情报工程 ISSN 2095-915X
第2卷 第3期 80-90 2016年6月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.3 80-90 Jun 2016

doi:10.3772/j.issn.2095-915x.2016.03.010

Stanbol 系统及其在国外应用概述

陈田田，吴广印

（中国科学技术信息研究所 北京 100038）

摘要：web 中大量新闻网页、博客、电子邮件等非结构化信息中蕴含着大量的知识，对其进行处理以自动获得知识具有重要意义。目前，一些基于信息抽取等技术抽取简单关联关系的知识获取应用系统存在明显的局限性，本文引入 Apache Stanbol——Apache 下的一种从非结构化信息中自动获取知识的开源项目，它是一个为语义内容管理设计的模块化的软件集和可重用组件，旨在将传统内容管理系统（CMS）拓展为支持语义服务的语义内容管理系统（SCMS），在此基础上，为改善搜索引擎关于内容的搜索、分类，实体消歧及语义化查询等带来帮助。

关键字：Apache Stanbol，非结构化信息分析，知识获取

中图分类号：G250

Review on Stanbol System and Its Application in Foreign Countries

CHEN TianTian，WU GuangYin

（Institute of Scientific and Technical Information of China，Beijing 100038，China）

Abstract: There are a large number of news pages, blogs, e-mail and other unstructured information in the network, which contains tremendous knowledge, thus, processing these information and obtaining knowledge automatically become more important. Recently, application system of knowledge acquisition based on information extraction and other techniques exist obvious limitations. This paper introduced the Apache Stanbol: an open source project for automatic acquisition of knowledge from unstructured information under Apache.

作者简介：陈田田（1991–）女，硕士研究生，中国科学技术信息研究所；吴广印（1965–）男，中国科学技术信息研究所研究员，北京万方软件股份有限公司董事长，研究方向：云计算、知识组织、大数据挖掘与分析。

It is a modular software set and reusable component for semantic content management, and it can also extend traditional content management systems with semantic services. This open source project may benefit for the research of the search engine, classification, entity disambiguation and semantic query.

Keywords: Apache Stanbol，unstructured information analysis，knowledge acquisition

引言

信息爆炸时代一方面使人们置身于海量信息中，另一方面为准确获取目标信息带来挑战。网络信息资源已经成为全球规模最大的资源库，然而 web 上的数据大多以非结构化或半结构化形式存在，因而牺牲了数据间的结构信息和其中包含的语义，为了获取目标信息，我们需要处理大量无用信息。W3C 主席 Tim Berners-Lee 提出语义网的基本思想是为计算机提供可供其处理的包含语义的数据，从而使得计算机自动化处理过程和 web 信息集成更为方便。他设想的智能代理(agent) 为人类提供方便快捷的信息搜索、信息过滤等服务，充分发挥了 web 的潜力 [1]。

语义网远景的实现需要大量经过语义标注的文档。概况的说，语义标注是将文本中的概念通过 URI 与知识库中实体的映射。W3C 组织制定了一系列的标准和工具用于实现这一远景，如 RDF（Resource Description Framework）和 OWL 技术（OWL 被设计用来处理内容的信息而不仅仅是向人展示信息）。为了加快语义网的发展，以统一、规范的方式发布和互联 RDF 数据集，LOD （Linked Open Data）社区出现了。它旨在将 Web 上的开放数据源如维基百科、GeoNames、Musicbrainz、Wordnet 等以 RDF 的方式发布出来，同时生成数据源之间的 RDF 链接，以供搜索引擎以及更高级的应用程序使用，包括：DBPedia①、Freebase②、YAGO③等。

Web 中的非结构化信息经过语义标注处理后得到可被机器理解的、结构化信息，可以为语义检索和异构数据源间的互操作性等带来好处 [2]。由于一些基于信息抽取等技术抽取简单关联关系的知识获取应用系统存在明显的局限性 [3]，目前已有一些从软件工程方面，着手对非结构化数据进行处理的系统和框架，Apache Stanbol 是其中之一。Stanbol 支持对内容中出现的实体与对应在 DBPedia 中的实体关联，并在此基础上对内容信息进行索引、知识推理等服务。因此，本文希望通过对 Apache Stanbol 系统进行相关概述及系统内部客观分析，以期使 Stanbol 系统为更多人所用，更好的服务于图情领域的知识服务、数据挖掘等领域研究。

