

本期话题：知识抽取

非结构化文本中内容对象抽取的技术方法综述*

□ 张智雄 吴振新 / 中国科学院国家科学图书馆 北京 100190

□ 赵琦 洪娜 / 中国科学院国家科学图书馆 北京 100190

中国科学院研究生院 北京 100080

□ 徐健 / 中国科学院国家科学图书馆 北京 100190

中国科学院研究生院 北京 10008

中山大学资讯管理系 广州 510275

□ 刘建华 / 中国科学院国家科学图书馆 北京 100190

中国科学院研究生院 北京 100080

摘要：近年来，知识抽取技术在非结构化文本的处理中起到很重要的作用。文章在对当前知识抽取的相关文献、系统和项目分析研究的基础之上，提出了当前知识抽取研究中的主要抽取内容对象的分类，并对这些主要内容对象抽取的相关技术方法进行综述。主要总结了Web对象识别和集成、术语识别和抽取、主题发现和识别、概念层次关系的抽取、非概念层次关系的抽取、事实抽取、观点抽取和倾向识别等7种内容对象抽取的技术方法。并在此基础之上，对未来知识抽取的发展趋势进行了分析。该文为2008年第9期本期话题“知识抽取”的文章之一。

关键词：知识抽取，对象识别，术语抽取，主题发现，关系抽取，事实抽取，观点抽取，数字图书馆
DOI: 10.3772/j.issn.1673-2286.2008.09.001

随着文本信息的不断增长，人们需要一种高效的文献挖掘技术，能够有效地发现和收集被掩埋的非结构化文本文献之中的知识。

近年来，知识抽取技术在非结构化文本的处理中起到很重要的作用，研究人员提出了诸如自适应的信息抽取（Adaptive IE）、开放信息抽取（Open IE）、自动本体学习（Ontology Learning）、基于模式的标注（Pattern-Based Annotation）、语义标注（Semantic Annotation）、基于Ontology的信息抽取（OBIE）等新的思路和方法^[1]，开发出了诸如MnM^[2]、Ontomat annotizer^[3]、TEXTRUNNER^[4]、Text2Onto^[5]、OntoBuilder^[6]、KIM^[7]、AKTiveMedia^[8]、ArtEquAKT^[9]等知识抽取系统。

目前，很多领域都通过知识抽取来实现对自由文本的挖掘，从文本中抽取特定的内容对象。

Philipp等人对非结构化文本中的内容对象抽取进行了分类^[118-119]，将从非结构化的文本中提取出知识的工作划分为词语提取、同义词提取、概念抽取、概念层级抽取、关系抽取、关系层级生成、公理模式抽取、通用公理抽取等一系列自下而上的学习子任务。

通过笔者对当前知识抽取相关文献、项目和系统的调研分析，发现现实的研究工作基本上遵循Philipp所提出的层次分类，但内容对象的抽取更加关注现实的需要，有一些很实用的内容对象抽取并不包括在Philipp所提出的知识抽取层次之中。从简单的命名实体识别、Web对象抽取，到术语抽取、主题发现、概念层次关系的抽取、非概念层次关系的抽取、事实抽取，再到评价抽取和倾向识别，都是当前非结构化文本中内容对象抽取的主要研究内容，而这些不同层次的内容对象的抽取，都利用了

* 本文受国家自然科学基金项目“从数字信息资源中实现知识抽取的理论和研究方法研究”（05BT0006）和国家“十一五”科技支撑计划课题“网络科技信息监测与评价”（2006BAH03B05）的资金资助。

各种不同的知识抽取技术和方法。

本文在对当前知识抽取的相关文献、系统和项目的分析研究的基础之上，提出了当前知识抽取研究中的主要抽取内容对象的分类，并对这些主要内容对象抽取的相关技术方法进行综述。

1 Web对象的识别和集成

Web对象的识别是命名实体识别的具体化应用。在Web页面中，嵌入了各种各样的对象，如人、产品、文章、组织机构等。将这些对象从Web页中抽取出来并进行集成，可以实现功能强大的对象级别的内容揭示。

Zaiqing Nie^[10-11]等认为，Web对象是一种有关某一Web信息的数据单元，可以被用来收集、索引和排序。Web对象通常可以被看成是一种与应用领域相关的概念。一个Web对象可以通过一系列的属性表示，如 $A = \{a_1, a_2, \dots, a_m\}$ 。对象的属性集根据领域的需要预先设置。为了构建一个基于对象级别的搜索引擎，需要应用到以下相关技术：①Web对象的抽取，从多个来源的数据中抽取对象及其属性；②Web对象的标识和集成，将每一个抽取出的对象实例映射到现实世界中并存储到数据仓储之中，需要对同一对象进行识别，共用一个对象标识；③Web对象的检索，需要对每个对象进行索引和排序。

为了实现Web对象的抽取，Zaiqing Nie等提出了分两步骤实现Web对象抽取的方法，如图1所示。

第一步的Web对象抽取，实现对象记录级别

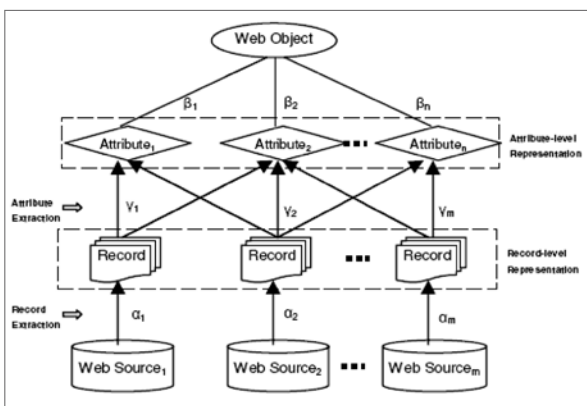


图1 Web对象抽取技术框架^[11]

(record-level)的抽取。Zaiqing Nie等将在Web页面上一系列有一定结构的相同条目（如产品的列表、服务的列表）称为数据记录。Web对象的抽取，先从数据源中抽取与领域相关的数据记录，形成对象记录级别的表示。每一个对象被表示成一系列被抽取出来的、与对象相关的数据记录，在此级别中不对对象的具体属性进行识别。

第二步的Web对象抽取，实现对象属性级别（attribute-level）的抽取。在这一过程中，需对上一步抽取出来的数据记录进行分析，将数据记录中的不同部分标识成为不同的属性，并且从多个来源的记录中，实现对同一对象的不同属性值的获取。

Zaiqing Nie等基于Web对象的抽取和集成的思路，应用在了Windows Live Product Search^[12]项目中。该项目从Web数据源中，自动地实现了大规模产品对象的抽取，在用户查询某一特定产品的时候，获得的并不是相关的Web页面，而是一系列与所查询内容相关的产品，每一个产品有明确的标题、图像、价格及特性等属性信息。对象级别的垂直搜索引擎系统Libra学术搜索^[13]也是一个基于Web对象抽取的系统，它可以帮助研究人员和学生查找科学论文、作者、会议和期刊。

Web对象的抽取对于Web资源的进一步发掘，如产品搜索、人物搜索、科学Web搜索、职位搜索、社团搜索等都有着重要的意义。近年来，数据记录的抽取^[14]、属性值的抽取^[15]以及Web对象的发现和标识^[16-17]等都引起了很多研究者的关注。

2 术语识别和抽取

科技文献之中的知识是围绕着领域内特定的概念来组织的。对科技知识领域进行描述的一个重要基础性工作就是从语言层面对科技领域中的概念进行表述。术语作为特殊主题领域内对某特定概念的“约定俗成”的名称，具有意义单一、低歧义、高专指性、相对固定的上下文环境等特点，表述了领域内的最为重要的一些概念，并且构成了文献的语义特征^[18]，对构建领域Ontology有着重要意义。发现专业领域内的术语，并建立术语之间的相关关系，是复杂的知识获取的基础。特别是近些年来，随着科学文献的大量出现，新术语层出不穷，抽取

