语义知识图谱关键技术与应用

Semantic Knowledge Graph and Application

张华平 博士 副教授

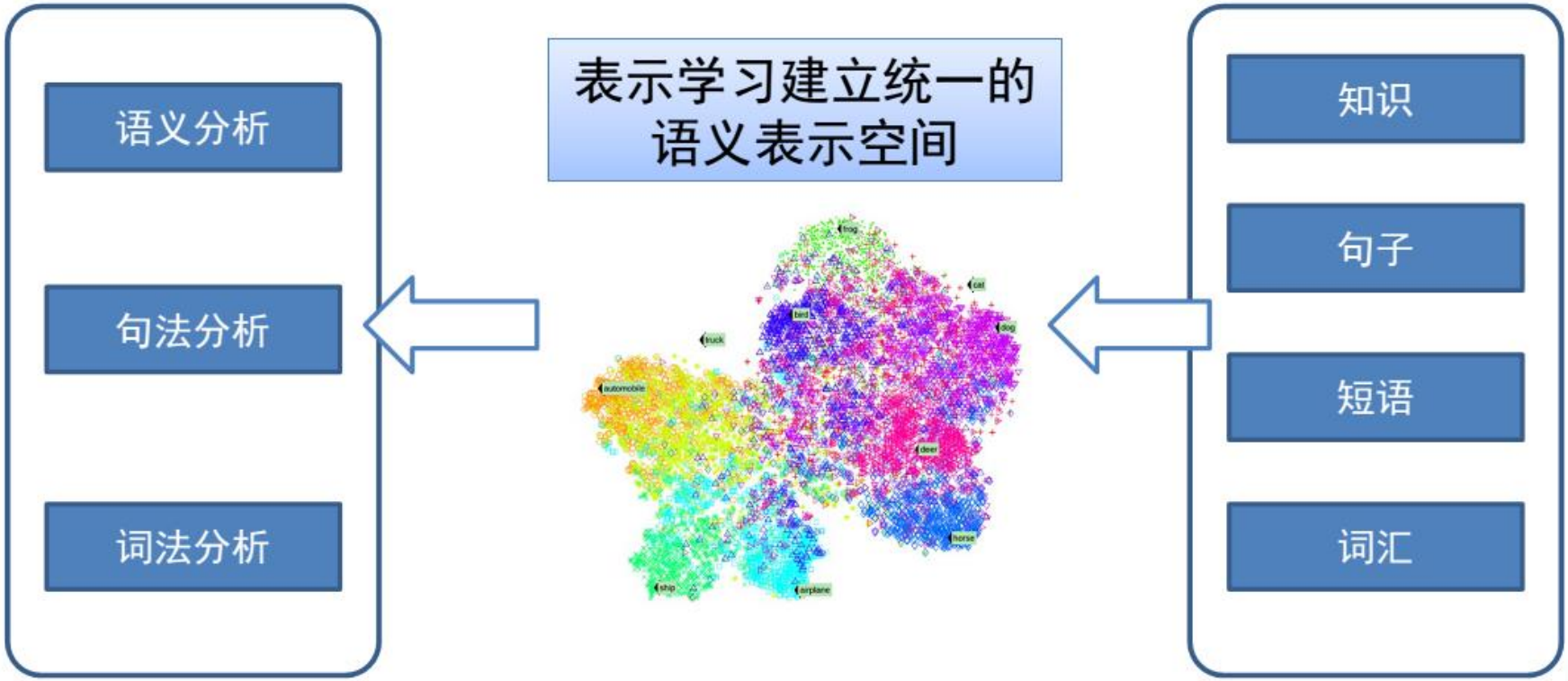
大数据搜索与挖掘实验室

kevinzhang@bit.edu.cn

@ICTCLAS张华平博士

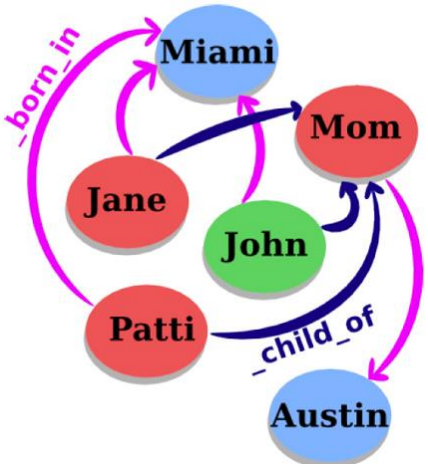
2016.12

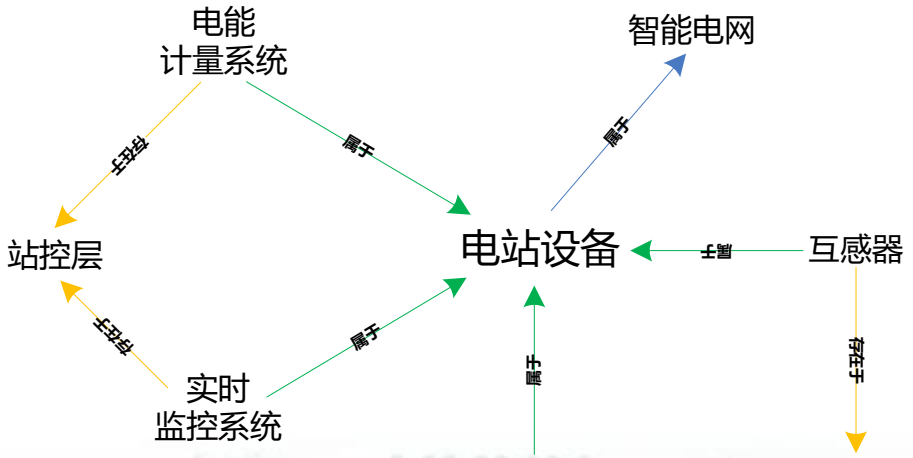
语义网/知识图谱概述


语义分析
句法分析
词法分析
表示学习建立统一的
语义表示空间
frog
cat
bird
dog
truck
automobile
deer
horse
ship
airplane
知识
句子
短语
词汇

语义网/知识图谱概述

- **知识图谱包括实体与关系**
  - 节点代表实体
  - 连边代表关系
- **事实可以用三元组表示**
  - (head, relation, tail)
- **代表语义网知识库**
  - WordNet: 语言知识
  - Freebase: 世界知识





语义网/知识图谱概述：词 VS 语义本体