1. 相关系统实现情况

非结构化信息是网络资源中的重要组成部分，并以快速方式增长。据统计，当今世界结构化数据增长率大概是 32%，而非结构化数据增长

① http://wiki.dbpedia.org/
② http://wiki.freebase.com/wiki/Main_Page
③ http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

则是 63%，因此，对其进行高效管理、分析和知识获取具有重要意义 [4]。目前，对非结构化信息进行开发利用在国内外引起广泛关注，无论在学术界还是工业界，都已出现一些有价值的研究成果和比较成熟的商业应用产品。

AlchemyAPI 主要提供实时文本分析和计算机视觉服务，可以从大量网页、文档、微博和图像等实时进行情感分析、关键词抽取、实体抽取、关系抽取、概念标注、图像标注等服务，同时支持对抽取结果的可视化展示，功能十分强大。如，开发者可通过 AlchemyAPI，可对新闻网页中实体进行抽取（包括人名、地名、公司等）或对图片中的内容进行分析等。对文本类型信息支持处理的语言包括英语、法语等，目前不支持中文。

Opencalais 是汤森路透（Thomson Reuters）下的一种商业应用解决方案。它能对非结构化文本中实体（人物、机构、产品等）、实体间关系（work-for，located-at）、事件信息进行抽取，并能根据文本内容抽取出相应的主题标签（教育、社会事件、政治等）。Open Calais 从文本中抽取的元数据信息可以用 RDF、JSON 等形式展示，用户可以方便的从抽取结果中获取对文本内容知识及其相关信息的获取，如根据抽取出的元数据信息进行浏览和搜索。目前，用户可以从公开的网站（http://www.opencalais.com/）获取免费的 Opencalais 服务，可以抽取诸如公司、人物、地名、产业类型等元数据信息，免费用户每日最多可上传 5000 份文档进行分析。与其他标注服务相比，Opencalais 具有更高的准确性，目前不支持中文。

Apache UIMA（Unstructured Information Management Architecture ）是 Apache 软件基金会支持的一个开源项目，它是一种对非结构化内容，如文本、视频和音频数据进行分析的软件框架。UIMA 的主要目标是通过对实体和实体关系的抽取，为非结构化数据世界和结构化数据世界间搭建一座桥梁。UIMA 由于在非结构化信息分析上表现出高效性、扩展性和通用性等特点，在企业的非结构化信息管理方面已有不少的探索和应用 [5]。UIMA 不支持使用 LOD，因此，对抽取结果缺乏搜索和推理（Search&Reasoning）的实现。

表 1 相关系统特性比较

	AlchemyAPI	OpenCalais	UIMA
主页网址	http://www.alchemyapi.com/	http://www.opencalais.com/	https://uima.apache.org
支持语言	英语、法语等，不支持中文	英语、法语和西班牙语，不支持中文	英语、德语等
服务分析	实体抽取、情感分析、关系抽取、文本分类、关键词抽取等等	实体抽取、主题分类	实体识别；关系抽取
抽取实体中包含的属性	实体类型；实体间关系；出现频次；文本字符串	实体类型；实体间关系；出现频次；文本字符串	实体类型；实体间关系

支持链接的知识库	Freebase; GeoNames; OpenCyc; YAGO 等	DBpedia; Wikipedia; Freebase; Reuters.com; GeoNames; Shopping.com 等	不支持
服务方式	分为免费和有偿服务	分为免费和有偿服务	开源项目
开发语言	——	——	Java;C++
分析结果数据格式	XML, JSON, RDF 等	RDF, JSON, or N3 格式	XML,XMI 等

以上三种对非结构化信息处理的工具，为非结构化分析提供一个通用的平台，减少了重复开发的可重用分析组件。从对知识挖掘的深度来看，AlchemyAPI 和 OpenCalais 支持对外部数据源的关联，提供对知识表示、推理检索的支持，而 UIMA 则缺乏相关实现。而从项目开源情况来看，前两种系统都为商业服务，对于需要大规模应用的用户来说需要一定成本。与上述系统相比，Apache Stanbol 项目是一种对非结构化信息进行自动知识获取的开源实现，灵活性好、安全性高。基于此，本文重点分析了 Stanbol 系统内在工作机理，以期为后续研究提供工作基础。

2. Stanbol 系统分析

2.1 Stanbol 概况

2.1.1 系统概况

Apache Stanbol④是 Apache 软件基金会（Apache Software Foundation）下的一个开源项目。该项目成立于 2010 年 10 月，由欧盟研究项目 IKS 发起，并在 2012 年底完成对该项目的开发。