和管理领域相关的术语变得越来越重要。

在英文中,术语的识别和抽取被称为Terminology Extraction、Term Extraction、Glossary Extraction、Automatic Term Recognition (ATR)。其目标在于自动地从特定的文件集中抽取出相关的术语。典型的自动术语抽取利用了语言分析器(如语法标记和短语分隔)来抽取候选术语,识别出文本中的名词短语(英文里的合成词,如“credit card”,形容词短语,如“local tourist information office”,以及介词短语,如“board of directors”)并通过统计和机器学习方法,对这些候选术语进行过滤。与术语抽取相关的主要问题有自动化的术语识别、术语变种处理、缩略语获取、自动相关术语发现以及术语聚类等等。

在术语的识别和抽取中, Kageura和Umino定义了判断某一语言单位是不是术语的两个重要指标^[19]: 单元性(Unithood)和术语度(Termhood)。单元性指的是组成某一语言单位的多个单词在句法结构或顺序上的稳定性和强壮性,也就是在词语搭配上的稳定性和强壮性。术语度指的是这一语言单位与领域专指概念之间的相关性,更直接而言,也就是这一语言单位表示领域专指概念的能力。对于一个术语候选者,如果单元性越强,术语度越高,则被识别为术语的可能性就越大。目前的术语识别和抽取方法有多种。从思路上来看,这些方法本质上都是或者基于术语单元性进行计算,或者基于术语度进行语义分析,或者对两个方面都进行考虑。

对于具体的术语识别和抽取方法,主要有以下几大类:

(1) 基于已有术语体系资源进行术语识别的,如利用LocusLink^[20]、FlyBase^[21]、UMLS^[22]等术语库进行术语查找和识别;

(2) 基于语言分析的方法,如对一般术语的构成模式进行分析, Ismail Fahmi^[23]提出术语过滤的规则如下:

$$((Adj|N)+|(((Adj|N)*(N Prep)?)(Adj|N)*))N$$

根据类似的术语规则可进一步将经过预处理的字符串与各规则相匹配判定候选术语;

(3) 基于统计和信息论方法。如Church and Hanks 的互信息(Mutual Information, MI)^[24], Smadja等提出的Dice Coefficient方法^[25], Dunning的

Log-likelihood 方法^[26], Manning and Schutze提出的T-test方法^[27];

(4) 基于机器学习。Jérôme Azé提出了名为Roger (ROc-based GEnetic learnéR)的遗传算法^[28];

(5) 多种混合方法。如C-Value和NC-methods方法^[29], Joachim Wermter提出的P-Mod方法^[30]。

基于上述方法,目前已经有一些系统,如University of Salford的ATRAC^[31]、IBM的GlossEX^[32-33]、University of Roma的TermExtractor^[34]、英国曼彻斯特国家文本挖掘中心NaCTeM的TerMine^[35]等。

3 主题发现和识别

主题发现的目标在于在已经确定了文档集中的若干术语或概念的基础上,进行主题聚类,发现文档中的核心主题和子主题,并通过一定的结构化方法(如Ontology)保存。同时,在聚类的基础上,用户往往需要一个直观的总结,以判断某个主题是否是他们感兴趣的。因此,为了使主题更明确和更易理解,一般的主题发现任务还包括主题描述构建以及主题可视化两部分工作。

3.1 主题聚类

主题聚类^[36-37]的目标是:不依赖于任何背景信息的支持,直接从数据中发现特定主题领域,它是主题发现的核心步骤。基本思路是:首先将文档中的术语表现为一个可计算的元素(如向量),进而判断术语之间的相似度,并通过选择适当的聚类算法进行聚类。

相似度计算是主题聚类形成的基本依据。一般来讲,术语聚类可按照术语的相似性将领域中的术语划分成多个类组,同一类中的两个术语之间相似性应当大于不同类之间的两个术语的相似性。Nenadic^[38]提出了语境、语法和词法相似性的计算方法。

(1) 词法相似性是衡量术语相似性最直接的方法,是指有一个共同的主要词和不同修饰词组成的术语之间的相似性。如果两个术语都有相同的主要词,则这两个术语就会有一个直接(或间接)的上位词,两个术语都是这个上位词的下位词(如progesterone receptor 和oestrogen receptor)。而且如

某个术语比另外一个术语多一个附加词（如nuclear receptor和orphan nuclear receptor），则指示这个术语更加专指。词法相似性对于比较多个词构成的术语较为有用，但对于特殊名称或者缩略词，则不能起到太大的作用。

(2) 语法相似性计算是在某些指示并行应用术语的词法-语法模式的基础之上展开的一种方法。并行模式主要有以下几种：枚举表示、并列、等格以及首语重复。其主要思想是所有处于并行结构中的术语都在句子中有相同的语法（如宾语或主语），并且与同一个动词和介词一起应用。可以认为这种并行应用术语的出现，表示这些术语之间有着功能的相似性。语法相似性的值，取决于术语共同出现的模式，以及在这种模式下术语共同出现的次数。

(3) 在语境相似性中，术语出现的语境模式（Context Pattern）被纳入到了比较的范围之中。语境模式包括了语法类型以及其它的语法和词法信息。在语境相似度计算之中，并不是对所有的语境信息都进行比较，相反是通过自动的模式发掘方法，来找出最显著的语境。语境会根据一个名为CP-value的值来进行排序，超过某个域值的语境将被认为是显著语境，一个术语因此将包括一系列的语境，术语之间的相似性，也就可以通过语境之间的相似性加以计算。

目前的针对文本主题分析的聚类技术基本上可以分为两类^[39]：划分聚类法（Partitional）和层次聚类法（Hierarchical）。

(1) 划分聚类法：给定一个包含n个术语或者关键词的文档集，要生成k个主题簇。一个划分聚类方法则要形成k个划分（k小于等于n），其中每个划分代表一个簇。通常会采用一个基本的划分准则，例如基于距离的相似度计算等，以便在同一个簇中的概念是相似的，而不同簇中的概念是相异的。划分聚类法常见的算法有K-means^[40]、K-Medoids^[41]、CLARA^[42]、CLARANS^[43]、CURE^[44]等。

(2) 层次聚类法：将概念组成一棵聚类的树。根据层次分解是自底向上的，还是自顶向下形成，层次聚类方法可以进一步分为凝聚和分裂层次聚类^[45]。凝聚聚类（Hierarchical Agglomerative Clustering，简称HAC）方法是首先将每一个概念

视为一个小的主题聚类，然后在逐渐收拢。而分裂聚类（Divisive Hierarchical Clustering，简称DHC）方法是将所有概念视为一个大的主题聚类，进而按照一定标准进行拆分。层次聚类法典型的算法有BIRCH^[46]、ROCK^[47]、Chameleon^[48]等。

划分聚类法的优势是可以利用文档集全局的信息（如熵值、术语分布情况等）来进行聚类划分。而另一方面，凝聚聚类算法可以生成小的可理解的聚类，这是划分聚类法所无法实现的。但是凝聚聚类方法在聚类的初期易出现决策错误，加重计算负担。因此，在实际的主题发现应用中，可将层次聚类和划分聚类两种方法结合。Aurora Pons-Porrata就提到了这个思想^[49]。利用一个分层求解（Multilayered Clustering）的方法来实现层次。在每一个层次上应用划分聚类方法，将前一个层次的主题聚类形成上层主题。混合法的常见算法有Scatter/Gather^[50]和限制凝聚算法（Constrained Agglomerative Algorithms）^[51]等。欧洲的SKET（Semantically Enabled Knowledge Technologies）^[52]项目中应用的Polysecting-Hierarchical-K-Means（简称PH K-Means）^[53]也利用了这种混合算法的思想。

需要特别指出的是，对于主题发现来说，目标文档集往往并不是固定不变的。尤其是对于主题探测与追踪（Topic Detection and Tracking，简称TDT）这样的任务来说。常常会有新的文档不断地加入到文档集当中。这种情况下，上述的一些聚类方法往往需要全局重新计算，计算成本很大。因此，许多专家在划分聚类和层次聚类的基础上研究了一系列的增量算法，针对新加入的文档进行增量计算，把新发现的概念或者术语融入到现有的主题结构中。典型的算法有Suffix Tree^[54]、DC-Tree^[55]、ICT^[56]、UMASS^[57]等。