语义网/知识图谱概述：词 VS 语义本体

1 三大关键技术的支持：

（1）语法层——XML

（2）资源管理框架——RDF

（3）本体层——Ontology

2 现在&未来：

Logic、Proof、Trust

->国内:HowNet（1999）、CCD以BiFrameNet

->国外:WordNet、FrameNet、Roget's Thesaurus

语义网现状：学术

KG	Year	Resources	#Langs	#Entities	#Facts
DBpedia	2007	Wikipedia	126	38.3 M	3 B
BabelNet	2012	Wikipedia, WordNet	271	13 M	2 B
WikiData	2012	Wikipedia, User edits	287	14 M	--
YAGO3	2015	Wikipedia, WordNet, WikiData	10	4.6 M	8.9 M

语义网现状：工业

Google

超过5.7亿实体
超过18亿条事实（关系）

ProBase

2,653,873概念

百度知心

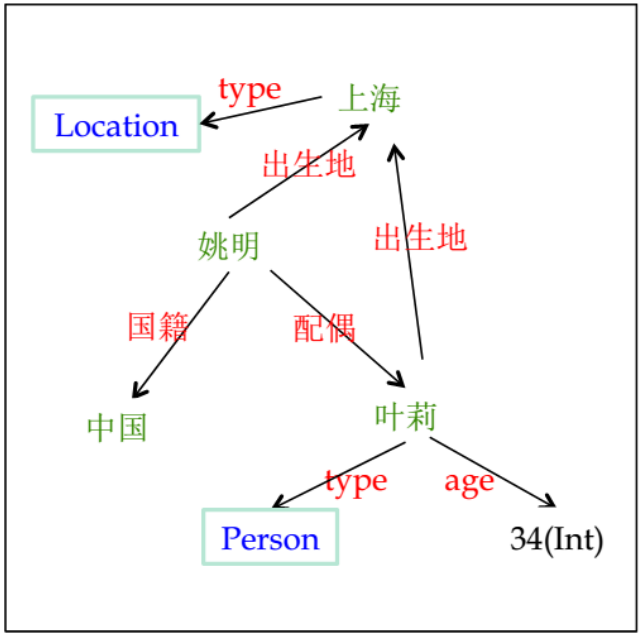
搜狗知立方

姚明的老婆出生在哪里?

