.]

基于语义推理的文本信息关联关系分析技术*

陈天莹**,苏智慧

(中国西南电子技术研究所,成都 610036)

摘 要:重点论述了文本信息中的知识发现及潜在关联分析技术。采用本体建模技术、信息抽取技术以及知识库上的语义推理技术等来完成并实现文本信息的关联关系发现和分析,最后给出了语义关联分析技术在文本信息处理系统中的应用,并简要描述了系统的处理流程。在信息处理领域的大数据环境下,该技术有利于信息分析人员快速获取关联线索,辅助完成信息挖掘,为指挥决策提供更全面的信息支持。

关键词:大数据;文本信息;数据挖掘;语义技术;信息抽取;关联分析
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1001-893X(2014)01-0068-06

Text Information Relationship Analysis Based on Semantic Reasoning

CHEN Tian-ying,SU Zhi-hui

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

Abstract:This paper mainly discusses knowledge discovery and potential relationship analysis in text information. By utilizing ontology modelling,information extraction and semantic reasoning technique running on knowledge bases,relationship discovery and analysis of text information are realized. Finally,the typical practice of semantic relation analysis in text information analysis system is given. Under the big data environment of information processing domain,the technique is beneficial for information analyzers to obtain relational clues quickly,and facilitates information mining which offers steady support for command and decision-making.

Key words:big data;text information;data mining;semantic technology; information extraction; related analysis

1 引 言

基于文本信息的数据挖掘和知识发现是当前信息处理的一大热点。文本信息中蕴含的潜在信息非常丰富,信息之间既具有语义性又具有关联性。文本信息的无结构性导致计算机对其理解、处理、分析较为受限,目前主要依托人工阅读、编辑、分析的方式来进行处理。因此,如何快速从文本信息中找到信息之间的所有直接和潜在关联,并快速对关联信息进行分析是辅助文本信息分析人员工作的重要技术。

关联关系属于知识发现的范畴,分别在数据挖掘和文本挖掘中有不同的内涵和处理技术,针对不同领域、不同信息处理对象其涉及的关键技术也大有不同。

在数据挖掘中的关联分析主要是指关联规则挖掘,它由 Agrawal 等人[1-2]提出,其处理对象主要是海量的有结构的数据库数据。关联规则挖掘主要是在有结构化的数据集上发现数据集中项之间的联系。现已发表的研究论文包括确定性关联规则的挖掘、量化关联规则的挖掘、增量式关联规则的挖掘、广义关联规则的挖掘等。最著名的关联规则算法是 Apriori[3]算法,其思想是通过多次迭代找出所有的频繁项目集。关联规则主要运用于交易数据库中发

* 收稿日期:2013-11-05;修回日期:2014-01-15 Received date:2013-11-05;Revised date:2014-01-15
** 通讯作者:ctiany@163.com **Corresponding author**:ctiany@163.com

现各数据项之间的关联关系,从而生成形如“$X \Rightarrow Y$”的规则。

文本挖掘中的关联分析主要是指知识关联,它是利用各项智能分析技术对非结构化文本进行信息提取、存储、分析后获取有用知识和信息的技术。文本信息中的关联性指对象之间的关联性,如(A 和 B 相关)、(B 和 C 相关)、(C 和 D 相关);检索希望实现 A 到 D 的查询,推理希望告诉用户 A 和 D 具有路径关联关系,这是人们基于语义的一种推理过程。同时,知识之间存在很多有用的关联性,在知识组织中,如果将知识视为一种网状结构,那么这种特定意义上的知识就是由众多的结点(知识)和结点间关系组成的[4]。有人将知识关联定义为,知识关联就是指大量的知识点之间存在的知识序化的联系,以及所隐藏的、可理解的、最终可用的关联,它超出信息检索的范畴,主要是揭示知识之间隐含的关联与寓意,发现更有价值的知识[5]。

