信息抽取研究综述

郭喜跃[1,2] 何婷婷[1]
(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)[1]
(兴义民族师范学院信息技术学院 兴义 562400)[2]

摘 要 信息抽取的任务是从大量数据中准确、快速地获取目标信息,提高信息的利用率。目前,信息抽取已经成为NLP领域的一个重要分支。随着互联网应用的发展,其价值也正日益显现,学术界和工业界对此都寄予厚望。首先回顾了信息抽取的发展历程;接着从命名实体识别、指代消解、关系抽取和事件抽取4个方面总结了信息抽取关键技术的研究进展;然后分析了信息抽取目前面临的若干主要问题;最后对信息抽取的研究趋势作了预测。

关键词 信息抽取,命名实体识别,指代消解,关系抽取,事件抽取
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.003

Survey about Research on Information Extraction

GUO Xi-yue[1,2] HE Ting-ting[1]
(National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China)[1]
(School of Information Technology, Xingyi Normal University for Nationalities, Xingyi 562400, China)[2]

Abstract The task of information extraction (IE) is obtaining the objective information precisely and quickly from a large scale of data. Nowadays IE has already been an important branch of NLP, and its value is also becoming increasingly apparent. As a result, the industry and academia are putting more and more emphasis on it. This paper first reviewed the development process of information extraction, then summarized the new research progress about IE from 4 aspect: named-entity recognition, anaphora resolution, relation extraction and event extraction. What's more, the paper analyzed some primary problems that the IE is facing with, and finally predicted the researching trend of IE in the future.

Keywords Information extraction, Named-entity recognition, Anaphora resolution, Relation extraction, Event extraction

1 信息抽取概述

随着互联网应用的迅猛发展,通过网络能够获取的数据量也呈指数级增长,如何从这些海量数据中快速、准确地分析出真正有用的信息,显得尤为关键和紧迫。而这正是信息抽取这一研究领域力图解决的问题。信息抽取(Information Extraction)的主要功能是从文本中抽取出特定的事实信息,这些文本可以是结构化、半结构化或非结构化的数据。通常,信息抽取利用机器学习、自然语言处理(NLP)等方法从上述文本中抽取出特定的信息后,保存到结构化的数据库当中,以便用户查询和使用[1]。整体上看,信息抽取的方法路线分为两大类:一种是基于KDD和数据挖掘的方法,主要从结构化、半结构化数据中抽取信息;另一种采用NLP和文本挖掘的方法,目标是从非结构化的开放文本中发现新知识,并将其转换为可理解的有用信息。

早期的信息抽取研究肇始于上世纪60年代中期,以美国纽约大学的Linguistic String和耶鲁大学的FRUMP这两个长期项目为代表。然而直到80年代末期,信息抽取的研究与应用才逐步进入繁荣期,这得益于消息理解系列会议(Message Understanding Conference, MUC)的召开。从1987年到1997年,MUC会议共举行了7届,MUC为信息抽取制定了具体的任务和严密的评测体系,该会议提出了一套完整的基于模板填充机制的信息抽取方案,核心内容包括命名实体识别、共指消解、关系抽取、事件抽取等具体内容。该会议吸引了世界各地的研究者参与其中,从理论和技术上促进了信息抽取的研究成果不断涌现。MUC为信息抽取在NLP领域中成为一个独立分支做出了重大贡献[2]。

继MUC之后,1999年至2008年美国国家标准技术研究所(NIST)组织的自动内容抽取(Automatic Content Extraction, ACE)评测会议成为另一个致力于信息抽取研究的重要国际会议。与MUC相比,ACE评测不针对某个具体的领域或场景,它采用基于漏报(标准答案中有而系统输出中没有)和误报(标准答案中没有而系统输出中有)的一套评价体系,还对系统跨文档处理(Cross-document Processing)能力进行评测。这一新的评测会议把信息抽取技术研究引向新的高度[3]。