Apache Stanbol 采用了基于组件的软件开发方式和面向对象的灵活编程。它是一个为语义内容管理设计的模块化的软件集和可重用组件，每个组件通过 RESTful 应用编程接口（API）对外提供功能服务，采用纯 Java 语言开发的免费开源软件，遵循 Apache license⑤。它旨在将传统内容管理系统（CMS）拓展为支持语义服务的语义内容管理系统（SCMS）。它的核心功能之一为对非结构化文本中的实体（人物、地点、机构）进行抽取，并将这些实体自动链接到 web 上的链接开放数据源（LOD），如 DBPedia。

通过 Stanbol 系统对文档内容进行语义标注后，文档中的内容不再是毫无意义的字符串，实体属性及实体间的关系丰富了原有内容信息，进而实现对非结构化信息的萃取。Stanbol 系统能作用于文本、视频、图像等多种内容类型，最常用的是对文本内容的信息抽取。其对语言种类的支持丰富，包括：英语、法语、意大利语等多种语言，

④ https://stanbol.apache.org/
⑤ https://stanbol.apache.org/ses/LICENSE-2.0

目前不支持对中文信息的处理。为了便于叙述，本文接下来将以 Stanbol 对文本内容类型的处理为内容主线进行展开。

2.1.2 Chain 工作模式

经 Stanbol Enhancer 组件处理后，内容信息中实体即与知识库中对应在真实世界中实体关联，得到的包含语义信息的增强结果以 RDF 形式表示。这种增强结构（Enhancement Structure）主要带来了两方面的好处：

（1）内容增强引擎（EnhancementEngines）之间的互操作性的好处。Stanbol 内容增强引擎按功能划分，可以分为三大类，预处理引擎、内容分析引擎和语义提升引擎。Stanbol Enhancer 通过链（chain）的方式将这三类组件组合，（如图 1 所示）实现对内容的语义增强。一条链中包含有多个对内容处理的引擎，后一个功能引擎在前一个功能引擎抽取的信息基础上继续完成该功能模块。这种增强结构确保了不同组件间数据信息的共享性，一个引擎处理后产生的元数据信息可以被后面的组件使用。如在预处理引擎中语言识别组件识别出文本内容中的语言类型，紧随其后的内容分析引擎根据识别出的语言类型选择正确的