文本信息的潜在关联关系分析技术主要引入语义技术,将信息抽取处理的结果采用本体进行知识表示,并结合知识检索技术、推理技术来实现文本信息挖掘。当前,国内研究将文本挖掘的方法集中在分类、聚类、机器学习等传统技术上,对信息抽取的结果采用关联规则提取的方式完成文本信息的挖掘,而本技术在信息抽取结果表示、处理上均采用语义技术,保留数据间的语义关系,在语义关系上进行知识检索和推理实现潜在关联关系发现。

2　文本信息中目标的关联关系分析

技术以文本信息的关联关系分析为研究对象,主要模拟文本信息处理和分析人员的需求,将信息的关联关系分析限定为目标的关联关系分析和潜在关联关系发现。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得地域、集群、设施、部队、装备、能力、功能、个人、人群、系统、实体或者行为[6],研究的目标主要是文本信息中的个人、设施、地域、机构。为了完成文本信息中目标的关联关系分析,首先,采用基于本体的信息抽取技术对文本内容进行信息提取,获取语义关系;其次,将提取的信息和关联关系存储到知识库中;最后,在知识库上进行知识检索和推理完成两种关联关系的分析。

2.1　关联数据抽取

本技术采用基于本体的信息抽取技术来完成关联数据和关联关系的获取。关联关系抽取首先要确定抽取信息的范畴,即确定哪些信息是有价值的。抽取对象是目标对象及目标对象之间的关系。经过仔细分析,在文本信息中目标对象之间的关联关系通常是和目标的动向情况进行直接关联的。目标动向事件是指目标的行为,例如目标的参与活动、发表言论等,将动向事件简称为动向。研究的范畴定义如下:

目标:={人物、机构、设施、地域}

动向事件:={时间<发生时间、涉及时间>、地点<发生地点、涉及地点>,参与者<目标对象>,内容<文字描述>}

因此,“目标-动向”是目标关联的重要信息,其关系图及示例如图 1 所示。

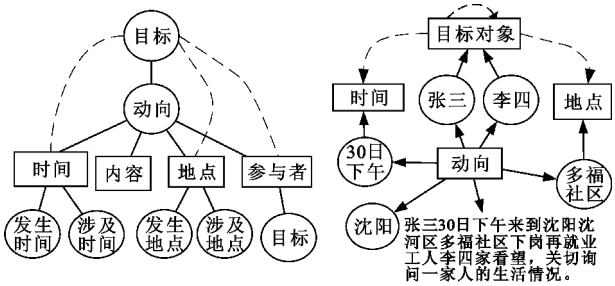

图 1　目标对象-事件”关系图及示例

Fig. 1 Diagram of target-event relationship with an example

由图 1 可以看出,目标的关联关系包括“目标-动向”、“动向-时间”、“动向-地点”以及间接的“目标-时间”、“目标-地点”、“目标-目标”6 种关系。文本采用基于本体的信息抽取技术来提取关联关系,流程如图 2 所示。

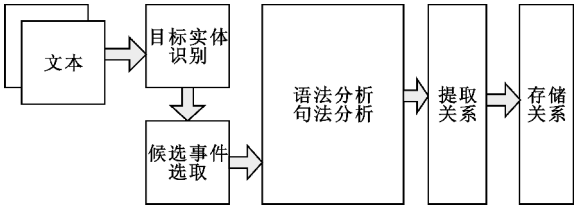

图 2　关联关系抽取流程

Fig. 2 Relationship extraction workflow

信息抽取首先对待处理文本进行目标实体识别,将目标实体识别的位置和分句结果相结合选取候选事件,为保证动向事件的可读性和完整性,我们将一个完整的包含动向事件的语句作为一个动向;在候选动向事件中进行语义分析,语义分析主要包括语法分析和句法分析,当候选动向事件包含的要素满足事件定义时,将其确定为动向事件,简称动向;将动向事件按照本体模型进行关联关系提取;最后将提取出来的关系按照本体模型的 schema 进行存储。

2.2　关联本体模型构建

本体模型的构建是信息抽取、知识库存储、知识检索和知识推理的依据。下面重点介绍如何对文本信息中的目标对象及目标对象关联关系进行建模。

首先,确定领域本体的建模范围,即建模对象(概念)有哪些,并对其关系进行描述和建模。本研究中的概念和关系如下:

Concept(概念):={时间、地点、目标、动向}

Relation(关系):={ 动向-时间、动向-地点、动向-目标、目标-时间、目标-地点、目标-目标}

其次,分别对 Concept 概念和关系进行建模。本体模型分为两个部分:一个是对概念及概念之间关系的描述,在描述逻辑中通常称为 TBox;另一个可以简单看成是对 TBox 进行实例化后的关系模型,称为 ABox。采用 Topbraid Composer 本体建模工具进行建模。

(1)概念模型

概念模型按照本体构建的标准和规范,主要定义了 Class,以及 Class 之间的分类关系。由图 3 可看出,我们定义了目标、动向 2 个 Class,并在目标下细分人物、机构、设施、地点 4 个子类。如此层层细分,将我们所需要研究的概念分层分类进行表示。

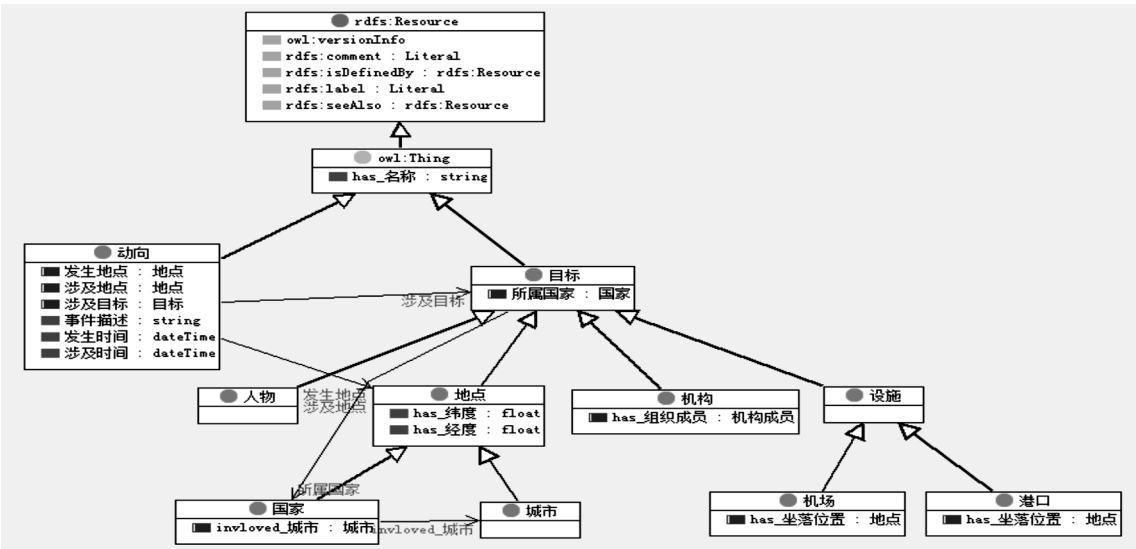

图 3 概念模型图

Fig. 3 Diagram of conceptual model

(2)关系模型

如图 4 所示,关系模型同样是在本体构建得标准和规范下,定义每个 Class 之间的关系,以及这些关系的数据模型和逻辑描述模型。所有定义规范遵循 W3C 的规范标准,同时引用了 RDF/RDFs、OWL 标准。关系模型表如表 1 所示。