语义解析

查询

```
SELECT DISTINCT ?x
WHERE {
      ?y 出生地 ?x.
      res:姚明 配偶 ?y.
}
```

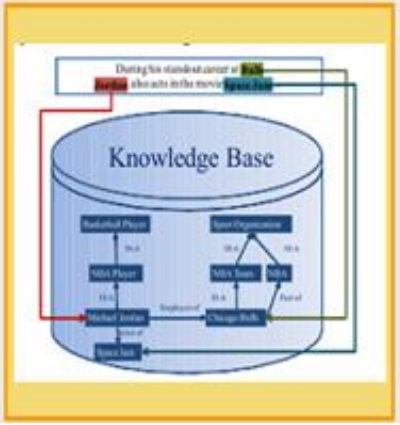


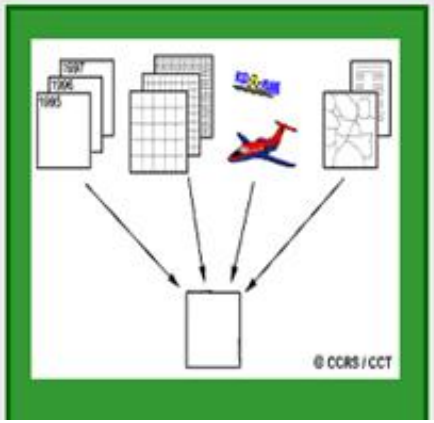
语义网/知识图谱概述：问题与挑战

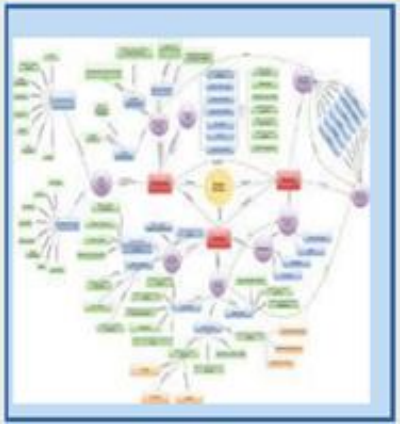
1. 完备性？如何应对知识增长？本体库扩建？
2. 推理计算的准确性与效率：好看不中用？
3. 构建与维护成本：人工密集型？

语义网自动构建系统流程

概念发现 → 关联计算 → 关系抽取 → 集成验证







语义网概念发现-Step1: 格式解析

针对PDF、Word、XML等主流文档，采用我们的信息抽取组件，抽取出结构化的文本信息


第五章 其他变配电设备运行与异常处理
（一）防误操作装置的操作使用
1．进入系统
双击桌面系统图标，选择相应用户和口令，即可登录系统，点击鼠标右键自动弹出主菜单。
2．设备对应
对位用于整定设备状态使图形中设备状态与现场情况一致，图形中设备状态与现场情况一致时可略过。
（1）自动对位。“五防”系统与监控系统设有通信接口，“五防”系统接收后台通信并实现实时自动对位。后台无法采集位置信号的设备一般通过钥匙记忆回传对位，经钥匙回传后仍无法正常对位的需要进行手工对位。
（2）手工对位。点击图形中的手形图标，按现场实际情况进行对位，对位完成后再选择退出。
3．模拟预演及电脑钥匙操作步骤
模拟分为两种情况：自由模拟开票（此种模式是按照满足“五防”逻辑条件的序列进行非特定间隔的模拟预演）和典型票开票模拟（是指严格按照已有的操作票序列进行模拟预演）。
（1）自由模拟开票。
1）点一次图形中的手形图标，进入模拟状态，同时弹出任务及步骤窗口。
2）按照操作票顺序，在图形中点相应设备的图标。
正确操作项出现手指图标，同时发出“叮”声并自动改变设备状态，错误操作项发出语音报警，提示错误旅行；同时弹出对话框提示错误旅行及错误所在，不能继续操作，点击对话框中的正确取消该步操作，方可继续进行后续步骤。
3）结束后再次选择图形中手形图标完成模拟过程。
（2）典型票开票模拟。
1）一次图形中的文件进入操作票选择界面，左侧的窗口显示已审核票以及未审核票的操作任务目录，右侧显示已选择任务的操作票内容及相关序列。
2）在左侧的窗口选择要模拟的操作任务，并选择开票图标。
3）系统弹出选择操作窗口，随即弹出当前任务窗口，同时窗口下方显示“下一步”，此时必须严格按照提示进行操作，否则将提示操作步骤不对，再次错误则弹出正确操作：*****的提示项。
4）模拟全部结束后，弹出窗口模拟操作结束，选择手形完成模拟过程。
4．传送
传送向电脑钥匙传送操作票。