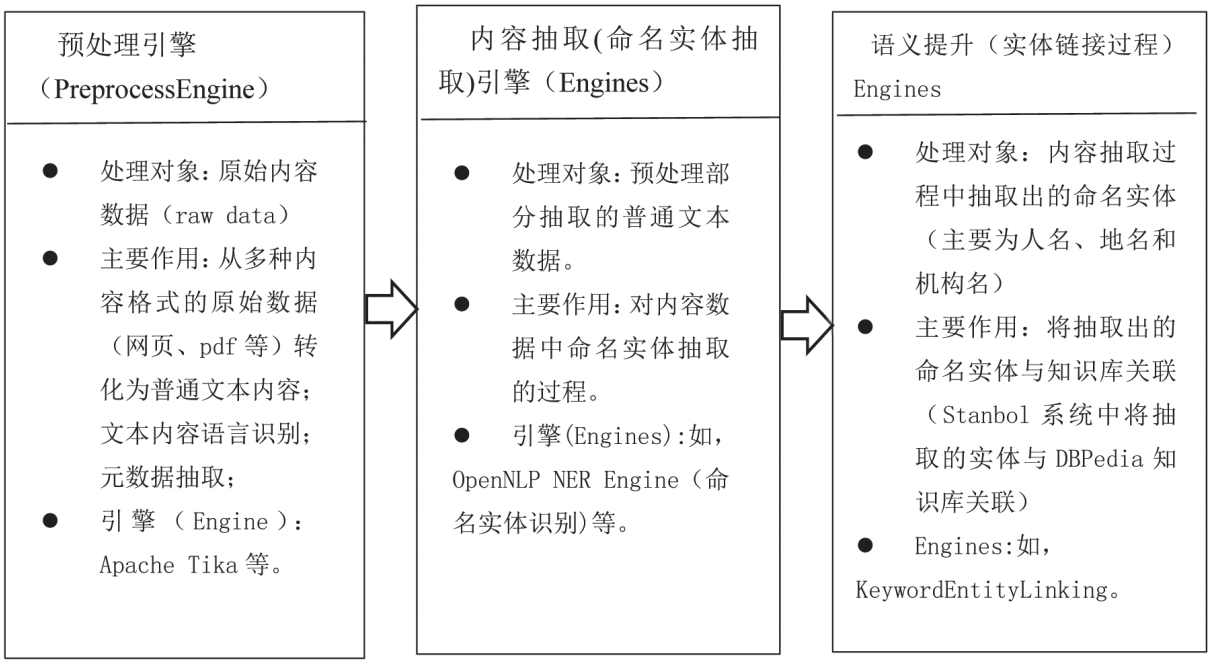

图 1　语义增强引擎（Enhancer&Enhancement Engines）分类及其工作流程

命名实体识别（NER）模型。

（2）为用户使用信息抽取的结果提供方便。互联网中各种新闻、博客等网页中包含大量的实体，它们散乱无规则地分布，尽管其中包含了经济、政治、文化等有重要战略价值和意义的信息资源，但是人们并不能直接从非结构化网页中挖掘出有利用价值的信息。互联网中这些海量信息成为亟待开发利用的巨大宝库。Stanbol 系统对发送来的文本信息，经内容分析引擎和语义提升引擎处理后，使文本内容中的信息不再只是字符串，而是经过处理后具有结构化信息、包含丰富语义的大量实体信息。这就使得以下功能得以实现：

● 对文档内容分类（明确文档内容包含的实体类型）

● 确定实体类型（实体消歧）

● 统计文本中命名实体出现频次

正如谷歌高级副总裁埃米特·辛格博士所言：“构成这个世界的是实体，而非字符串（things,not strings）”，文本内容中出现的实体与知识库中相应词条链接将搜索引擎从字符串匹配跳跃到对实体及实体间关系和属性的层面，为下一代搜索引擎用户提供的应用场景提供了无限遐想。

2.2 Stanbol 的作用

（1）实体标注（Entity tagging）

命名实体是文本中的基本元素，也是传递信息最重要的要素。一般来说，文本中命名实体类型主要包括人名、地名、机构名、时间、日期、货币等类型。Stanbol 系统能将文本字符串“Bob Marley”（人名）抽取，并将该实体关联到

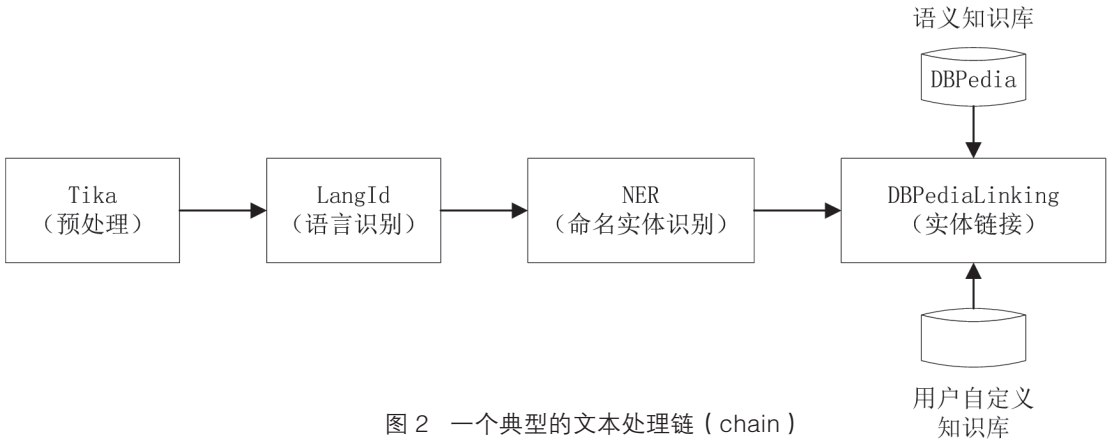

图 2 一个典型的文本处理链（chain）

RDF 知识库（如，dbpedia）中对应的实体 URI（dbpedia:Bob Marley），从而获取到该实体对应在知识库中所包含的结构化信息，实现语义扩充。

这将会改善搜索引擎关于内容的搜索。知识库中包含众多实体、实体属性及实体间关系的信息。利用文本中众多链接过的实体及其属性（元数据）信息，搜索引擎可以对内容的“理解”从字符串匹配推进到实体层面，也就是将其与现实世界中的一个真实事物进行映射。如对查询词“李娜大满贯”，如果基于字符串匹配，会产生搜索引擎无法判断用户查询词中的“李娜”对应的是中国女子网球名将还是中国女子跳水运动员等歧义现象。因而机械地返回所有含有“李娜”这个关键词的网页。而查询词中的实体与知识库映射后，搜索引擎可以利用实体及其属性（“大满贯”）能准确“理解”到“李娜”对应为中国女子网球运动员这一实体，因此可以准确理解用户查询意图。从字符串到实体，搜索引擎不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