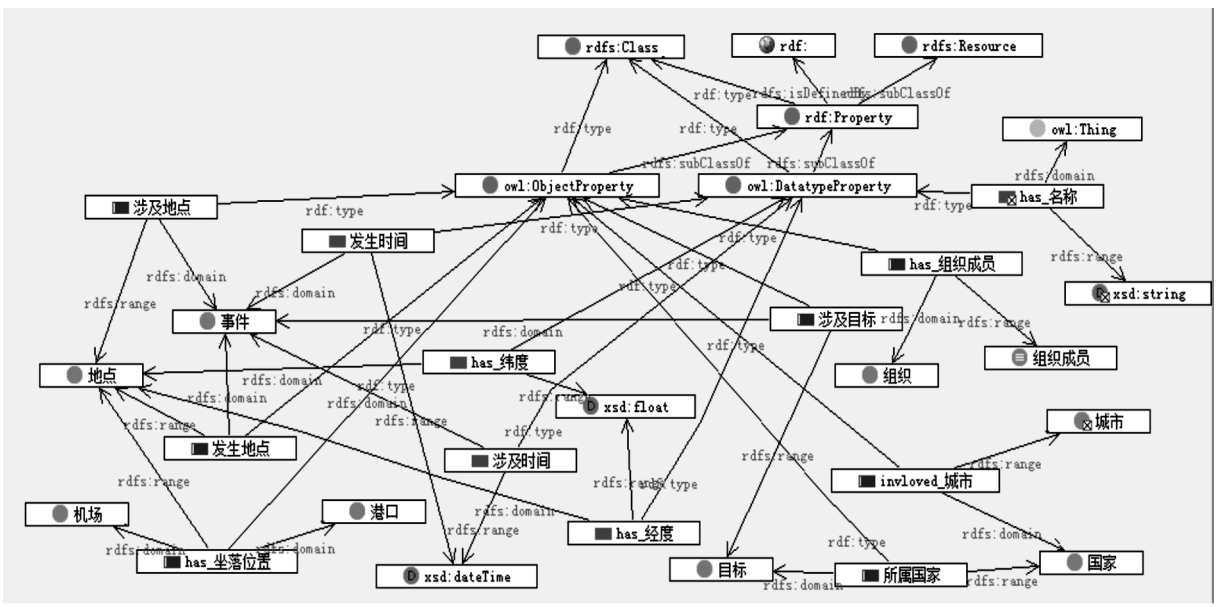

图 4 关系模型图

Fig. 4 Diagram of relation model

表 1　关系模型表
Table 1 Table of relation model

名称	类型	domain	range
涉及地点	Owl:ObjectProperty	class:事件	class:地点
发生时间	Owl:DatatypeProerty	class:事件	xsd:dataTime
has_经度	Owl:DatatypeProerty	class:地点	xsd:float
has_纬度	Owl:DatatypeProerty	class:地点	xsd:float
发生地点	Owl:ObjectProperty	class:事件	class:地点
涉及时间	Owl:DatatypeProerty	class:事件	xsd:dataTime
所属国家	Owl:ObjectProperty	class:目标	class:国家
involved_城市	Owl:ObjectProperty	class:国家	class:城市
has_组织成员	Owl:ObjectProperty	class:组织	class:组织成员
涉及目标	Owl:ObjectProperty	——	class:目标

2.3　关联检索及推理

关联检索及推理是在知识库的基础上,运用知识检索技术和知识库推理技术来对知识库中的知识进行关联关系挖掘和发现的一种基于业务驱动的应用性技术。关联分析主要解决目标的知识检索、目标的路径关联分析和目标的潜在关联关系发现三个方面。

目标的知识检索区别于关键词检索的不同在于,关键词检索使用户只能查询哪些文本中出现了该目标,返回的结果集大,从结果集中需要人工定位后通过上下文获取到该目标的信息;目标的知识检索是从目标出发,在网状结构的知识中将目标关联的所有事件聚合后返回给用户。因此,目标的知识检索是基于语句的检索,而关键词检索是基于文章的检索,目标的知识检索返回的结果更加精确。同时,在知识检索的结果上可以按时间、地点排序和统计,以实现对目标的简要分析,如目标动向、目标活动轨迹以及活动预测等。图 5 用某人物为示例展示了知识检索和关键词检索的结果及可扩展的分析能力。