到稿日期:2014-02-18 返修日期:2014-05-17 本文受国家语委"十二五"重点项目(ZDI125-1),国家"十二五"科技支撑计划课题(2012BAK24B01),教育部国家外国专家局高等学校学科创新引智计划项目(B07042),湖北省自然科学基金重点项目(2011CDA034)资助。
郭喜跃(1983-),男,博士生,实验师,主要研究方向为信息抽取、自然语言处理,E-mail:ihnlaoyao@126.com;何婷婷(1964-),女,博士,教授,博士生导师,主要研究方向为自然语言处理、数据库与数据挖掘。

除了 MUC 和 ACE 外，还有多语种实体评价任务会议(Multilingual Entity Task Evaluation，MET)、文本理解会议(Document Understanding Conference，DUC)等与信息抽取相关的国际学术会议，它们为信息抽取在不同领域、不同语言中的应用起到了很大的推动作用。

基于中文的信息抽取研究起步较晚，中文与西方字母型文字的巨大差异，导致中文信息抽取研究进展较慢，早期工作主要集中在中文命名实体识别方面，在 MUC-7、MET 等会议的支持下，取得了长足进步[4]；当前中文信息抽取研究在继续优化命名实体识别效果的基础上，已经向着共指消解、关系抽取、事件抽取等更高阶段发展。虽然当前信息抽取通常还只是面向特定领域开展，能够真正实现大规模应用的信息抽取系统仍然未出现，但是应当看到，近年来信息抽取领域呈现出更为活跃的态势，从理论到应用都有一些新进展。

2 信息抽取关键技术研究进展

信息抽取的具体实现方法可分为两类：基于规则的方法和基于统计的方法。早期的研究主要采用基于规则的方法，也曾促进了信息抽取的明显进步。但是基于规则的方法有其自身的局限性，如：人工编制规则的过程较复杂、通过机器学习得到的规则效率较低、系统通用性差等，所以后来的研究逐渐又转向基于统计的方法。基于统计的信息抽取，虽然可以从一定程度上弥补基于统计方法的缺陷，但是随着研究的深入，人们发现基于统计的方法并不是完美的，所以现在的研究又开始考虑采用将基于规则和基于统计的方法相结合的策略来寻找效果更佳的信息抽取方案。信息抽取的具体实现过程在一定程度上要依赖机器学习算法，近年来，机器学习算法在一些方面的突破，为信息抽取关键技术的进步提供了直接支持。

2.1 命名实体识别

命名实体识别(Named Entity Recognition，NER)是信息抽取的基础性工作，其任务是从文本中识别出诸如人名、组织名、日期、时间、地点、特定的数字形式等内容，并为之添加相应的标注信息，为信息抽取后续工作提供便利[5]。

早期研究人员通常创建基于语法的语言模型，利用人工编写规则进行命名实体识别，这种方法有较好的准确率，但是召回率不理想，而且规则的编写通常需要领域的专业人士耗时几个月甚至更长时间才能完成。鉴于基于规则的方法效率不高，人们考虑将统计模型引入到命名实体识别的实现上，利用机器学习的方法习得领域知识库，然后再对测试文本进行分析，这种方法取得较好的效果，一定程度上弥补了前述方法的不足。命名实体识别研究日趋成熟，当前的研究热点集中在应用阶段。

法国 EURECOM 的 Giuseppe Rizzo 等人在 2011 年介绍了他们开发的一款名为 NERD 的应用于 Web 的综合性命名实体识别系统。通过系统接口，NERD 能将 10 种较为流行的 Web 命名实体抽取器(主要有 DBpedia ontology、YAGO、AlchemyAPI、Content Extraction、YAHOO! 等)整合在一起，由能蕴含大量规则信息结构的本体来支撑运行，依据本体还可以在特定环境中调整各种工具的分类情况[6,7]。笔者认为，这种优势整合的思想能够在一定程度上提高系统的通用性，只是系统整合的复杂度较高。