3.2 主题描述构建

在实现了主题聚类之后，需要对主题聚类进行适当的描述，使用户可以迅速清晰地了解主题的内容，判断该主题是不是自己感兴趣的，决定要不要进一步深入了解。构建主题描述的主要方法是从主题内的文档集中将与主题密切相关的词、语句甚至片段抽取出来，并合理组织。形成主题描述的典型技术之一是多文档自动文摘（Multidocument

Summarization)^[58-59]。该技术通过相似度计算可以判断多文档集合中冗余信息的多少,在句子的抽取时根据句子的相似度抽取冗余性最小的句子组成文摘句集合,可以看到句子相似度的值将在多文档文摘各项技术中发挥作用。进而判断哪些句子可以纳入到概述当中,又如何组织成可理解的概述。模式识别领域的Testor Theory^[60]算法可以用来实现概述,与其他的多文档自动文摘不同,对于句子是否出现的判断是基于各个聚类中出现的不同数据的频繁度而计算出的,这样计算的效率较高。SEKT的OntoGen^[61]系统中也考虑到了生成主题背景信息对于用户发现主题的重要性。系统选择的方法是提供描述主题的关键词。为了得到这些关键词,采用了图心向量法(Centroid Vector)、支持向量机(Support Vector Machine, SVM)^[62]两种方法。

3.3 主题可视化

可视化是发现文档中主题的有效方法,利用可视化图形可以迅速和清晰地获得文档集主题领域的概览,帮助用户短时间内了解文档集全貌。同时通过一系列可视化的交互操作,能有效提高用户搜索的特定主题效率。由于对于文档主题多是以在高维空间内的向量来表示,因此降维是在这部分涉及的主要技术,可以将文档映射到平面上。代表性的算法有SOM(Self-Organizing Maps)^[63]、线性子空间法(Linear Subspace Method)^[64]和多维排列法(Multidimensional Scaling)^[65]等等。

主题发现目前多应用在主题探测与识别(TDT)、本体自动构建等领域。典型系统包括SEKT项目的OntoGen^[66]和组件TextGarden^[67],以及美国马萨诸塞州大学的TDTLighthouse^[68]和Aurora Pons-Porrata等人开发的JERARTOP系统^[69]等。

4 概念层次关系的抽取

在文本之中,有着各种各样通过术语表现的概念,而这些概念之间存在着多种的关系。概念层次关系(Concept Hierarchy、Taxonomy、Thesauri)反映概念的分类和归属,是对领域内知识的一种显现的表达方式,能够用于知识的推理,在信息检索、文本聚类 and 分类中起着重要的作用,对于知识系统

有着重要的意义。因此,从文中抽取出概念层次关系是当前知识抽取的一个研究重点。

按照Gruber^[70]的框架Ontology,概念之间的主要关系有:概念的继承(子类)关系(subclass-of);概念的实例关系(instance-of);概念的属性关系(relation);概念的内涵和外延(domain and range restrictions);概念的组成部分关系(mereological);概念的等同关系(equivalence)。

当然以上六种关系并不是所有概念之间的关系,也并不是所有六种关系都是概念层次关系抽取考虑的内容。从目前来看,概念层次关系的抽取,主要是对以下几种关系的抽取:子类关系的抽取(is-a),实例关系(instance-of),部分和整体关系(part-whole)以及等同(equivalence)关系的抽取。

在概念层次关系的抽取上,主要有三个思路和方法:

(1) 基于模式匹配的方法,通过发现文本中存在的层次关系表示语言,来查找概念之间的关系。这一方法代表性的例子有Hearst-pattern方法,因Hearst^[71]开创性的工作而得名。除此之外,很多概念层次抽取的工作利用首语重复法来抽取同位概念、利用合成词(修饰词+主要词)来抽取上下位类概念,利用枚举表示、并列、等格等来抽取同位概念,一些系统如PANKOW和C-PANKOW^[72]都主要基于这种方法。

(2) 基于分布式假设(Distributional Hypothesis)^[73]的方法。Harris提出的分布式假设认为,在同一上下文环境中出现的词趋向于有一个确定的相似意义,其本质意义就是一个词是通过它周围环境中的其它词来描绘其特点的,更进一步,一个词的意义是通过与它一同出现的词,以及这些词与该词同时出现的频次决定的。分布式假设构成了统计语义学的基础,同样也构成了语义关系抽取的一条重要理论方法。基于分布式假设的概念层次关系抽取方法又可分为两种,一种基于相似性计算,通过两词或术语之间的相似性关系计算来确定他们是否可以被聚为一类;另一种基于集合理论,它部分地利用了概念属性集之间的包含关系来对概念进行排序。

(3) 基于文献频次计算术语包含关系的方法。

Mark Sanderson^[74]认为,概念层次关系的构建是一个将术语按从综合性到专业性进行排序的一个过程,他参照Forsyth和Rada的研究成果^[75],认为一个术语的综合性或专业性可以通过它的文献频次(DF)来判定,在同一文献集中,如果出现某个术语的文章数越多,这一术语就可能会越综合。因此Mark Sanderson认为对于两个术语x和y,如果它们满足 $P(x|y) = 1, P(y|x) < 1$,则x包含y。也就是说,如果有术语y出现的文献集是有术语x出现的文献集的一个子集,则术语x包含y。

目前,也有一些工作,将以上三种方法都有机整合,形成一个统一的体系。如Philipp Cimianot^[76-77]等,将Hearst-pattern、WordNet^[78]词典匹配、主要词匹配启示法('head matching'-heuristic)^[79]、基于语料的概念属性集之间的包含关系计算、以及基于文献频次计算术语包含关系计算都综合在一起,从多个角度来构建概念层次关系,取得了较好的效果。

5 非概念层次关系的抽取

不同的两个实体之间有着多种多样的关系,以is-a为代表的概念层次关系仅仅是其中的一小部分。除此之外,两个实体之间存在着更多的是非概念层次关系,如科研人员和研究机构的“工作”关系,科研人员与论文之间的“撰写”关系等。对这些非概念层次之间的关系的识别和抽取,是当前知识抽取的重要内容之一。

在自动内容抽取(Automatic Content Extracting,简称ACE)^[80]测评会中,非概念层次之间的关系抽取是其中的一项重要内容,被称为关系探测与描述(Relation Detection and Recognition,简称RDR)。其目标在于发现物理空间关系、社会/个人的关系、雇佣关系、成员资格关系、制造商与代理(包括所有权)关系、从属关系等。对于每一个关系,测评系统都给出两个主要的变量(arguments)(即两个相互联系的实体)以及关系的属性。

Sophia Katrenko等人^[81]认为,关系抽取可以看作是具有两个步骤的过程:①识别存在关系的证据;②检查是否存在关系。而Gumwon Hong^[82]将自由文本关系抽取问题界定为对同一个句子中的一对实体之间关系的判断,可以表达为:

$$(e1, e2, s) \rightarrow r$$

其中,e1和e2是句子s中的两个实体,r是它们之间存在的关系的标签,三元组(e1, e2, s)称为一个关系候选。在r的取值集合确定(要抽取的关系类型确定)的情况下,至少存在三种关系抽取任务:①仅探测关系候选是否存在关系;②对关系候选是否存在关系进行探测,并执行N+1种关系类型的分类,N是关系类型的数量,另外不存在关系的情况也被看作分类时的一种类型;③对于每一个关系,仅执行N种分类。

在非概念层次关系的抽取上,主要有六种思路和方法:

(1) 基于模式匹配的方法。这种抽取方法通过运用语言学知识,在执行抽取任务之前,构造出若干基于语词、基于词性或基于语义的模式集合并存储起来。当进行关系抽取时,将经过预处理的语句片段与模式集中的模式进行匹配。一旦匹配成功,就可以认为该语句片段具有对应模式的关系属性。这类方法的典型例子是Douglas E. Appelt等人^[83]提出的基于“宏”概念的领域依赖规则通用方式表达方法。用户只需要修改相应“宏”中的参数设置,就可以快速配置好特定领域任务的关系模式规则。此外,Roman Yangarber等人^[84]提出了基于样本泛化的关系抽取模式构建方法。用户通过模式构建界面,对含有某种关系的例句进行分析,识别出所含关系的要素,并将这些要素泛化,最后经用户确认存储经泛化表达的模式。