（2）实体消歧（Disambiguation）和内容分类

网络用户的信息查询行为主要包括信息检索行为和信息浏览行为[6]。一方面，从非结构化信息中检索实体（人名、地名、机构等）信息在信息检索、问答系统、推荐系统等领域有重要应用。近几年的研究成果显示，实体相关搜索占互联网查询的很大部分，并且这个比例还在不断上升[7]。由于自然语言表达的多样性和丰富性，实体歧义现象普遍存在。而搜索引擎网页中对实体查询的结果缺乏层次性，往往包含多个共享同一名称对应为不同对象的信息，造成返回信息鱼龙混杂，检索准确性大打折扣。而通过实体标注（Entity tagging），可以准确的将歧义的实体与真实世界中的实体一一映射，准确关联。

另一方面，对 web 中信息按内容类型进行高效分类，减少用户以浏览方式获取信息成本。传统的对内容分类需要经人工对文献、新闻、博客等内容信息进行标引，不仅效率低，准确率也难以保证 [8]。而基于文本自动分类的研究也存在处理过程复杂、耗时、计算量庞大、缺乏对内容的语义理解层次等弊端。如，被标记为“李娜”标签的博文，不能被自动分类到“网球运动员”，尽管二者存在相关关系；而对博文中的实体标注后，利用知识库中存在的实体关系（职业，“李娜”，“网球运动员”），即可将该博文正确分类到相关标签下。因此，基于知识库的内容分类得以使用户能真正按照知识结构脉络有层次、有深度地对内容信息的获取。

2.3 Stanbol 系统的整体架构

Stanbol 系统中，每个组件（Component）是相互独立的，可经由其自身的 RESTful 接口获取服务，因此用户可以依据不同的应用场景对各组件进行灵活组合。

Apache stanbol 核心组件及其功能说明如下：

1. 语义增强引擎（EnhancerEngine）

增强组件及其引擎（Enhancer&Enhancement Engines）使用户能将内容提交给 Apache Stanbol 及获取经过实体抽取和实体链接后返回的结果。对该过程的实现主要分为三个阶段，见图 1。通过分析文本内容，识别在非结构化文本中的实体（人物、地点、机构），并将实体链接到开放数据源中，最后以 RDF 数据结构的形式返回抽取的信息。

2. 内容库（Contenthub）

内容库是基于 Apache Solr 的文档库，可以用来存储文本文档及在此基础上的自定义语义搜索功能。存储在其中的文档称为内容项（即 Content items，包括文档元数据和文本内容）。它包括两个子组件：Store 和 Search。存储在 Content Store 中的文档及其元数据通过 Apache Solr 建立索引后，每个文档被分配唯一的 ID，通过该 ID 即可从 Contenthub 中对文档进行检索（Retrieval）和删除（Delete）等操作。用户上传的文档首先存储在 Contenthub 中，并将其文本内容转送到增强组件（Enhancer），经内容提升的过程，获得语义增强后的结果。

3. 实体库（Entityhub）

实体库用来管理与特定领域实体相关的信息。它是一个通用的组件，可以连接到一个开放链接数据库，如 DBPedia，从而获取到各种来源丰富的、与实体关联的信息。

4. 本体管理（Ontology Management）

本体是用来定义描述内容元数据的知识模型。此外，元数据的语义信息也可通过本体定义。本体管理组件提供对本体的管理。

5.CMS 适配器（CMS Adapter）

CMS 适配器为内容管理系统（CMS）和 Apache Stanbol 之间交互提供桥梁。主要功能包括：双向映射（Bidirectional Mapping）和 Contenthub Feed。双向映射，即 CMS 可用 RDF 数据结构表示其内容库的结构；另一方面，存在于 web 的 open linked data 可以映射到 CMS 内容库中。Contenthub Feed，提供对 Contenthub 中内容数据的管理，主要包括两种操作：提交和删除。

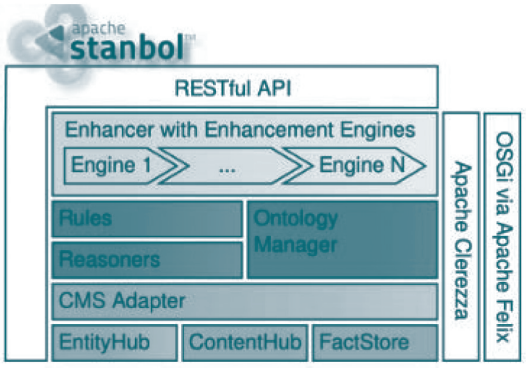

图 3 Apache Stanbol 系统架构 [9]