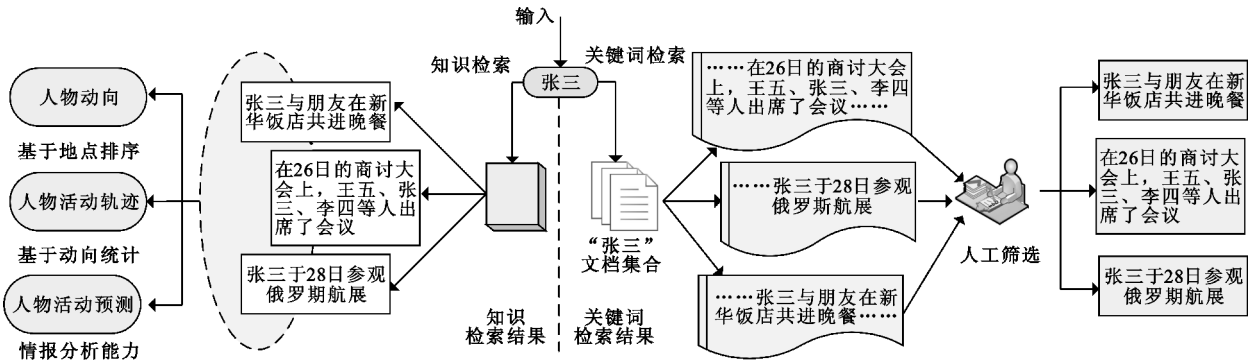

图 5　知识检索和关键词检索结果对比图
Fig. 5 Comparison between knowledge search result and keyword search result

目标的关联关系分析分为路径关联分析和潜在关联关系发现两种,前者主要是基于知识检索进行的路径关联查询,后者是基于知识推理规则进行的知识发现。下面我们将根据一个实际的示例来主要描述潜在关联关系发现得的分析方法和模型及结果。首先示例 ABox 用 triples 形式描述如图 6。