(2) 基于词典驱动的方法。与基于模式匹配的关系抽取方法相比,基于词典驱动的关系抽取方法显得非常灵活。新的关系类型能够仅仅通过向词典添加对应的动词入口而被抽取。该类方法的一个典型例子是Chinatsu Aone等人^[85]在REES(Large-Scale Relation and Event Extraction System)系统中应用的词典驱动方法。REES的词典驱动方法需要对于每一个事件指示词设置一个词典入口,而这个词通常是动词。词典入口具体化了该动词参数的句法和语义限制。

(3) 基于机器学习的方法。该方法实质是将关系抽取看作是一个分类问题。通过具体的学习算法,在人工标引语料的基础上构造分类器,然后将其应用在领域语料关系的类别判断过程中。该类方

法近年来得到了广泛关注,并出现了各种具体实现方法。典型的例子有:Zelenko^[86]等人提出了利用核函数来从自然语言语料中抽取关系的方法。Intel中国研究中心的ZHANG Yi-Min和ZHOU Joe F等人^[87]提出了利用MBL算法获取规则用以抽取命名实体及它们之间关系的方法。Zhu Zhang^[88]提出了基于SVM的弱监督关系抽取方法。Michele Banko等人^[89]提出了Open IE方法,通过自动学习和统计算法,实现了对网络中海量异构信息中可能存在的关系的抽取。

(4) 基于Ontology的方法。该方法借助已有的本体层次结构和其所描述的概念之间的关系来协助进行关系的抽取。Marta Sabou和Mathieu d' Aquin等人^[90]在SCARLET系统中应用了通过自动选择和查询本体来发现概念实体之间关系的方法。当要确定两个概念实体之间的关系时,SCARLET先识别网络上能够提供上述概念实体相关信息的本体,然后综合这些信息并进行推理来获取概念实体之间的关系。

(5) 混合抽取方法。随着关系抽取研究的不断深入,研究者逐渐意识到,单纯的抽取方法在识别特征和识别模式方面难以避免地会具有局限性。为了将更多的已有关系识别特征有机地组织到关系抽取过程中来,一些将多种现有关系抽取方法相结合的混合抽取方法被提出来,其中具有代表性的是Lucia Specia和Enrico Motta^[91]提出的一个抽取语义关系的混合方法。该方法通过管道(pipeline)方式引入了解析器(parser),词性标注器(part-of-speech tagger),命名实体识别系统,基于模式的分类器以及词义辨析模块,并用到了领域本体、知识库以及词语数据库等资源。

关系抽取技术在很多领域具有应用价值。在自动问答系统中,关系抽取自动关联相关问题和答案;在检索系统中,关系抽取使类似于“北京有哪些公司?”这样的语义检索功能的实现成为可能;在本体学习过程中,关系抽取能够发现新的实体间关系来丰富本体结构;在语义网标注任务中,关系抽取能够自动关联语义网知识单元。

6 事实抽取

在自由文本之中,隐藏着很多有用的事实。在知识工程中,事实是一种知识类型,是某些被认为

可信的东西,或是真实存在的东西,或是可根据某些已经建立起来的判断标准进行检验的东西^[92]。事实可以直接回答用户提出的很多日常问题(如“美国2007年的GDP是多少”),对于提高信息检索的针对性和准确性有着重要的作用;另一方面,事实库的建设,对于提高问题解答系统的功能起着非常重要的作用。

事实抽取(Fact Extract)很早就成了文本理解领域内一个关注的重点。上世纪90年代初Carnegie Group就已经为路透社开发名为JASPER(Journalist's Assistant for Preparing Earnings Reports)的事实抽取系统^[93]。这一系统利用了一个模板驱动的方法,以路透社所采集的信息作为输入,通过部分理解技术以及相关启发规则,能够从文本中抽取出特定的关键事实,并将这些事实重新组合,形成新的新闻故事,以供栏目编辑继续修改和完善。

目前从文本中实现事实的抽取仍然是知识抽取关注的一个重点。例如2004年荷兰科学研究组织(NWO)资助的交互式多模态信息抽取(IMIX)项目中,就有一项工作需要实现一个名为FactMine的事实抽取器^[94-95]。它利用了手工方法和机器学习规则进行信息抽取,利用了聚类技术以发现文本中短语之间有用的关系,从文本中抽取出事实,构建事实库和本体信息。

Google公司所做的工作是目前大规模事实知识抽取的代表^[116-117]。Google认为,事实由两个命名实体以及它们之间的二元关系构成,例如“澳大利亚的首都是堪培拉”,“莫扎特出生于1756年”。他们认为由于噪音文本处理难度较大,作为分散人类知识仓储的Web仍然没有得到充分的开发和利用。构建一定规模(如上亿个事实数据)的事实库,能够有效促进信息的检索,提供另一种模式实现检索结果的展示。为了达到这一目标,他们进行了从Web文件中实现大规模事实知识抽取的实验,并构建了百万级的事实知识库。

Google公司提出的从Web中实现大规模事实抽取的基本思路和Sergey Brin提出的反复迭代的模式关系扩展(DIPRE - Dual Iterative Pattern Relation Expansion)方法十分相似^[96]。它先从一系列小规模的事实种子集出发,通过匹配和分析等步骤发现可以用于从文档集中抽取事实的上下文模式,然后

通过这些模式搜索网页, 根据这些模式从这些网页中抽取更大规模的候选事实种子, 对候选事实种子排序, 并且将效果最好的事实种子候选者加入到种子集中, 再进一步对这些种子集进行模式分析, 形成更多的模式, 再进而根据这些模式抽取候选事实, 如此循环, 不断扩大模式和事实种子集, 直到满足需要。

Google公司在实验中强调的是大规模的抽取, 希望能够处理上亿篇的文档, 实现一百万事实的抽取, 并且达到80%的正确率。并且由于手工编写的抽取规则具有领域限定性, 而大规模的种子事实难以编辑, 因此, 实验要求, 事实抽取从10个种子事实开始。抽取的集合和最初的种子事实集合在数量上的增长比值达到100,000:1。就目前所知, Google公司事实抽取集合的增长率和数量等指标比在它之前所有事实抽取研究的相应指标都要高。

在实际的实验中, Google公司以10条Person-BornIn-Year的种子事实开始, 成功地从3亿多个Web页面中抽取出了1,000,000条Person-BornIn-Year的事实, 并且准确率达到了90%的水平, 为构建大型事实仓储、辅助Web检索和提供问题解答提供了有益的尝试。

7 观点抽取和倾向识别

有人认为, 各种文本资料中都存在着两类混合在一起的信息: 事实和观点^[97]。观点 (Opinion) 不同于事实, 它是一个人对其事物的意见和想法, 是对某件事物的估计、判断和评论, 与事实相比, 观点可能具有不符合真实情况、尚未被证明的特点^[98]。

近些年来, 随着Web应用的不断发展, 人们表述观点的方式方法已经急剧地改变。人们可以将他们的观点发布在Internet论坛、讨论组、博客等被称为用户产生内容 (或用户产生媒体) 的Web之上^[99]。因而我们可以看到, 当前的网络上充斥着个人的观点信息, 如对某个产品的评论, 对某项国策的意见, 对某些服务的抱怨等。如何从这些Web文本之中, 识别和抽取人们对某个产品、对某项国策、对某些服务的观点, 是当前知识抽取领域一个重要的研究课题, 同时也是当前知识抽取的一个研

究热点。

除了Opinion Extraction之外, 对于观点抽取, 还有多种不同的英文称谓, 如Opinion Mining, Opinion Detection, Sentiment Analysis, Sentiment Classification等。当前的观点抽取主要集中在以下3个领域: 情感分类、事物属性评价和比较评价挖掘。