近两年来，社交媒体应用在互联网上呈井喷状发展，由于其内容较短，句子成分复杂和随意，导致传统的 NLP 方法在分析微博内容时效果不佳，因此微博文本的分析成为当前的一个研究热点[8-10]。美国华盛顿大学(University of Washington)的 Alan Ritter 等人在对微博文本特点进行分析后，提出了一种基于文本分类和词性标注相结合的命名实体识别方法。首先对短博文进行分类，尽可能降低文本特征维度；在 Penn TreeBank 标注集的基础上，人工编制适用于微博文本的词性标注集，对文本进行词性标注；采用条件随机场(Conditional Random Fields，CRF)和交叉验证(Cross Validation)的方法对文本进行浅层语法分析，以识别出非递归短语；最后利用基于支持向量机(Support Vector Machine，SVM)的机器学习方法，对未标注的领域内和领域外数据进行训练，并结合 Gibbs 采样方法，获取命名实体的分布情况。与传统的文本处理方法相比，该方法能够减少 41%的识别错误率[11]，引起了较大关注。

在国内，命名实体识别的研究也已经进入到实用阶段。2006 年中科院计算所俞鸿魁、张华平等人提出了一种基于层叠 HMM 模型的中文命名实体识别方法，该方法首先在词语粗切分的结果集上采用底层 HMM 模型识别出普通无嵌套的人名、地名和机构名等，然后依次采取高层 HMM 模型识别出嵌套了人名、地名的复杂地名和机构名。该方法成为中文词法分析工具 ICTCLAS 的核心实现算法[12,13]，目前该工具已成为中文词法分析效果最好的软件之一。

苏州大学鞠久朋等人于 2011 年提出一种 CRF 与规则相结合的地理空间命名实体识别方法，该方法以丰富的知识(行政区划及其层级关系、地名通用词典、黄页中的公司名、特殊的句型句式等)作为触发条件，用 CRF 对满足条件的片段作地名及机构名识别，识别出来的命名实体又被解构(即解析出命名实体的内部结构，找出其中包含的通名、专名、饰名及扩展单元)，CRF 及知识用来进一步判断该命名实体是否表示事件发生地的地理空间信息，在实验中该方法的 F 测试值能达到 91.87%[14]。

2.2 指代消解

指代是一种常见的语言现象，通常分为回指和共指两种，回指是指当前的照应语与上文出现的词、短语或句子存在密切的语义关联性；共指则主要是指多个名词(包括代名词、名词短语)指向真实世界中的同一参照体。指代消解可以简化、统一实体的表述方式，对提高信息抽取结果的准确度有很大的促进作用。

早期的指代消解主要基于语言结构或领域知识，由人工建构消解规则来实现。较有影响力的研究成果主要有 Brennan、Strube 和 Tetreault 分别提出的基于中心理论和分类算法的方法、ZHOU 等提出的基于限制规则的多代理策略等[15]。

基于机器学习的共指消解研究采用机器学习中成熟的算法或模型对指代语和文本进行处理，从概率的角度消解指代。Carina 和 Anette 引入潜在角色链接(Implicit Role Linking)的方法，该方法可简单概括为：检测事件未填充的语义角色，确定上下文语境下这些角色可能被理解为哪些指向性信息，该文给出的解决思路是将实体模型嵌入到非监督的 CR(Coreference Resolution)框架中，消解效果有较为明显的提

升[16]；2009 年，布朗大学（Brown Universit）的 Eugene Charniak 与 Micha Elsner 提出的一种基于最大期望（Expectation Maximization，EM）算法的共指消解方法引起广泛关注，该方法将指代消解作为分类问题，利用 EM 模型在参数估计上的优势，从最简单的模型结构开始迭代学习参数，并借助越来越庞大的参数集使消解模型逐步复杂化，从而得出某一指代词在特定语境下最有可能的真实所指，基于此方法已经开发出相应的应用系统[17]；基于最大熵分类器，德国海德堡大学（University of Heidelberg）于 2010 年研发了一款名为 BART 的跨语言共指消解系统，该系统是一个高度模块化的工具包，依赖以最大熵为基础的论及对（Mention Pairs）分类器，针对不同的语言使用不同的论及对获取方法，并以论及类型（Mention Type）、语义级别、字符匹配度以及距离等作为选取特征，最终通过点对建模算法（Pairwise Modeling）、Ranker、语义树等算法实现消解，该系统能够从英语、德语、意大利语中进行共指消解[18]。