3. Stanbol 系统实现英文信息抽取

3.1 Stanbol 实现英文信息抽取过程

从宏观的角度看，Stanbol 系统对文本信息抽取过程可以分为两个子过程：自然语言处理（NLP）子过程（抽取出命名实体）和语义提升子过程（实体链接处理）。对于自然语言处理子过程，Stanbol 采用 Apache OpenNLP 作为默认的 NLP 处理框架。OpenNLP 支持句子切分、分词、词性标注（POS，Part of Speech）、分块（Chunking）和命名实体识别（NER，Named Entity Recognition）等通用的处理。该系统支持对多种语言的处理，包括英文、法语等，目前对中文的支持不理想，仅能做到对中文句子的切分，远不能实现对中文语义提升的任务。本文以对英文信息处理为例说明整个处理流程。

Stanbol 系统对该任务的完成采用流水线的工作方式，即逐个处理抽取环节。具体来说，对待一篇待处理文档，经过类似流水线一般的处理，严格按照规定顺序经过英文分句、英文分词、英文词性标注、命名实体识别之后，将抽取到的命名实体与知识库中对应的实体进行关联。获取到实体对应在知识库中的 URI 后，即可在对内容集中文档数据建立语义索引的基础上进行语义检索（知识发现、推理的过程）。整个处理流程如图 4 所示：