情感分类 (Sentiment Classification) 是目前观点抽取中研究最多的领域。情感分类又基本上可包括3个方面内容: 主观性 (Subjectivity) 判断, 如何区分文章中的事实和观点, 对文章、句子以及语词的主观性和客观性进行判断, 区分哪些是主观的, 哪些是客观的; 倾向性 (Orientation, Polarity) 判断, 也称语义倾向性判断, 主要的任务是判定文本中的某些内容 (整篇文章、句子、词语) 是肯定的、否定的、还是中立的; 等级强度 (Gradability) 判断是对主观性和倾向性的强度进行计算, 判定肯定或否定的等级强度。

词语的语义倾向性判断是情感分类的基础。与事实抽取关注名词短语的识别和抽取不一样, 词语的语义倾向性通常借助于形容词和副词实现, 然而词语的语义倾向性计算同样是基于Harris提出的分布式假设, 一个词语如果和一些人工标注的属于肯定倾向的“种子”词共同出现的频率越高, 那么这一词属于肯定的语义倾向性则越高。

Peter D. Turney提出的PMI-IR算法是词语语义倾向性计算的一个主要方法^[100-101], 它利用了点式互信息 (Pointwise Mutual Information, 简称PMI) 和信息检索 (Information Retrieval, 简称IR) 来测定两个词语的相似性, 并利用了某一词语和一个参考肯定词 (此处为“excellent”) 的相似性, 以及此词语与一个参考否定词的相似性 (此处为“poor”), 来测定这一词语的语义倾向性 (Semantic Orientation, 简称SO), 如下公式所示:

$$SO(\text{phrase}) = \text{PMI}(\text{phrase}, \text{“excellent”}) - \text{PMI}(\text{phrase}, \text{“poor”})$$

另一个词语语义倾向性计算的方法是Hong Yu等提出的修正对数似然比 (log-likelihood ratio) 方法^[102], 它通过某一词 W_i 和肯定种子集合 ADJ_p 中的词一同在句子中出现的频次, 以及这一词和否定种子集合 ADJ_n 中的词一同在句子中出现的频次的对数比来计算这一词的语义倾向性。计算方法如下:

$$L(W_i, POS_j) = \log \left(\frac{\text{Freq}(W_i, POS_j, ADJ_p) + \epsilon}{\text{Freq}(W_{all}, POS_j, ADJ_p)} \right)$$

其中 $L(W_i, POS_j)$ 表示词 W_i 作为某个词性(如动词、形容词、副词等)时的语义倾向性,而 $\text{Freq}(W_{all}, POS_j, ADJ_p)$ 表示所有词性为 POS_j 的词,与 ADJ_p 出现的频次,而 ϵ 是一个常量。

事物属性评价是在词语语义倾向性识别的基础之上,在句子层面识别和发现人们对某个事物的某些方面的喜好。例如对某型号的数码相机镜头、外观、性能方面的评价和意见。这方面的研究有Nozomi Kobayashi等从Web文献中实现<evaluated subject, focused attribute, value>三元组抽取的尝试^[103]。也有一些专家针对产品评价进行信息抽取的系统出现,如Ana-Maria开发的OPINE系统,能够对产品的特性,如属性、部件、部件属性等进行抽取,同时还可抽取出人们对这一产品不同属性的评价意见^[104]。

比较评价挖掘针对文献中记录是有关不同对象之间的比较关系进行挖掘,例如从文本中识别出某款数码相机的外观比另一款数码相机的外观更加漂亮。比较评价挖掘主要有两个核心步骤:①识别出文本中的比较语句;②对比较语句中的比较关系进行识别。从目前的比较评价挖掘来看,对比较词汇的识别等自然语言分析在比较关注的识别中起到了很大的作用^[105]。而N. Jindal等认为,利用83个比较关系词,就可以识别出98%的英文比较句子^[106],这些比较关键词包括形容词和副词的比较级和最高级,一些类似于same、similar、differ等表示相同或不同的词汇,以及一些类似于favor、beat、win、exceed等的,应用于主语和宾语为同一类型对象的动词。

8 未来趋势分析

知识抽取的内容对象和技术方法还在不断地完善和丰富中。Web对象抽取、术语识别、主题发现、概念和非概念层次关系抽取、事实抽取、观点抽取等不同思路的提出对知识抽取技术的发展做出

了有益尝试,机器学习和自然语言分析两大技术思路的相互融合已经成为知识抽取技术发展的主流趋势。在此基础上,今后的研究将会向更加深入和实用化的方向发展,其主要发展趋势有:

(1) 知识抽取内容对象的多样化。当前,除了以上主要的内容对象,不同的知识抽取任务还针对多种不同的内容对象进行抽取,如故事抽取(Story Capture)、条件抽取(Constrain Extraction)等。不同的知识抽取内容对象采用的技术方法也各不相同,如Andrew S. Gordon等人提出了一种基于统计学原理的算法实现文本分类,从而可以从网络博客中进行故事抽取的方法^[107]; Will Bridewell等人提出了一种基于机器学习原理的迭代方法,在流程建模过程中进行条件抽取^[108]。这些新的内容对象和抽取方法都为知识抽取理论和技术研究开创了新的思路。

(2) 知识抽取多种方法的集成。多种技术方法的集成使用将成为今后知识抽取研究的趋势之一,如英国Open University的KMI(Knowledge Media Institute)研究所开发的Espotter系统^[109],采用了基于词典的抽取方法、基于规则的抽取方法与Google域搜索相结合的方式知识抽取,该系统的最大特点是可以通过调节词典和规则来适应不同域下网页的内容对象抽取; Washington大学Turing center实验室的Kylin系统^[110]采用基于SVM的机器学习方法、启发式文档分类器、Infobox、自动链接生成等方法从Wikipedia上抽取出结构化数据。这些方法的集成能够帮助知识抽取任务更加高效、准确地完成特定内容对象的抽取目标。

(3) 跨文档、多来源、多媒体、多模态信息内容的抽取和集成成为重要研究领域。现实世界语义内容的揭示是复杂的、多途径的,有时仅从文本不能全面反映知识内容,需要从多媒体、多模态中来体现。因此,跨文档、多来源、多媒体、多模态信息内容的抽取和集成也将成为知识抽取研究的另一个主要方向。ACE测评会议提出跨文档、多来源的内容对象抽取已经成为当今知识抽取领域的研究重点。跨文档的实体识别任务需要在不同文档中建立指示相同对象的实体之间的关联,通过建立跨文档的唯一标识符来实现跨文档实体的一致性。跨文档的关系共指消解通过建立跨文档的、覆盖整个语料的唯一关系标识符实现。K-space^[110]主要通过构建一

个多媒体Ontology进行多媒体内容的分析和标注,采用基于多媒体处理技术(如颜色聚类)、基于规则的算法,从多媒体环境中进行语义内容对象的抽取。X-media^[111]提出了一种基于机器学习方法进行多媒体内容特征提取的知识抽取框架,它的目标是抽取多种媒体(文本、图片和数据库数据)中的显性和隐形知识。

(4) 嵌入式知识抽取将得到广泛关注。随着系统互操作和软件集成需求的不断提升,开发能够灵活嵌入到其他硬件或软件中的知识抽取系统成为未来发展的必然要求。嵌入式知识抽取系统是指系统可以以插件的形式方便地集成到现有的硬件系统或软件系统中,与原系统的功能组件无缝集成,以达到在原系统功能基础上实现知识抽取任务的目的。目前具有嵌入式功能的系统已有一些案例,如Magpie^[112]和Watson^[113]。

(5) 可移植性是知识抽取技术应用的一个重要要求。为了充分实现软件复用和系统的松散耦合,可移植的知识抽取系统将是今后的发展趋势,也是研究的难点。随着Ontology在知识抽取系统中的应

用,知识抽取技术向可移植性方面发展。可移植性主要是指系统以Ontology为起点和核心,各个组件与Ontology之间是松散耦合的,在没有人工干预和机器学习的情况下,系统能够随着Ontology的改变而自动适用于不同领域。这种知识抽取系统的实现难度较大,目前只能适用于部分对语义理解要求不高的场合,如OntoSyphon^[114]和ontoX^[115]。