国内方面，清华大学王厚峰教授于 2002 年从技术上总结了中文共指消解的基本原理和方法，并对 Hobbs 方法、中心理论的 BFP 算法进行介绍[19]；2009 年，哈尔滨工业大学的刘挺、秦冰等提出一种集成多种背景语义知识（包括 WordNet、维基百科中蕴含的语义知识）的共指消解方法，该方法集成多种背景语义知识作为基于二元分类的共指消解框架的特征，分别从各语义知识集上提取背景知识，利用句子中的浅层语义关系、常见文本模式以及待消解词上下文文本特征，取得了明显效果[20]；苏州大学周国栋、孔芳提出一种基于树核（Tree Kernel）的中文共指消解框架，它亦将消解的过程作为分类问题来看待，根据指代消解的一般过程，它分为零指代检测、回指判定和先行词识别等 3 个子任务，针对每个子任务分别构建不同的解析树，最后通过训练可以得到一个完整的消解器，实验结果表明，其 F 测试值能够达到 70.67%，该框架成为国内首个能够系统处理指代消解任务的平台[21]。

从整体来看，面向中文的指代消解研究虽然取得不小进步，但是与国际研究水平相比，还有待进一步提高。另外，针对跨语言、跨文本的指代消解研究相对较薄弱。

2.3 关系抽取

关系抽取的作用是获取文本中实体之间存在的语法或语义上的联系，关系抽取是信息抽取中的关键环节。MUC 和 ACE 两大会议推进了研究热点，已从最初的单纯语言学模型的应用发展到使用浅解析器或完全解析器的 NLP 技术的应用和复杂机器学习方法的应用，而关系抽取的性能也随之有了大幅提升。早期关系抽取主要采用模式匹配的方法，随后又产生了于基于词典驱动的方法，当前主要采用基于机器学习和基于 Ontology 的方法。这里主要介绍后两种方法的研究进展。基于机器学习的方法本质上仍然是根据文本特征进行分类，较为成熟的算法有 MBL 算法和 SVM 算法。

斯坦福大学（Stanford University）的 Mike Mintz 等人在 2009 年提出一基于远距离监督学习的无标注文本关系抽取方法。该方法以 Freebase 为训练数据进行远距离监督学习，设计出面向文本特征（5 种词汇特征、2 种句法特征、命名实体标签特征以及这些特征的组合特征）的分类器，其算法融合了有监督和无监督的信息抽取方法，结果表明他们的方法既能从无标注文本中抽取出实体关系，也在一定程度上脱离了对领域知识的依赖[22]。

2011 年，德国洪堡大学（Humboldt Univercity）的 Philippe Thomas 等人，为研究生物医学文献进展，提出一种利用整体学习方法（Ensemble Learning）抽取药物之间的相互作用的方法。他们的方法基于不同语言特征空间，构建多种机器学习方法对比机制（主要为基于 Kernel Tree 的方法和基于案例的方法），然后利用整体学习法选出效果最好的方法。该方法利用 DDI Extraction 2011 提供的数据进行实验，实验结果的 F 值比 DDI Extraction 2011 最好的结果高出 5.1%[23]。

传统的监督学习假定每个实例明确地映射到一个标签，但这与实际并不总是一致的。为此，斯坦福大学的 Mihai Surdeanu 等人在 2012 年将多实例多标记学习（Multi-instance Multi-label Learning）引入到关系抽取中，形成一种新的方法，它利用带有潜在变量的图模型，并将文本中实体对和其标记融合在一起。这一方法一定程度上克服了远距离监督学习的缺陷，而且实验表明它在两类不同领域的文本中性能表现不俗[24]。