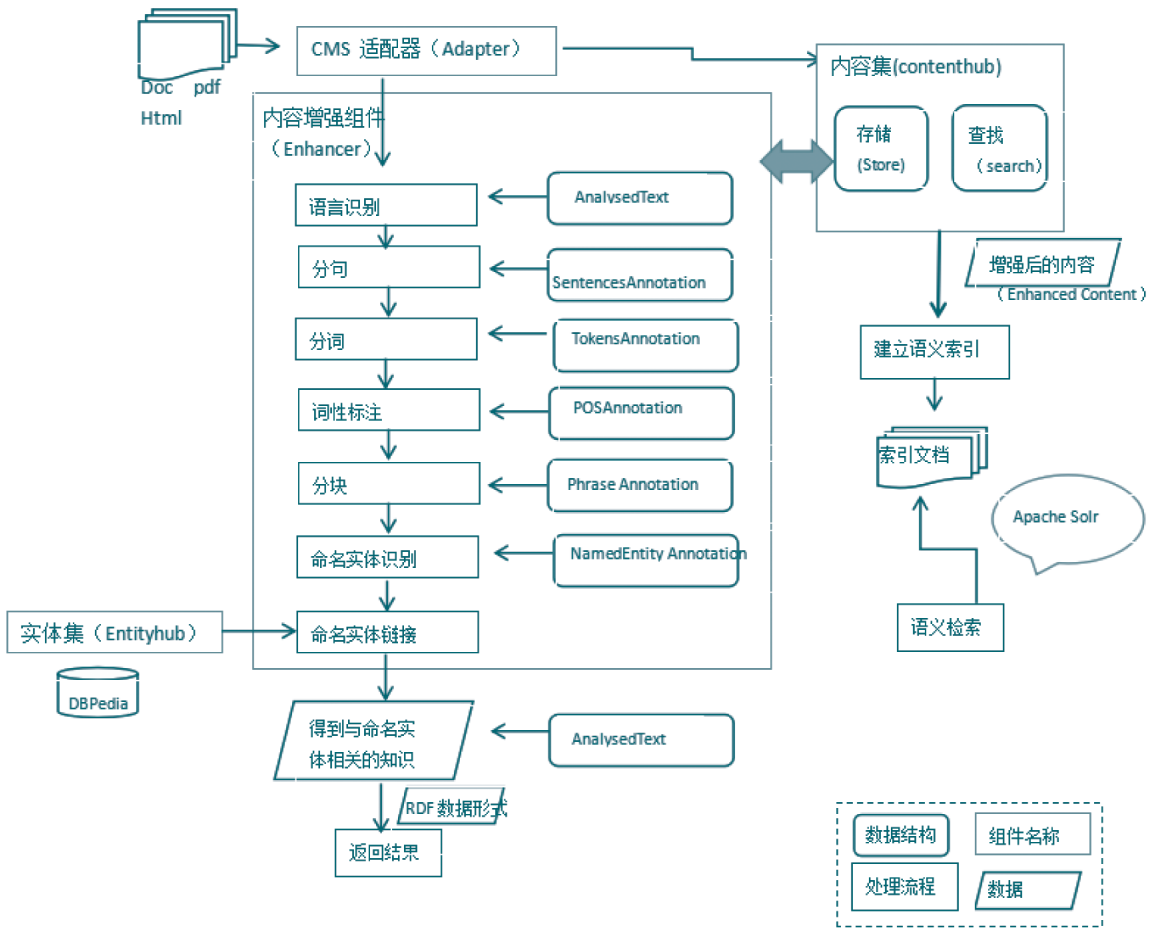

图 4　Stanbol 实现英文信息处理流程

3.2 Stanbol 系统标注集的数据结构分析

为了更进一步的了解 Stanbol 的英文信息抽取过程，需要分析在整个信息抽取过程中，输入数据被存储的数据结构，及对该数据结构进行操作的应用编程接口（API）。

文本信息数据在被 Stanbol 系统处理过程中，ContentItem 数据结构作为可被 Enhaner 处理对象，该对象存在于整个处理过程中；NLP 处理过程中主要通过 AnalysedText 对象和 NLPAnnotation 对象实现数据流动的；实体链接处理中，ContentItem 对象由 ContentItemFactory 创建，是 Stanbol Enhancer 可以处理的内容单元，该数据结构主要由两部分组成 Contentparts（包含 Blob 类型的原始输入数据和中间 NLP 处理过程中对内容信息处理后的结果等）和 Metadata（标注结果的元数据信息）。在整个语义提升处理过程中用来储存处理结果，因此 ContenItem 即为最后标注结果。

AnalysedText 数据结构用来存储 NLP 处理过程中的结果，如 Sentences、Tokens 和 Chunks。AnalysedText 对象中存储的数据需通过 URI 作为 ContentPart 注册到 ContentItem 中，因为在 Stanbol 系统处理中，ContentItem 对象是可被处理对象。

NLPAnnotation 数据结构定义了对典型的 NLP 处理结果的标注模型，POS tags、chunks 和命名实体识别（NER）的标注等。

下面以一段文本处理结果的实例来说明标注结果如何存储在 Stanbol 的数据结构中。

例如文本“Stanford University is located in California.”

Text

[Stanford University]$_{Organization}$ is located in [California]$_{Place}$.

表 3 包含分析结果的标注集信息

Selected-text	language	Span-start	Span-end	Entity-type	Entity-label	Entity-reference
"Stanford University "	en	“0”	“19”	"http://dbpedia.org/ontology/Organization"	"Stanford University"	"http://dbpedia.org/resource/Stanford_University" "http://dbpedia.org/ontology/EducationalInstitution" "http://dbpedia.org/ontology/Organisation" "http://dbpedia.org/ontology/University"
"California"	en	“34”	“44”	"http://dbpedia.org/ontology/Place” "http://dbpedia.org/ontology/PopulatedPlace" "http://dbpedia.org/ontology/AdministrativeRegion"	"California" "Baja California" "Southern California"	"http://dbpedia.org/resource/California" "http://dbpedia.org/resource/Baja_California" "http://dbpedia.org/ontology/PopulatedPlace" "http://dbpedia.org/ontology/Settlement" "http://dbpedia.org/resource/Southern_California"

从表 3 中我们可以看到，selected-text 是从文本信息中抽取到的命名实体，language 字段为识别出文本的语言类型，span-start 表示抽取到的命名实体起始位置，span-end 表示抽取到的命名实体结束位置，entity-label 表示实体的名称信息，entity-reference 表示实体在知识库中对应的 URL。

4. 国外应用情况

Apache Stanbol 系统在设计之初就与图情领域的应用有着紧密的结合。早在 2002 年，就有学者指出“互联网上的检索技术和知识发现技术”是情报学未来新的研究方向和热点 [10]。 Apache Stanbol 是一个全栈式的语义内容管理框架，基于 RDF 知识库，系统可以对不同内容类型（文档、视频、图片等）的资源进行语义标注，进而为实现在此基础之上的语义检索和知识检索提供基础，目前国外已有相关应用成果。