(6) 知识抽取技术和其它知识技术相结合,将出现更加丰富多样的应用。将多种内容对象抽取之后的结果做深层次的、细致和深入的分析,继而投入到应用领域中去,这是知识抽取任务的最终价值体现。除了用于通常的语义知识领域,如语义Web、知识工程、趋势分析、主题发现、舆情监测、自动问答等,还将有多种更深层次的应用,如用于知识领域描绘(Knowledge Domain Mapping)、突发事件的探测(Burst Detection)、新兴发展趋势的探测(Emerging Trends Detection)等。我们相信,随着知识抽取技术方法的不断完善,知识抽取必将更加深远地影响到诸多与语义应用密切相关的领域。

参考文献

[1] 张智雄. 当前知识抽取的主要技术方法解析[J]. 现代图书情报技术, 2008(8):2-11.
 [2] VARGAS-VERA M, et al. MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup[C]//The 13th International Conference on Knowledge Engineering and Management (EKAW 2002). Berlin: Springer, 2002:379-391.
 [3] Ontomat[EB/OL]. [2008-07-05]. <http://annotation.semanticweb.org/ontomat/index.html>.
 [4] TEXTRUNNER[EB/OL]. [2008-07-05]. <http://www.cs.washington.edu/research/textrunner/>.
 [5] Text2Onto[EB/OL]. [2008-07-05]. <http://ontoware.org/projects/text2onto/>.
 [6] OntoBuilder. [EB/OL]. [2008-07-05]. <http://iew3.technion.ac.il:8080/OntoBuilder/>.
 [7] POPOV B, et al. KIM Semantic Annotation Platform[J]. Natural Language Engineering, 2004, 10(3/4):375-92.
 [8] AKTive Media[EB/OL]. [2008-07-05]. <http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html>.
 [9] ArtEquAKT[EB/OL]. [2008-07-05]. <http://www.artequakt.ecs.soton.ac.uk/>.
 [10] NIE Zaiqing. Web object retrieval[C]// International World Wide Web Conference Committee. Proceedings of the 16th international conference on World Wide Web, 2007:81-90.
 [11] NIE Zaiqing. Object-level ranking: bringing order to Web objects[C]// Proceedings of the 14th international conference on World Wide Web, 2005:567-574.
 [12] Windows Live Product Search[EB/OL]. [2008-07-05]. <http://products.live.com>.

[13] Libra学术搜索[EB/OL]. [2008-07-05]. <http://libra.msra.cn>.
 [14] LIU Bing, GROSSMAN R, ZHAI Y. Mining Data Records in Web Pages[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2003.
 [15] ZHU Jun, NIE Z, WEN JR, ZHANG B, MA WY. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2006.
 [16] YAO Conglei, YU YangJian, SHOU S, LI XM. Towards a Global Schema for Web Entities[C]//WWW 2008 / Alternate Track: WWW in China - Chinese Web Innovations April, 2008:21-25.
 [17] TEJADA S, KNOBLOCK C A, MINTON S. Learning domain-independent string transformation weights for high accuracy object identification[C]// Knowledge Discovery and Data Mining (KDD), 2002.
 [18] Terminology_extraction[EB/OL]. [2008-07-05]. http://en.wikipedia.org/wiki/Terminology_extraction.
 [19] KAGEURA K, UMINO B. Methods of automatic term recognition: A review[J]. Terminology, 1996, 3(2):259-289.
 [20] MAGLOTT D R, et al. NCBI's LocusLink and RefSeq[J]. Nucleic Acids Research, 2000, 28(1):126.
 [21] DRYSDALE R A, CROSBY M A. FlyBase: genes and gene models[J]. Nucleic Acids Research, 2005, 33(Database Issue): D390.
 [22] HUMPHREYS B L, LINDBERG D A B. The UMLS project: making the conceptual connection between users and the information they need[J]. Bulletin of the Medical Library Association, 1993, 81(2):170.
 [23] FAHMI I. C-value method for multi-word term extraction[EB/OL]. [2008-07-05]. <http://odur.let.rug.nl/~fahmi/talks/statistics-c-value.pdf>.
 [24] CHURCH K W, HANKS P. Word association norms, mutual information,

- and lexicography[J]. Computational Linguistics, 1990, 16(1):22-29.
- [25] SMADJA F, MCKEOWN R, HATZIVASSILOGLU V. Translating collocations for bilingual lexicons: a statistical approach[J]. Computational Linguistics, 1996, 22(1):1-38.
- [26] DUNNING T. Accurate methods for the statistics of surprise and coincidence[J]. Computational Linguistics, 1993, 19(1):61-74.
- [27] MANNING C, SCHUTZE H. Collocations[M]. Cambridge, MA: MIT Press, 1999:165-184.
- [28] Az é J. Preference Learning in Terminology Extraction: A ROC-based approach [C]// ASMDA'05 (Applied Stochastic Models and Data Analysis), 17-20 may, Brest, France. Arxiv preprint cs.LG/0512050, 2005.
- [29] NENADIC G, SPASIC I, ANANIADOU S. Terminology-driven mining of biomedical literature[J]. Journal of Biomedical Informatics, 2003(33):1-6.
- [30] WERMTER J, HAHN U. Finding new terminology in very large corpora[C]//Proceedings of the 3rd international conference on Knowledge capture, 2005. New York: ACM, 2005:137-144.
- [31] ATRACT[EB/OL]. [2008-07-05]. <http://cat.inist.fr/?aMode=afficheN&cpsidt=1020164>.
- [32] PARK Y, BYRD R J, BOGURAEV B K. Automatic glossary extraction: beyond terminology identification[C]//International Conference on Computational Linguistics, Proceedings of the 19th international conference on Computational Linguistics Taipei, Taiwan, 2002. New Jersey: Association for Computational Linguistics, 2002:1-7.
- [33] NAVIGLI R, VELARDI P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites[J]. Computational Linguistics, 2004, 30(2):151-179.
- [34] SCLANO F, VELARDI P. TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities[C]//Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA, 2007:28-30.
- [35] TerMine[EB/OL]. [2008-07-05]. <http://www.nactem.ac.uk/software/ctermine/>.
- [36] AYAD H, KAMEL M. Topic discovery from text using aggregation of different clustering methods[C]//AI' 2002: The Fifteenth Canadian Conference on Artificial Intelligence, 2002:27-29.
- [37] HAN J, KAMBER M. Data Mining: Concepts and Techniques[M]. 2 ed. [S.l.]: Morgan Kaufmann, 2006.
- [38] 同29.
- [39] 同37.
- [40] HARTIGAN J A, WONG M A. A K-means clustering algorithm[J]. Applied Statistics, 1979, (28):100-108.
- [41] KRISHNAPURAM R J, YI A L. A fuzzy relative of the k-medoids algorithm with application to webdocument and snippet clustering[C]//Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE'99. 1999 IEEE International, 1999. 3.
- [42] LEE B, ASSOCIATES G, SANTA C A. A new algorithm to compute the discrete cosine transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(6): 1243-1245.
- [43] NG R T, HAN J. CLARANS: A Method for Clustering Objects for Spatial Data Mining[J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2002, (14): 1003-1016.
- [44] GUHA S, RASTOGI R, SHIM K. Cure: an efficient clustering algorithm for large databases[J]. Information Systems, 2001, 26(1): 35-58.
- [45] 同37.
- [46] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases[C]//Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 1996: 103-114.
- [47] GUHA S, RASTOGI R, SHIM K. Rock: A robust clustering algorithm for categorical attributes[J]. Information Systems, 2000, 25(5): 345-366.
- [48] KARYPIS G, HAN E H, KUMAR V. Chameleon: A hierarchical clustering algorithm using dynamic modeling[J]. IEEE Computer, 1999, 32(8):68-75.
- [49] PONS-PORRATA A, BERLANGA-LLAVORI R, RUIZ-SHULCLOPER J. Topic discovery based on text mining techniques[J]. Information Processing & Management, 2007, 43(3):752-768.
- [50] CUTTING D R, et al. Scatter/Gather: a cluster-based approach to browsing large document collections[C]//Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992: 318-329.
- [51] ZHAO Y, KARPIS G, FAYYAD U. Hierarchical Clustering Algorithms for Document Datasets[J]. Data Mining and Knowledge Discovery, 2005, 10(2): 141-168.
- [52] SEKT Project[EB/OL]. [2008-07-05]. <http://www.sekt-project.com/>.
- [53] GRCAR M, MLADENIC D, GROBELNIK M. User Profile Inference Module[R]. Technical report, SEKT project deliverable D5.5.2.
- [54] ZAMIR O. Fast and intuitive clustering of web documents[C]//Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997: 287-290.
- [55] WONG W C. Incremental document clustering for web page classification[D]. Hong Kong: Chinese University of Hong Kong, 2000.
- [56] YU M Q. ICT's Approaches to HTD and Tracking at TDT2004[G]. TDT2004 Workshop, 2004.
- [57] CONNELL M. UMass at TDT 2004[C]. TDT2004 System Description, 2004.
- [58] MANI I, BLOEDORN E. Multi-document Summarization by Graph Search and Matching[OL]. Arxiv preprint cmp-lg/9712004, 1997.
- [59] BARZILAY R, ELHADAD N, MCKEOWN K R. Inferring strategies for sentence ordering in multidocument news summarization[J]. Journal of Artificial Intelligence Research, 2002(17): 35-55.
- [60] LAZO-Cortés M, RUIZ-SHULCLOPER J, CABRERA E A. An overview of the concept testor[J]. Pattern Recognition, 2001, 34(4):13 21.
- [61] OntoGen[EB/OL]. [2008-07-05]. <http://textgarden.org/>.
- [62] FORTUNA B, MLADENIC D, GROBELNIK M. Semi-automatic construction of topic ontology[C]//Proc. of ECLM/PKDD Workshop KDO, 2005.
- [63] KOHONEN T. Self-Organizing Maps[M]. Springer, 2001.
- [64] JEPSON A, HEEGER D. Linear subspace methods for Recovering translational direction[C]//Spatial Vision in Humans and Robots: The Proceedings of the 1991 York Conference on Spatial Vision in Humans and Robots, 1993. New York: Cambridge University Press, 1994: 39-62.
- [65] COX T F, COX M A A. Multidimensional Scaling[M]. CRC Press, 2001.
- [66] IOKASOC M I, et al. Design of a user study and evaluation of software tool OntoGen V 2.0[R]. Technical Report FAS Rijeka 2006:12.
- [67] TextGarden[EB/OL]. [2008-07-05]. <http://textgarden.org/>.
- [68] FREY D. Monitoring the news: a TDT demonstration system[C]//Proceedings of the first international conference on Human language technology research, 2001:1-5.
- [69] PONS-PORRATA A, et al. JERARTOP: A New Topic Detection System[J]. Progress in Pattern Recognition, Image Analysis and Applications, 2004, (3287):446-453.
- [70] GRUBER T R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human-Computer Studies, 1995, (43):907-928.
- [71] HEARST M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th conference on Computational linguistics. New York: Association for Computational Linguistics, 1992(2): 539-545.
- [72] PANKOW/C-PANKOW[EB/OL]. [2008-07-05]. <http://km.aifb.uni-karlsruhe.de/pankow/>.
- [73] HARRIS Z. Distributional structure[J]. The Philosophy of Linguistics, 1985:26-47.
- [74] SANDERSON M, CROFT B. Deriving concept hierarchies from text[M]//Research and Development in Information Retrieval. New York: ACM, 1999:206-213.
- [75] FORSYTH R, RADA R. Adding an edge in Machine Learning: applications in expert systems and information retrieval[M]. Ellis Horwood Ltd, 1986:198-212.