整体上，国内在关系抽取方面的研究方法正逐步接近国际前沿，这方面的研究也不少，但是近几年内鲜有突破性进展。主要将特定领域的、带标注的文本作为分析对象；在不断完善基于本体、各类图模型的关系抽取方法的基础上，越来越多的研究采用机器学习的方法来获取关系特征。

2.4 事件抽取

在信息抽取中，事件是指在某个特定的时间片段和地域范围内发生的、由一个或多个角色参与、由一个或多个动作组成的一件事情，一般是句子级的。事件抽取（Event Extraction）主要研究如何从含有事件信息的非结构化文本中抽取出用户感兴趣的事件信息，将用自然语言表达的事件以结构化的形式呈现出来。在事件抽取研究的发展过程中，ACE 会议给予的影响最为深远。事件抽取大体上可分为元事件抽取和主题事件抽取两个层次，其中元事件抽取是基于句子的基础级的事件抽取，是指一次动作的发生或状态的转变，其抽取目标包括时间、地点、人物、动作等；主题事件抽取是指围绕某一确定的主题，获取与其相关的一系列事件，通常由多类元事件组成。当前的研究主要还是集中在元事件抽取阶段，且已取得良好成果；另一方面，目前事件抽取研究使用的语料基本上还是以新闻、生物医学等个别领域的文本为主，面向开放文本的事件抽取研究还较少。未来事件抽取研究将在不断完善元事件抽取的基础上，继续向主题事件抽取方向迈进[25]。

事件抽取应该属于信息抽取领域中的深层次研究内容，它需要以前述几项研究作为基础，涉及 NLP、机器学习、模式匹配等多个学科的方法与技术，在信息抽取、情报学、NLP 等领域都有很好的应用前景。

David Ahn 在 2006 年提出一种基于分治思想的事件抽取方法。他将 ACE 会议中关于事件抽取的任务分解为一系列基于分类的子任务，包括：锚记标识、论元识别、属性赋值和事件共指，其中每一个子任务由一个机器学习分类器负责实施。基于句法分析和词法分析，此方法综合运用多种分类方法，主要包括 K 近邻分类算法、最大熵分类器、MegaM 算法和 Timbl 算法等，来提取事件属性及特征。最后他们通过实验表明，此方法的效果整体上都优于 ACE2005 的评测结果[26]。

2007年YAHOO!研究院的Tye Rattenbury等人提出了一种基于图像标签信息的事件抽取方法，引起了较高的关注。他们以Flickr网站上的图片及其元数据标签(通常包含时间、地点、经度和纬度等)为研究对象，以尺度结构识别(Scale-structure Identification)算法为基础，提出了抽取地点及事件语义信息的方法，首先利用改进的突发探测方法Naive Scan Methods和Spatial Scan Methods获取事件特征，然后利用尺度结构识别算法获取事件的地址信息。并且他们的方法还可以有效地应用于包含有特定地理标注信息的网页事件抽取，能够根据网页中出现的地理领域术语信息抽取出对应的事件，说明此方法有一定的普适性[27]。

在ACL2011年会上，来自芬兰图尔库大学(University of Turku)的Jari Björne介绍了他们研发的一款应用于生物医学领域的事件抽取系统，该系统能够有效描述生物分子之间相互作用的一些细节。此系统基于SVM模型工作，将词法、语句、词之间的依赖关系等作为选择特征，其事件抽取实现过程大致为：首先，在句子中识别出所有的实体；然后，预测实体之间的属性关联；最后，将实体/属性集分离成为独立的事件。该系统的评测结果也很显著，是所有参与BioNLP11任务的系统中唯一一个获得4项最佳性能的系统。可以看出，此系统也属于元事件抽取[28]。