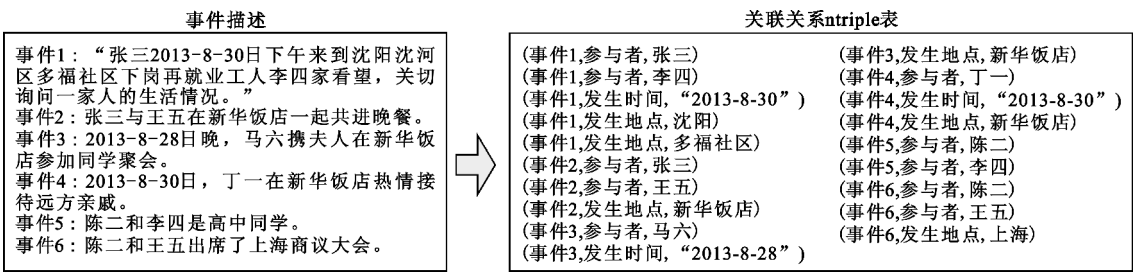

图 6　事件描述及抽取关联关系
Fig. 6 Event description and extraction relationship

目标对象的潜在关联关系发现模型及示例如下：

(1)关联规则 1 定义：如果两个目标 A 和 B 在同一时间、同一地点出现，则目标对象 A 和 B 具有潜在关联。

Prolog 规则模型如图 7 所示。

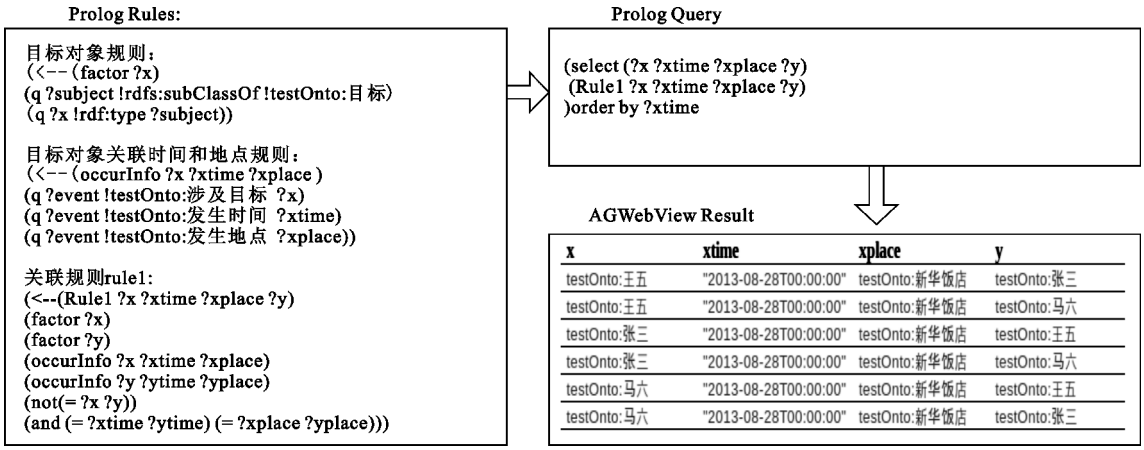

图 7 规则 1 描述图

Fig. 7 Description diagram of rule 1

(2)关联规则 2 定义：如果两个目标对象 A 和 B，分别检索并得到 A 和 B 的直接关联目标对象集合，直接目标对象中超过两个以上相同，则 A 和 B 具有潜在关联性。

Prolog 规则描述如图 8 所示。

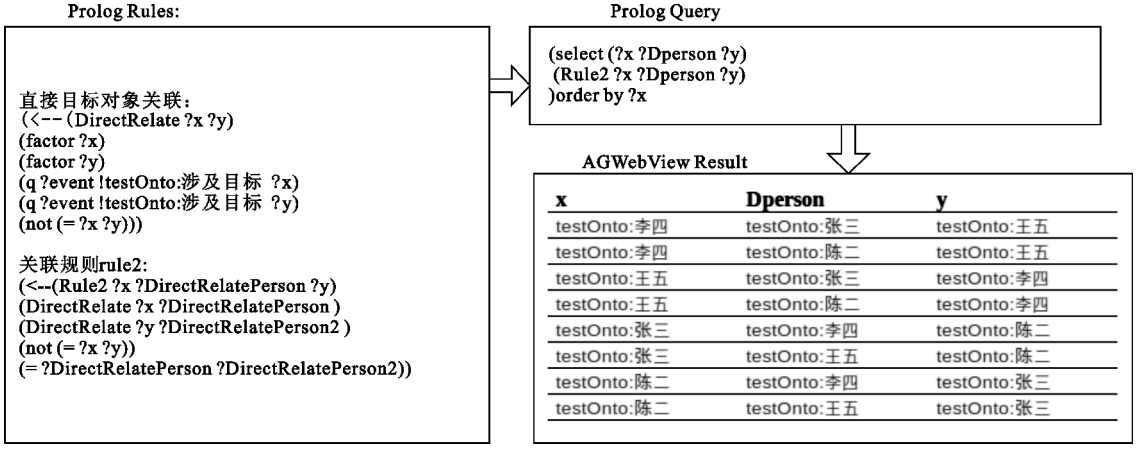

图 8 规则 2 描述图

Fig. 8 Description diagram of rule 2

3 系统主要流程

信息关联分析系统主要实现基于语义的知识检索，并在知识检索的结果上进行知识分析；在信息知识库的知识上通过基于语义的知识推理来完成目标对象的路径关联分析和目标对象的潜在关联关系发现。系统处理流程如图 9 所示。

首先将文本信息接入到系统，系统通过本体模型中的概念来确定需要在该文本信息中识别和提取哪些目标，以及判别这些目标实体的类型；通过目标实体识别结果、类型及位置来获取候选事件集；将候选事件集进行语法、句法分析来进行检测，选取符合条件的事件；在抽取的事件集中，结合本体模型的关系模型来提取目标实体之间的关联关系；将抽取的目标实体关联关系存储到实例知识库中；在实例知识库、本体知识库上进行知识检索；在实体知识库、本体知识库和规则库上进行知识推理；最后给出关