[76] CIMIANO P. Learning taxonomic relations from heterogeneous sources of evidence[J]. *Ontology Learning from Text: Methods, Evaluation and Applications*, 2005: 59-73.

[77] GIMIANO P, MAGNINI B. *Ontology Learning from Text: Methods, Evaluation and Applications*[J]. *Frontiers in Artificial Intelligence*, 2005(7):59-73.

[78] FELLBAUM C. *WordNet: an electronic lexical database*[M]. Cambridge, Mass: MIT Press, 1998.

[79] CIMIANO P. Learning taxonomic relations from heterogeneous sources of evidence[M]//BUILELAAR P. *Ontology Learning from Text: Methods, Evaluation and Applications*. [S.l.]: SciTech Book News, 2005: 59-73.

[80] Automatic Content Extraction[EB/OL]. [2008-07-05]. <http://www.nist.gov/speech/tests/ace/>.

[81] KATRENKO S, ADRIAANS P. Learning Relations from Biomedical Corpora Using Dependency Tree Levels[C]//Proc. BENELEARN conference, 2006.

[82] HONG G. *Relation Extraction Using Support Vector Machine*[M]. Berlin: Springer Berlin/Heidelberg, 2005.

[83] APPELT D E, HOBBS J R, BEAR J, et al. SRI International FASTUS System: MUC-6 Test Results and Analysis[C]//Proceedings of the 6th Message Understanding Conference (MUC-6), 1995:237-248.

[84] YANGARBER R, GRISHMAN R. NYU: Description of the Proteus/PET System as Used for MUC-7 ST[C]//Proceedings of the 6th Message Understanding Conference (MUC-7), 1998.

[85] AONE C, SANTACRUZ M R, Rees: A large-scale relation and event extraction system[C]//Proc of the 6th Applied Natural Language Processing Conference, New York, 2000:76-83.

[86] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. *J. Mach. Learn. Res.*, 2003(3):1083-1106.

[87] ZHANG Yimin, ZHOU J F. A trainable method for extracting Chinese entity names and their relations[C]//Proceedings of the second Chinese Language Processing Workshop, ACL, 2000:66-72.

[88] ZHANG Zhu. Weakly-supervised relation classification for information extraction[C]// Proceedings of the Thirteenth ACM conference on Information and knowledge management, Washington D.C., 2004:581-588.

[89] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open Information Extraction from the Web[C]//Proceeding of the International Joint Conferences on Artificial Intelligence, 2007.

[90] SABOU M, d' AQUIN M, MOTTA E. SCARLET: Semantic relAtion discoverY by harvesting onLiNe onTologies[C]//Proceedings of the 5th European Semantic Web Conference, June, 2008.

[91] SPECIA L, MOTTA E. A hybrid approach for extracting semantic relations from texts[EB/OL]. [2008-05-30]. http://www.dcs.shef.ac.uk/~lucia/publications/SpeciaMotta_OLP2-2006.pdf.

[92] Fact[EB/OL]. [2008-07-05]. <http://en.wikipedia.org/wiki/Fact>.

[93] ANDERSEN P M, et al. Automatic extraction of facts from press releases to generate news stories[C]//Proceedings of the third conference on Applied natural language processing, 1992: 170-177.

[94] TJONG K S E. Introduction to FactMine[C]//CLIN-2004.

[95] FactMine: Fact and Ontology Mining for Question Answering[EB/OL]. [2008-07-05] <http://ifarm.nl/erikt/factmine/>.

[96] BRIN S. Extracting Patterns and Relations from the World Wide Web[C]//WebDB Workshop at 6th International Conference on Extending Database Technology, 1998.

[97] YU H, HATZIVASSILOGLU V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]//Proceedings of EMNLP, 2003(3).

[98] Opinion [EB/OL]. [2008-07-05]. <http://en.wikipedia.org/wiki/Opinion>.

[99] LIU Bing. Opinion Mining & Summarization: Sentiment Analysis[C]//WWW-2008 Beijing, 2008.4.

[100] TURNEY P D. Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002.

[101] TURNEY P D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL[C]//Proceedings of the Twelfth European Conference on Machine

Learning, 2001: 491-502.

[102] 同97.

[103] KOBAYASHI N, et al. Collecting Evaluative Expressions for Opinion Extraction[C]//Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP), 2004: 584-589.

[104] POPESCU A M. Information Extraction from Unstructured Web Text[D]. Washington: University of Washington Ph.D. Dissertation, 2002.

[105] 同99.

[106] JINDAL N, LIU B. Mining Comparative Sentences and Relations[C]//Proc. of National Conference on Artificial Intelligence (AAAI' 06), 2006.