3 信息抽取面临的问题

通过对现有研究的分析可以发现，当前信息抽取研究在语料的加工与选择、理论模型的改进与创新、应用范围的拓展等方面都取得了一定进展。特别是近年来，社会化网络、电子商务应用的迅猛发展，带动信息抽取的研究与应用取得了相应进步。但整体来看，信息抽取，特别是中文信息抽取的一些深层次研究与应用仍有较大提升空间，主要面临以下几方面问题。

1)中文篇章分析研究有待突破。篇章分析旨在研究自然语言文本的内在结构并理解文本单元是句子从句或段落间的语义关系，篇章分析技术在信息抽取的模板生成阶段将发挥重要作用。相对于英文篇章分析技术，中文篇章分析研究才刚刚起步。如果中文篇章分析技术能够取得突破性进展，则会有力促进信息抽取研究的进步[29]。

2)大数据时代带来的挑战。大数据意味着信息抽取对象的海量性，传统的面向特定领域、特定数量的文本信息抽取方法在大数据中应用时可能会出现各种不适应的问题。这是一个较为紧迫的课题，应引起研究者的注意。

3)事件抽取能力的局限。如前所述，事件抽取分为较低层次的元事件抽取和较高层次的主题事件抽取，而当前的研究成果主要还是集中在元事件抽取，这在一定程度上为从全局获取语义信息造成困难。

4)跨语言处理能力不足。随着人类交流活动日益广泛与深入，包含有多种语言的文本会越来越多地出现，这对信息抽取在处理跨语言文本方面提出更高要求。目前在这方面的研究还较少，研究成果也不太显著。

5)通用性较差。前文已经提及，当前信息抽取研究主要还是面向特定领域的文本进行，个别研究成果也仅能在相关的1-2个领域内进行抽取，这说明信息抽取系统的通用性还处于较低层次，影响了信息抽取应用的普及。

4 信息抽取的研究趋势

1. 知识表示结构的研究。当前信息抽取工作通常都需要一定量的领域知识库作为支撑，可辅助实现规则构建、机器学习等。本体(Ontology)是以往被较多采用的一种知识表示结构，但是随着处理数据量的急剧增加，本体的构建过程越来越困难，所以目前学术界开始考虑利用全新的知识表示形式——知识图谱来描述现实世界的知识存在。目前，这一研究在个别领域已经有成功的模式与经验出现。

2. 面向开放文本。信息抽取的语料来源起初为结构化的数据(如数据库中的数据)，后来发展到半结构化数据(如HTML网页、XML文件等)，这为从无结构的开放文本中进行信息抽取积累了丰富的经验，可以预计今后信息抽取研究会越来越多地将开放文本作为语料来源。这一研究也将间接促进信息抽取通用性的提升。

3. 理论模型的创新。理论模型是NLP研究中最核心的问题，学术界一直以来都非常重视理论模型的构建与完善，未来还将继续把成功的模型或方法(特别是机器学习领域新的研究成果，如Deep Learning等)借鉴到信息抽取的研究中来。

4. 应用领域的扩展。任何一种学术研究的价值最终都要体现到实际应用中。随着信息抽取理论研究的不断发展与成熟，其研究成果将越来越多地应用到不同的实际领域中，并在这一过程中进一步完善。

结束语 作为NLP领域的一个分支，信息抽取是比信息检索更深层次的文本挖掘研究，其研究价值也正得到越来越多的认可和重视。在重视其基础研究的同时，既要从纵向了解信息抽取的研究阶段和发展空间，也要进行横向比较，总结分析NLP其它领域甚至是其它与NLP相关的领域的研究阶段与研究成果，以创新理念引领信息抽取研究不断取得进步。

文章总结了近几年来信息抽取关键技术的研究进展，同时指出了国内外所处的研究阶段与存在的差异；分析了信息抽取当前所面临的主要问题与困难，并预测了信息抽取今后的研究趋势。