联分析的结果。

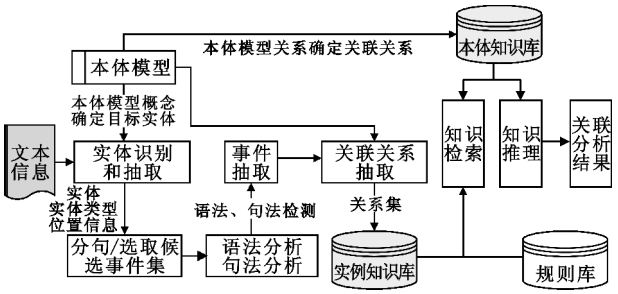

图 9　系统流程图

Fig. 9 Diagram of system workflow

文本关联关系分析技术其目的在于为文本信息处理人员提供快速的关联关系检索,并辅助其完成关联关系发现。结合工程系统应用,本技术对接入的文本信息中人物目标的相关信息进行提取,在抽取结果上引入语义技术进行人物目标的信息聚合,采用知识检索技术实现人物目标关联信息的快速检索,运用知识推理技术完成指定人物目标的潜在关联人物发现等功能,为信息分析人员进行人物跟踪监控、多人物间关系分析等提供辅助决策信息。

4　结　论

文本关联关系分析技术针对文本信息处理领域中文本信息的关联关系自动提取、快速检索、潜在关联关系发现等重大处理需求进行研究和设计,采用语义技术抽取并表示文本信息的关联关系,运用知识检索和推理技术实现信息聚合检索和潜在关联关系发现。基于语义进行文本信息的挖掘是一个新的研究方向,仍需要对每个处理环节进行持续研究,包括如何提取有价值的关联信息,如何更加合理、灵活地保留其语义信息和表示,语义信息的推理技术是否可以有效结合非语义数据从而演变新的技术来满足业务的处理需求等。

参考文献:

[1] Gao J. Resolution and accuracy of terrain representation by grid GEMs at a micro scale[J]. International Journal of Geographical Information Science,1997,11(2):199-212.

[2] 汤国安,杨勤科,张勇,等. 不同比例尺 DEM 提取地面坡度的精度研究——以在黄土丘陵沟壑区的试验为例[J]. 水土保持通报,2001,21(1):53-56.
TANG Guo-an,YANG Qin-ke,ZHANG Yong,et al. Research on Accuracy of Slope Derived From DEMs of Different Map Scales[J]. Bulletin of Soil and Water Conservation,2001,21(1):53-56. (in Chinese)

[3] 吴强,刘宗田,强宇. 基于本体的知识库推理研究[J]. 计算机应用研究,2005,21(1):55-57.
WU Qiang,LIU Zong-tian,QIANG Yu. Ontology based knowledge reasoning research [J]. Application Research of Computers,2005,21(1):55-57. (in Chinese)

[4] 曹锦丹. 基于文献知识单元的知识组织—文献知识库建设研究[J]. 情报科学,2002,20(11):1187-1189.
CAO Jin-dan. The knowledge organization based on the document knowledge unit[J]. Information Science,2002,20(11):1187-1189. (in Chinese)

[5] 卢宁. 面向知识发现的知识关联提示及其应用研究[D]. 南京:南京理工大学,2007.
LU Ning. Knowledge discovery oriented knowledge relationship reveal and application research[D]. Nanjing:Nanjing University of Science and Technology,2007. (in Chinese)

[6] 中国电子科技集团公司第十研究所. 联合情报[J]. 电讯技术,2012,52(suppl. 1):1-132.
The 10th Institute of CETC. Joint Information[J]. Telecommunication Engineering,2012,52(Suppl. 1):1-132. (in Chinese)

[7] 于龙,蹇强. 面向主题的信息抽取需求描述与分析[J]. 计算机工程,2012(23):57-59.
YU Long,QIAN Qiang. Theme oriented information extraction requirement description and anaylsys[J]. Computer Engineering,2012(23):57-59. (in Chinese)

[8] 高强,游宏梁. 事件抽取技术研究综述[J]. 情报理论与实践,2013(4):118-121,132.
GAO Qiang,YOU Hong-liang. Summery of event extraction technology research[J]. Information Studies:Theory & Application,2013(4):118-121,132. (in Chinese)

作者简介:

陈天莹(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

苏智慧(1965—),女,四川西昌人,1986 年于电子科技大学获计算机工程学士学位,现为研究员,主要研究方向为情报侦察及情报处理。

SU Zhi-hui was born in Xichang, Sichuan Province, in 1965. She received the B. S. degree from University of Electronic Science and Technology of China in 1986. She is now a senior engineer of professor. Her research concerns information reconnaissance and processing.