PDF格式内容：7.10 GB
文本大小：159.34MB
总词数：21,545,528

电力语义网概念发现-Step2: 分词标注

NLPIR-ICTCLAS分词系统可以融合已有本体库，实现专业领域的分词标注，15年历史，成功应用于华为、人民网、中国邮政、央行、中央网信办。

语义网概念发现-Step3: 新词发现

演示地址： http://ictclas.nlpir.org/nlpir/#box-1 右边的新词发现

NLPIR-ICTCLAS新词发现可直接从原始语料中直接发现新词、新概念。

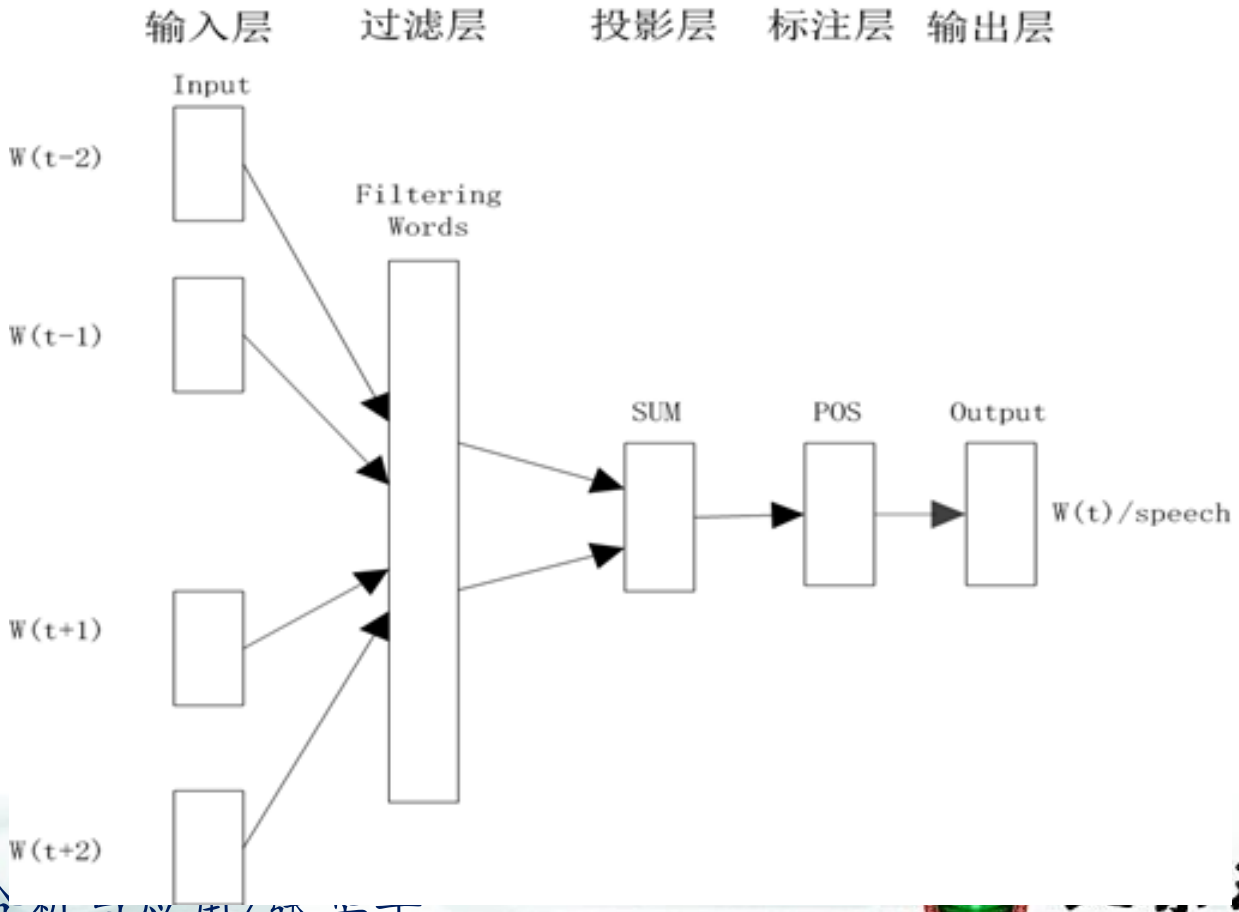
语义网概念发现-Step4: 概念发现

NLPIR-ICTCLAS从新词中过滤筛选本体概念

数据格式	PDF	PDF（加入专业词）	XML（加入专业词）	备注	
文件数量	351个	351个	351个	提供总文件个数	IN
总大小	约7.1GB	约7.1GB	约185M	文件本身大小	
语料集大小	约160M	约237M	约125M	萃取、清洗、归档后的文本大小	
专业词（英大提供）	无	1865个	1865个	现有的专业词个数	
专业词（包含新词）	4000个	5865个	5865个	发现新词与已有的专业词的总数	
基于词性的连续词袋模型(POS-CBOW模型)	约81.8M	约62M	约30M	生成的模型文件的大小	TOOL
词数	70180个	53013个	25379个	最终得到的词的数量	OUT