参考文献

[1] 李保利，陈玉忠，俞士汶. 信息抽取研究综述[J]. 计算机工程与应用，2003(10)：1-5

[2] Wikipedia：Message Understanding Conference [EB/OL]. 2013-12-27. http://en.wikipedia.org/wiki/Message_Understanding_Conference

[3] 张素香. 信息抽取中关键技术的研究[D]. 北京：北京邮电大学，2007

[4] Wikipedia：Named Entity Recognition [EB/OL]. 2013-12-28. http://en.wikipedia.org/wiki/Named_Entity_Recognition

[5] 张晓艳，王挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005(4)：44-48

[6] Rizzo G，Troncy R. NERD：Evaluating Named Entity Recognition Toolsinthe Web of Data[J]. Lecture Notesin Computer Science，2012(7295)：39-55

[7] Rizzo G，Troncy R. NERD：A Framework for Unifying Named Entity Recognition and Disam biguation Extraction Tools[C]// 13th Conference ofthe European Chapter of the Association for ComputationalL inguistics. 2012：73-76

[8] Li Chen-liang，Weng Jian-shu. TwiNER：Named Entity Recognition in Targeted Twitter Stream[C]//SIGIR. 2012：721-730

(下转第38页)

ding: Substrate Support for path splitting and migration[J]. ACM SIGCOMM on Computer Communication Review, 2008, 38(2): 17-29

[6] Szeto W, Iraqi Y, Boutaba R. A multi-Commodity flow based approach to virtual network resource allocation[C]// Proceedings of the IEEE Global Telecommunications Conference. San Francisco, USA, 2003: 3004-3008

[7] 姜明，王保进，吴春明. 网络虚拟化与虚拟网映射算法研究[J]. 电子学报，2011，39(6)：1315-1320

[8] Even G, Medina M, Schaffrath G, et al. Competitive and Deterministic Embeddings of Virtual Networks[J]. Theoretical Computer Science, 2013, 496: 184-194

[9] Hou Y, Zafer M, Lee K, et al. On the mapping between logical and physical topologies[C]// Proceedings of the 1st International Conference on Communication Systems and Networks (COMSNETS'09). Bangalore India, 2009: 483-492

[10] Zhu Y, Ammar M. Algorithms for assigning substrate network resources to virtual network components[C]// IEEE International Conference on Computer Communications (INFOCOM). Barcelona, Spain, 2006: 1-12

[11] Botero J F, Hesselbach X, Fischer A, et al. Optimal mapping of virtual networks with hidden hops[J]. Telecommjunications Systems, 2012, 51(4): 273-282

[12] Mosharaf Kablr Chowdhury N M, Muntasir Raihan R, Raouf B. ViNEYard: Virtual network embedding algorithms with coordinated node and link mapping[J]. IEEE/ACM Transactions on Networking, 2012, 20(1): 206-219

[13] Zhang S, Qian Z Z, Wu J, et al. An Opportunistic Resource Sharing and Topology-Aware Mapping Framework for Virtual Networks[C]// IEEE International Conference on Computer Communications (INFOCOM). Orlando, USA, 2012: 2408-2416

[14] 李小玲，郭长国，李小勇，等. 一种基于约束优化的虚拟网络映射方法[J]. 计算机研究与发展，2012，48(9)：1601-1610

[15] Jens L, Holger K. A virtual network mapping algorithm based on subgraph isomorphism detection[C]// Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures. Barcelona, Spain, 2009: 81-88

[16] Cheng X, Su S, Zhang Z B. Virtual Network Embedding Through Topology-Aware Node Ranking[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(2): 39-47

[17] Alkminm G P, Batista D M, Fonseca N L S. Optimal mapping of virtual networks[C]// Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM'11). Houston, USA, 2011: 1-6

[18] Hu Q, Wang Y, Cao X J. Resolve the virtual network embedding problem: A column generation approach[C]// Proceedings IEEE INFOCOM. Turin, Italy, 2013: 410-41488