针对当前 web 中多媒体资源仍未包含有描述该类型资源的元数据信息这一问题，Damjanovic 等人 [11] 讨论了将 cms 中内容赋予语义信息所使用的方法和技术（包括了语义搜索和浏览；语义标注；语义分析和知识发现）以及将 Stanbol 与 LMF（Linked Media Framework）整合的设想。LMF 是一个基于 Linking Open Data(LOD) 的开源项目，为多媒体内容和它的元数据发布为关联数据（linked data）提供了一个框架。因此，二者的整合意义在于通过对大量丰富的数据分析其文本和媒体内容，并用一种统一的方式为媒体资源提供内容和元数据的存储和检索工作，可以提供更好的内容增强功能。

Gönü [12] 等人分析认为，CMS 对管理的内容不支持任何的语义检索、查找、浏览功能，或者语义查找的功能是有限制的。在该研究中他们利用 Apache Stanbol 项目弥补了现有 CMS 内容中语义描述的不足。利用从关联数据（LOD cloud）中获取的信息作为 CMS 管理的内容标注的信息来源，且这种语义索引结构允许在此之上进行语义有意义的搜索功能，达到从 cms 的结构和其中的真实内容中抽取隐藏的知识内容管理系统的目的。

Drupal 作为一种强大的内容管理系统之一，也可以利用 Stanbol 对内容管理系统功能的扩展。 Hangl 等人 [13] 介绍了一种 Drupal 内容管理系统的扩展功能。它将语义技术整合到 CMS 中，为改善搜索、整合和智能管理内容带来明显的益处。

上述都是关于 Stanbol 在国外应用的相关情况，目前国内还没有关于 Stanbol 应用的文献。其原因为目前 Stanbol 系统对中文支持程度不足以开展大规模应用。Stanbol 系统中对文本信息处理并进行语义提升是以一种 pipeline 的方式进行，而其缺乏对中文自然语言处理（即分词、分句、命名实体识别、词性标注等）相关功能组件的支持。因此，为了能实现 Stanbol 对中文信息处理的支持，未来的工作重点在于利用一些自然语言处理（Natural Language Processing, NLP）工具，实现对中文语言处理相关模块的支持。

5. 总结与展望

目前，对非结构化信息开发及管理得到广泛关注，Stanbol 系统作为一种全栈式的语义内容管理框架，其主要作用是通过一系列可重用的语义内容管理组件，将传统内容管理系统（CMS）拓展为支持语义服务。其支持将非结构化信息中的实体与 LOD 社区中的 DBPedia 知识库关联，进而为用户提供知识推理、语义检索等智能知识服务。

值得关注的是，该系统对中文信息处理不够

理想，因此，需要利用一些自然语言处理（Natural Language Processing）工具，实现对中文语言处理相关模块的支持。

参考文献

[1] Berners-Lee T, Hendler J, Lassila O, et al. The Semantic Web", Scientific American[J]. Lecture Notes in Computer Science, 2001, 284(October):34—43.

[2] Uren V, Cimiano P, Iria J, et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art[J]. Web Semantics Science Services & Agents on the World Wide Web, 2005, 4(1):14-28.

[3] 邹志鹏，饶若楠．一种面向非结构化信息的知识获取框架 [J]. 微型电脑应用 ,2010,26(8):18-21.

[4] 非结构化数据来袭 50%~75% 数据来自人与人互动 [EB/OL]. [2011-07-13].http://tech.sina.com.cn/i/2011-07-13/11545774485.shtml.

[5] 邱明辉，彭强．非结构化信息管理框架的原理与应用 [J]. 情报探索，2011(1):93-95.

[6] 曹双喜，邓小昭．网络用户信息行为研究述略 [M]// 网络用户信息行为研究．北京：科学出版社，2010:79-81.

[7] 姜丽丽．实体搜索与实体解析方法研究 [D]. 兰州：兰州大学，2012.

[8] 钱爱兵，江岚．基于标题的中文新闻网页自动分类 [J]. 现代图书情报技术，2008(10):59-68.

[9]Apache Stanbol Components[EB/OL].[2016-01-05].https://stanbol.apache.org/docs/trunk/components/

[10] 赖茂生，张莉扬．情报学的学科发展与教育问题 [J]. 情报学报，2003, 22(1):3-9.

[11] Damjanovic V, Kurz T, Westenthaler R, et al. Semantic Enhancement: The Key to Massive and Heterogeneous Data Pools[J]. 2010:413-416.

[12] Gönül S, Sinaci A A. Semantic content management and integration with JCR/CMIS compliant content repositories[J]. 2012:181-184.

[13]Hangl S, Toma I, Thalhammer A. Introducing a Diversity-Aware Drupal Extension[C]//I-SEMANTICS (Posters & Demos). 2013: 20-24.