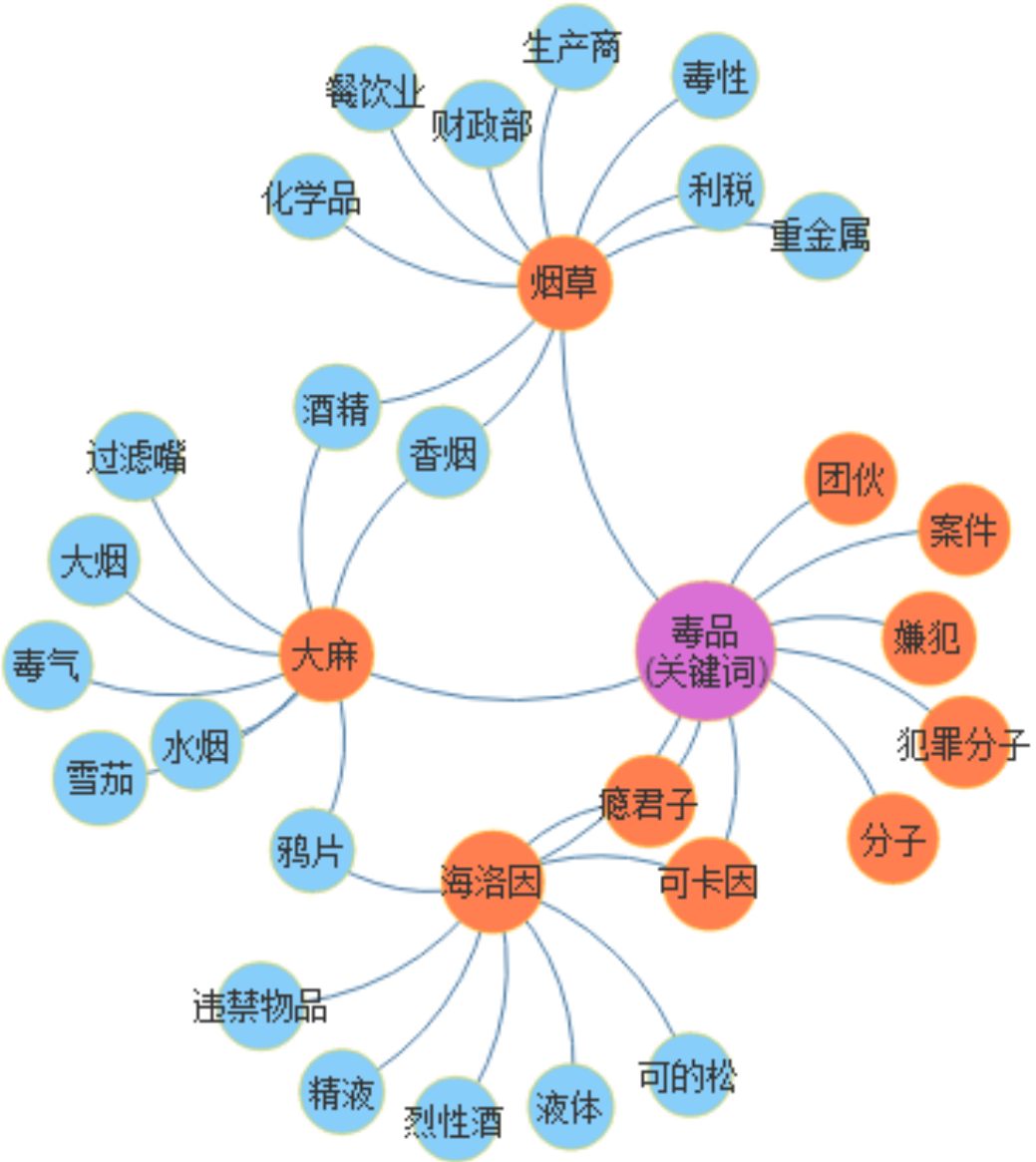
语义关联关系计算- POS-CBOW深度计算

POS-CBOW语言模型：加入新概念后的五层神经网络



语义关联关系计算- POS-CBOW深度计算

演示地址： http://ictclas.nlpir.org/nlpir/#box-7



语义网关系抽取- 短语抽取

- 通过识别表达语义关系的短语来抽取实体之间的关系
  - *(华为，总部位于，深圳)，(华为，总部设置于，深圳)，(华为，将其总部建于，深圳)*
- 同时使用句法和统计数据来过滤抽取出来的三元组
  - 关系短语应当是一个以动词为核心的短语
  - 关系短语应当匹配多个不同实体对
- 优点：无需预先定义关系类别
- 缺点：语义没有归一化，同一关系有不同表示

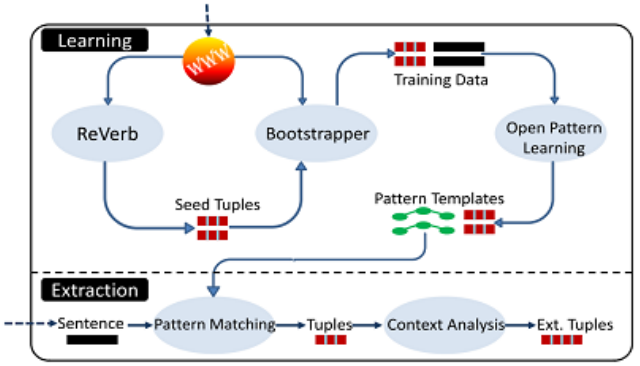
$$V \mid VP \mid VW^*P$$

V = verb particle? adv?