[19] 刘新刚，怀进鹏，高庆一，等. 一种保持结点紧凑的虚拟网络映射方法[J]. 计算机学报，2012，35(12)：2492-2504

[20] Zhang Z B, Cheng X, Su S, et al. A unified enhanced particle swarm optimization-based virtual network embedding algorithm [J]. International Journal of Communication Systems, 2013, 26 (8): 1054-1073

[21] 黄彬彬，林荣恒，彭凯，等. 基于粒子群优化的负载均衡的虚拟网络映射[J]. 电子与信息学报，2013，35(7)：1753-1759

[22] Chang X L, Mi X M, Muppala J K. Performance evaluation of artificial intelligence algorithms for virtual network embedding [J]. Engineering Applications of Artificial Intelligence, 2013, 26 (10): 2540-2550

[23] Borodin A, Ei Yaniv R. Online Computation and Competitive Analysis[M]. New York: Cambridge University Press, 1998: 1-19

[24] Jon Michael Kleinberg. Approximation algorithms for disjoint paths problems[OL]. http://www.citeulike.org/user/djlin/article/271384

（上接第 17 页）

[9] Liu Xiao-hua, Zhang Shao-dian, et al. Recognizing Named Entitiesin Tweets[C]// ACL. 2011: 359-367

[10] Finin T, Murnane W. Annotating Named Entitiesin Twitter Datawith Crowdsourcing[C]// ACL. 2010

[11] Ritter A, Clark S, Etzioni M O. Named Entity Recognitionin Tweets: An Experimental Study[OL]. http://aclweb.org/anthology/D/D11/D11-D1141.pdf

[12] 俞鸿魁，张华平，等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报，2006(02)：87-93

[13] 李渝勤，孙丽华. 面向互联网舆情的热词分析技术[J]. 中文信息学报，2011(1)：48-54

[14] 鞠久朋，张伟伟，等. CRF 与规则相结合的地理空间命名实体识别[J]. 计算机工程，2011(7)：2010-214

[15] 孔芳，周国栋. 指代消解综述[J]. 计算机工程，2010(8)：33-36

[16] Silberer C, Frank A. Casting Implicit Role Linkingasan Anaphora Resolution Task[C]// ACL, 2012

[17] Charniak E, Elsner M. EM Works for Pronoun Anaphora Resolution[C]// ACL, 2009: 148-156

[18] Broscheit S, Poesio M. BART: A Multilingual Anaphora Resolution System[C]// ACL, 2010: 104-107

[19] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报，2002(06)：9-17

[20] 郎君，忻舟，刘挺，等. 集成多种背景语义知识的共指消解[J]. 中文信息学报，2009(3)：3-9

[21] Kong Fang, Zhou Guo-dong. A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution[C]// Proceedings of the 2010 Conference on Empirical in Natural Language Processing. 2010: 882-891

[22] Mintz M, Bills S, Snow R, et al. Distan tsupervision for relation extraction with out labeled data[C]// IJCNLP. 2009: 1003-1011

[23] Thomas P, Neves M, Solt I. Relation Extraction for Drug-Drug Interactions using Ensemble Learning[OL]. http://ceur-ws.org/vol-761./paper1.pdf

[24] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance-Multi-labelLearningforRelationExtraction[C]// Proceedings of the 2012 Joint Conference on Emprical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 455-465

[25] 赵妍妍，秦兵，车万翔，等. 中文事件抽取技术研究[J]. 中文信息学报，2008，22(1)：3-8

[26] Ahn D. Thestages of event extraction[C]// ACL. 2006: 1-8

[27] Rattenbury T, Good N, Naaman M. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags[J]. SIGIR, 2007

[28] Björne J, Salakoski T. Generalizing Biomedical Event Extraction [C]// ACL. 2011: 183 101

[29] 徐凡，朱巧明，周国栋. 篇章分析技术综述[J]. 中文信息学报，2013(3)：20-32