[107] GORDON A S, CAO Q, SWANSON R. Automated story capture from internet weblogs[C]//Proceedings of the 4th international conference on Knowledge capture, 2007: 167-168.

[108] WILL B, STUART R B, LJUPCO T. Extracting constraints for process modeling[C]//Proceedings of the 4th international conference on Knowledge capture. ACM: Whistler, BC, Canada, 2007.

[109] ZHU J, UREN V, MOTTA E. ESpotter: Adaptive Named Entity Recognition for Web Browsing[C]//Proc. of Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference, Kaiserslautern, Germany, April, 2005: 11-13.

[110] K-space[EB/OL]. [2008-07-05]. <http://kspace.qmul.net:8080/kspace/kspacewp4.jsp>.

[111] IRIA J, et al. Enhancing enterprise knowledge processes via cross-media extraction[C]//Proceedings of the 4th international conference on Knowledge capture, 2007: 175-176.

[112] DOMINGUE J, DZBOR M, MOTTA E. Magpie: Supporting Browsing and Navigation on the Semantic Web[EB/OL]. [2008-07-05]. <http://kmi.open.ac.uk/people/dzbor/public/2004/domingue-dzbor-motta-iui2004.pdf>.

[113] Watson[EB/OL]. [2008-07-05]. http://watson.kmi.open.ac.uk/editor_plugins.html.

[114] MCDOWELL L K, M.C. Ontology-driven Information Extraction with OntoSyphon[C/OL]//The 5th International Semantic Web Conference(2006)[2008-03-05]. <http://turing.cs.washington.edu/papers/iswc2006McDowell-final.pdf>.

[115] YILDIZ B, MIKSCH S. ontoX-A Method for Ontology-Driven Information Extraction[M]. *Computational Science and Its Applications-ICCSA 2007*, Springer-Verlag, LNCS 4707:660-673.

[116] PASCA M. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge[C]//Proc. of AAAI-2006, 2006.

[117] PASCA M. Names and similarities on the web: Fact extraction in the fast lane[C]// Procs. of ACL/COLING, 2006.

[118] MAEDCHE A, PEKAR V. Ontology Learning Part One-On Discovering Taxonomic Relations from the Web [EB/OL]. [2008-07-05]. http://projekte.l3s.uni-hannover.de/pub/bscw.cgi/d7841/Maedche_Pekar_Staab-Ontology_Learning-Web_Intelligence_Sub.pdf.

[119] CIMIANO P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*[M]. Springer US, 2006:19-34.

作者简介

张智雄 (1971-), 男, 中国科学院国家科学图书馆研究员、博士生导师, 发文60余篇。通讯地址: 北京市海淀区中关村北四环西路33号, 中国科学院国家科学图书馆 100190

吴振新 (1968-), 女, 中国科学院国家科学图书馆, 副研究员, 发文40余篇。通讯地址: 同上

赵琦 (1983-), 女, 中国科学院国家科学图书馆, 在读硕士研究生, 发文5篇。通讯地址: 同上

洪娜 (1980-), 女, 中国科学院国家科学图书馆, 在读博士研究生, 发文5篇。通讯地址: 同上

徐健 (1977-), 男, 中国科学院国家科学图书馆, 在读博士研究生, 中山大学资讯管理系讲师, 发文8篇。通讯地址: 同上

刘建华 (1984-), 女, 中国科学院国家科学图书馆, 在读硕士研究生, 发文8篇。通讯地址: 同上

(下转36页)

13	肥胖是骨关节炎的高危因素	OBESITY(肥胖), HYPERTENSION(高血压), BODY MASS INDEX(BMI),
14	透明质酸治疗骨关节炎的机制	HYALURONIC ACID(透明质酸),
15	骨关节炎中关节软骨损伤的动物实验模型	ARTICULAR CARTILAGE(关节软骨), DAMAGE(损伤), HORSES(马),
16	骨密度与骨关节炎的关系	BONE(骨), BONE MINERAL DENSITY(骨矿物质密度),
17	风湿性关节炎与骨关节炎的关系	RHEUMATOID ARTHRITIS(风湿性关节炎),
18	骨关节炎基因疗法中通过抑制炎症性细胞因子、基质金属蛋白酶的表达以及促进生长因子的表达的研究	EXPRESSION(表达), CYTOKINES(细胞因子), IL1-BETA(白细胞介素), MATRIX METALLOPROTEINASES(基质金属蛋白)
19	传统非甾体类抗炎药治疗骨关节炎	CYCLOOXYGENASE INHIBITORS(环加氧酶抑制剂), NSAIDS
20	手关节炎	HAND(手), HAND OSTEOARTHRITIS(手关节炎)

参考文献

[1] 刘华. 基于文本分类中特征提取的领域词语聚类[J]. 语言文字应用, 2007(1):139-144.
 [2] Essential Science Indicators[EB/OL]. [2007-08-01]. <http://www.esi-topics.com/RFmethodology.html>.
 [3] POTTENGER W M, KIN Yong-Bin, MELING D D. HDDITM: Hierarchical Distributed Dynamic Indexing[EB/OL]. [2007-08-01]. <http://www.cse.lehigh.edu/~billp/pubs/HDDIFinalChapter.pdf>.
 [4] KLEINBERG J. Bursty and hierarchical structure in streams[EB/OL]. [2007-08-01]. <http://www.cs.cornell.edu/home/kleinber/bhs.pdf>.
 [5] 魏晓俊. 基于科技文献中词语的科技发展监测方法研究[J]. 情报杂志, 2007(3):34-39.
 [6] 数据挖掘中聚类分析的技术方法[EB/OL]. [2007-08-01]. <http://bidwhome.itpub.net/post/20871/155927>.
 [7] 梁立明, 武夷山. 科学计量学: 理论探索与案例研究[M]. 北京: 科学出版社, 2006.S

作者简介

殷蜀梅 (1977-), 女, 北京大学医学图书馆, 馆员, 发文7篇。
 通讯地址: 北京市海淀区学院路38号北京大学医学图书馆 100083
 张智雄 (1971-), 男, 中国科学院国家科学图书馆研究馆员、博士生导师, 发文60余篇。通讯地址: 北京市海淀区中关村北四环西

路33号, 中国科学院国家科学图书馆 100190

A Method for Topic Extraction and Clustering Based on Medical Literature

Yin Shumei / Peking University Health Science Library, Beijing, 100083
 Zhang Zhixiong / National Science Library, Chinese Academy of Sciences, Beijing, 100190

Abstract: Important keywords in academic papers reflect topics of the literature. Therefore, the extraction of topics turns to be the extraction of keyword groups. This paper first investigates techniques for topic extraction and clustering used by overseas, then the researchers propose a technical scheme for extracting topics in text information resources in the medical field and for topic area identification. A detailed explanation of the techniques for topic clustering is given. To verify the validity of the method, this paper applies the scheme to the field of osteoarthritis research. The result proves the validity of the proposed method.

Keywords: Knowledge extraction, Topic extraction, BM25F, MMTx, Text mining, Medical data mining, Digital library

(收稿日期: 2008-07-13; 责任编辑: 贾廷霞)

(上接12页)

Review of the Technologies and Methods for Extracting Content Objects from Unstructured Text

Zhang Zhixiong, Wu Zhenxin / National Science Library, Chinese Academy of Sciences, Beijing, 100190
 Zhao Qi, Hong Na / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049
 Xu Jian / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049; Department of Information Management, Sun Yat-Sen Univ., Guangzhou, 510275
 Liu Jianhua / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049

Abstract: In recent years, knowledge extraction plays a very important role when dealing with unstructured text. In this paper, based on the analysis of current relevant literature, systems and projects, it proposes the classification of the current knowledge extraction objects and reviews the relevant technologies and methods. The major themes include web object identification and integration, terminology extraction, topic discovery, conceptual hierarchy relation extraction, non-conceptual hierarchy relation extraction, fact extraction and opinion extraction. This paper also analyzes trends of knowledge extraction in the future.

Keywords: Knowledge extraction, Object identification, Terminology extraction, Topic discovery, Relation extraction, Fact extract, Opinion extraction, Digital library

(收稿日期: 2008-07-13; 责任编辑: 贾廷霞)