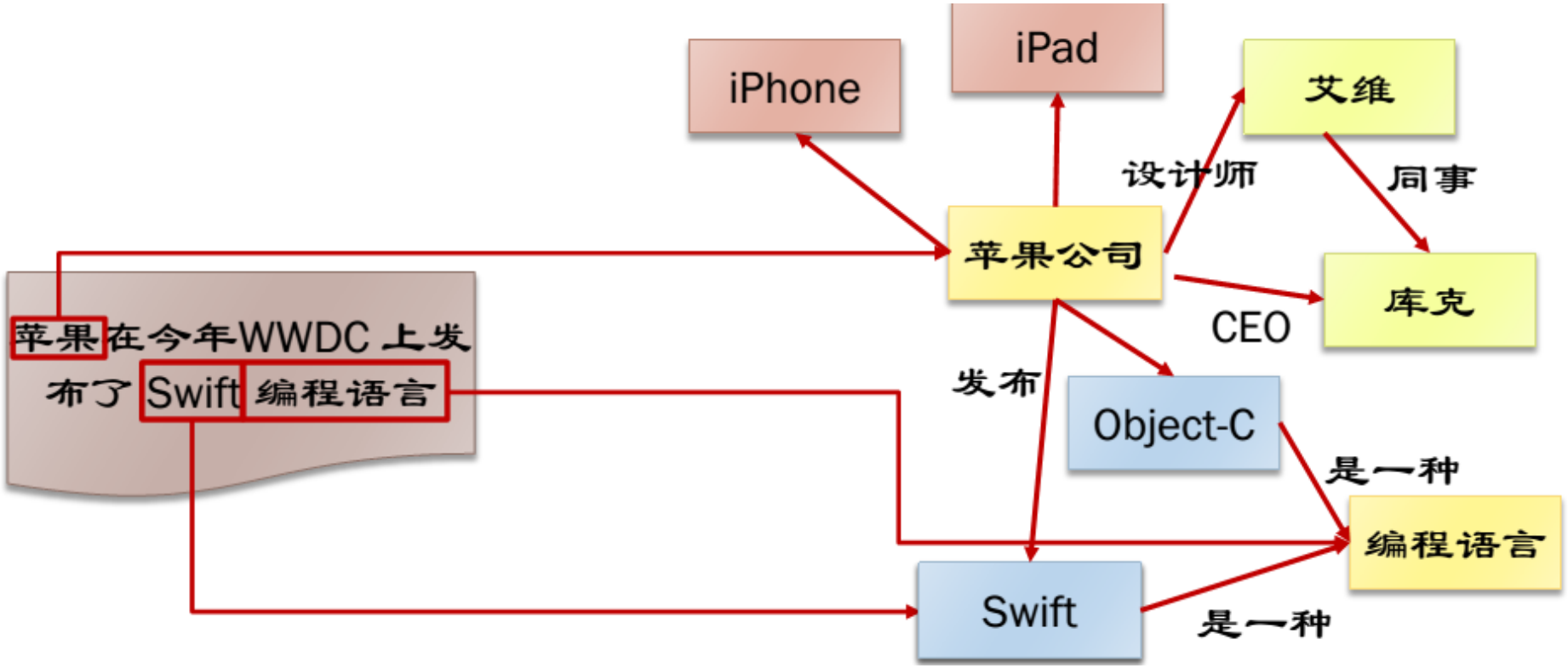
W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

关系短语的句法结构约束



语义网关系抽取- 依存句法分析



语义网关系抽取- 篇章关系抽取

语义网关系抽取- 篇章关系抽取

语义网关系抽取- 篇章关系抽取

语义网集成验证-OWL+ Protégé

1. OWL:网络本体语言,W3C开发的一种网络本体语言，用于对本体进行语义描述。

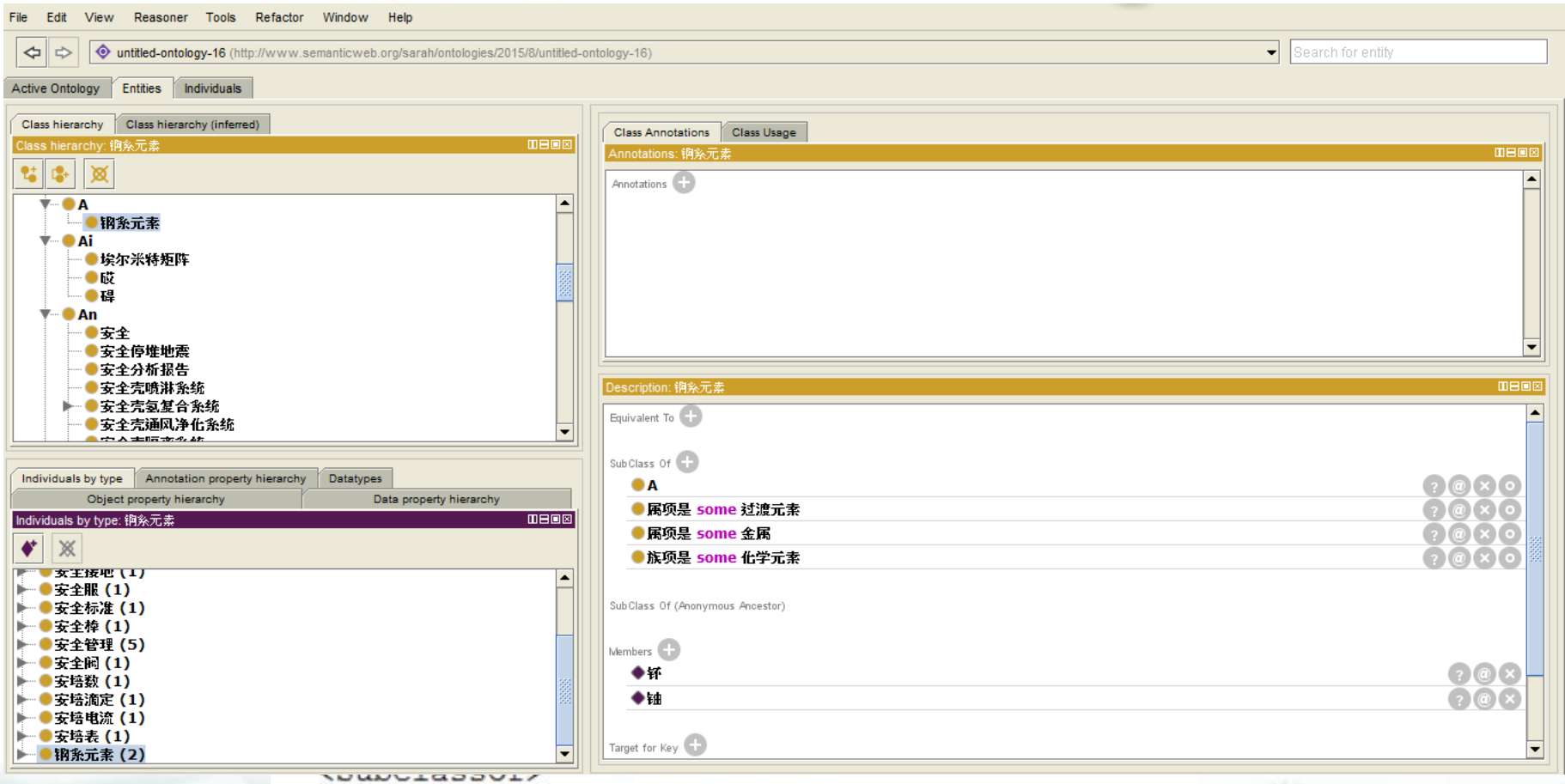
类(Class)、个体(Individual)、属性(Property)

2. Protégé: 斯坦福大学基于Java语言开发的本体编辑和知识获取软件，是语义网中本体构建的核心开发工具，现在的最新版本为5.0 Beta版本。

可视化界面+API调用开发

语义网集成验证-OWL+ Protégé

```
<Declaration>
    <AnnotationProperty IRI="#卤素"/>
</Declaration>
```




```
    <Class IRI="#化学元素"/>
    <Class IRI="#族系"/>
</SubClassOf>
```

方案总结：我们的优势与特色

1. 大数据技术：从原始语料中自动生成语义网
2. 深度计算：图计算与向量计算，便于推理
3. 投入产出：自动化技术为主，人工为辅

JZSearch语义精准搜索引擎

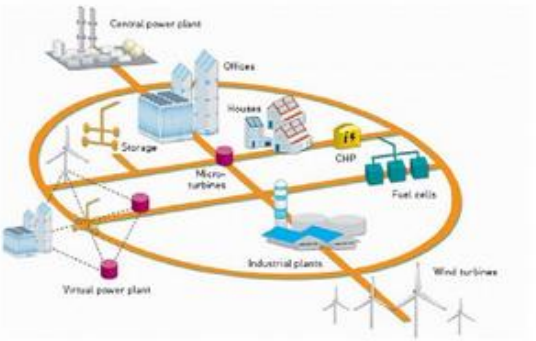
个人中心

国家电网公司 STATE GRID CORPORATION OF CHINA

综合搜索 组织结构 电网报 图片 电力地图 统计分析

刘振亚是谁 搜索

热搜词条： 刘振亚是谁 UHV是什么 智能电网 配电网 全球能源互联网



最新消息 国网大事记

- 严格依法治企 坚持以...
- 台风"灿鸿"肆虐 供...
- 奋发有为 扎实工作 ...
- 连续抢修3天
- 风狂雨骤显担当
- 一定把损失降到最低

JZSearch语义精准搜索引擎

聚类结果

特高压

教育实践活动

智能电网

公司

全球能源互联网

学习实践活动

经济社会发展

安全生产

语义结果

询问对象：刘振亚，

刘振亚，男，汉族，1952年8月生，山东郯城人，1984年加入中国共产党，1971年参加工作，山东工学院电力系电力系统及自动化专业毕业，大学学历，山东大学电气工程学院电气工程及其自动化专业硕士研究生毕业，电气工程硕士，教授级高级工程师，享受国务院政府特殊津贴。现任国家电网公司董事长、党组书记。

语义统计分析

分析对象：刘振亚 分析

智能搜索

返回检索结果约1906个结果...

1. 刘振亚总经理调研韩国STX集团大连造船厂项目的用电情况

来自栏目： 无设定栏目 **板块：** 要闻 **发布时间：** 2007/05/29 00:00:00 **作者：** 杜平

27日，刘振亚总经理(前排左四)一行在大连市市长夏德仁(前排左三)的陪同下，调研韩国STX集团大连造船厂项目的用电情况。

关键词：刘振亚　　命名实体--人物：刘振亚#夏德仁#

2. 刘振亚分别会见花旗 集团和通用电气高层

来自栏目： 无设定栏目 **板块：** 要闻 **发布时间：** 2009/04/28 00:00:00 **作者：** 岳文

语义自动计算

相关新概念发现

沙捞　公司党组　电力集团公司　能源互联网

相关人物聚类

习近平/61　李克强/22　舒印彪/17　李荣融/12　郑宝森/11　温家宝/10　俞正声/9　曹志安/9　帅军庆/7　高培/7

相关作者聚类

姚雷/295　陶思遥/108　张超义/54　江莹/41

JZSearch语义精准搜索引擎

语义统计分析

语义统计分析

分析对象：刘振亚 分析

分析对象：刘振亚是谁

时间：2007-2015

分析机构：国家电网公司

数据来源：10年国家电网报

说明：左图为关键词随着时间的变化，关键词也在发生变化。

2007 2008 2009 2010 2011 2012 2013 2014 2015 >>>

JZSearch语义精准搜索引擎

3. 刘振亚会见马来西亚沙捞越州首席部长

来自栏目： 无设定栏目 **板块：** 要闻 **发布时间：** 2010/04/14 00:00:00 **作者：** 姚雷

总经理刘振亚在公司总部会见了到访的马来西亚沙捞越州首席部长泰益玛目一行，双方进行了亲切友好的交谈，并就发挥各自优势加强合作深入交换了意见。 刘振亚对马来西亚客人的来访表示热烈欢迎，并详细介绍了中国电力工业尤其是电网的发展情况。他希望双方能进一步增进了解和信任，达成合作共识，并在更广泛领域积极开...

关键词：刘振亚 命名实体--人物：刘振亚#哈吉·拉旺#杜至刚#

4. 刘振亚会见东方电气集团董事长

来自栏目： 头条2 **板块：** 要闻 **发布时间：** 2009/07/16 00:00:00 **作者：** 姚雷

刘振亚对王计一行的到访表示热烈欢迎，对于东方电气广大干部员工在特大地震灾难面前顽强不屈的精神和恢复重建取得的成绩表示钦佩。刘振亚说，东方电气集团公司是我国最大的发电设备制造企业之一，国家电网公司一直十分关注东方电气集团灾后重建和发展，帮助东方电气集团恢复重建是国家电网的社会责任所在。经过大...

关键词：刘振亚 命名实体--人物：刘振亚#栾军#

5. 刘振亚会见法国电力集团公司董事长

来自栏目： 无设定栏目 **板块：** 要闻 **发布时间：** 2010/12/09 00:00:00 **作者：** 姚雷

总经理刘振亚在公司总部会见了到访的法国电力集团公司董事长兼首席执行官亨利·普格里奥一行。双方表示，将巩固合作基础，加强交流，进一步拓展双方在电力领域的合作，实现未来共同进步。 刘振亚对亨利·普格里奥一行的来访表示热烈欢迎。他说，国家电网公司非常重视与法国电力同行的友好关系，双方已有的合作是双赢...

相关概念词发现

基于语义的自动学习计算

相关人物计算

基于语义的自动学习计算

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会