A Linked Data Generation Method for Academic Conference Websites

Peng Wang, Mingqi Zhou, Xiang Zhang, Fengbo Zhou
汪鹏， 周铭琪，张祥，周凤波

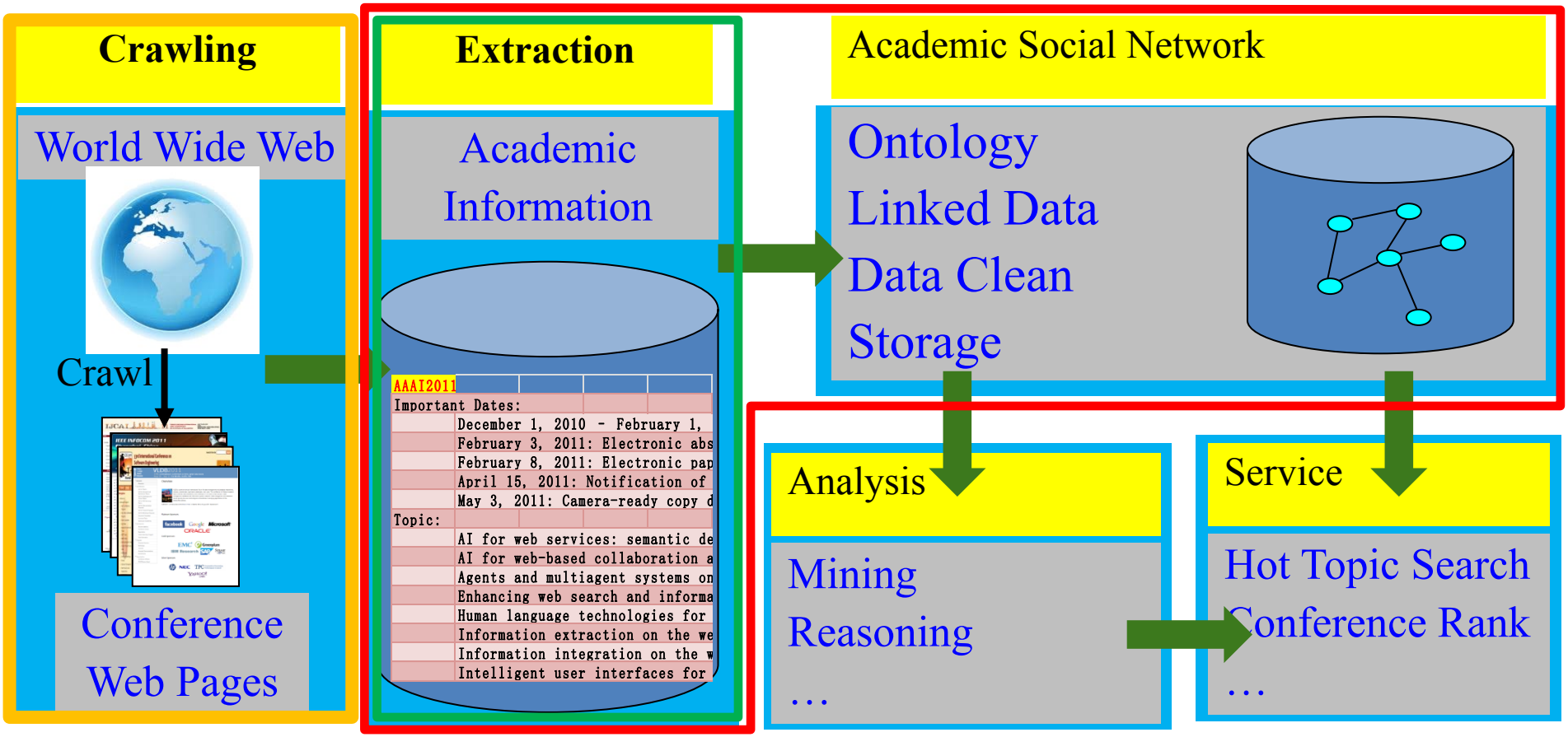
Southeast University
东南大学

报告提纲

- 1. 研究背景、目标和成果
- 2. 主要工作
  - Web信息抽取
  - 本体生成
  - Linked Data生成
- 3. 实验结果
- 4. 总结与展望

研究背景及目标

项目背景：学术社交网络的发现、构造和分析



研究背景及目标

研究目标

从爬取到的大量学术会议网站自动提取学术信息并生成Linked Data数据，该数据是构造学术社交网络的基础数据。

主要挑战

1. Web信息抽取：网页表述的多样性
2. 本体生成：半结构化数据构造语义数据
3. Linked Data生成：本体匹配，数据清洗

相关工作

维基百科:DBpedia　　谷歌:Knowledge Graph

微软:Entity Graph　　语义Web:LOD

本文主要贡献

■一、信息抽取

1. 提出一种改进的基于视觉的Web页面文本块划分算法，取得98.5%的划分精度

2. 利用视觉和文本特征，使用贝叶斯网络自动对文本块分类，取得96%精度和98%召回率

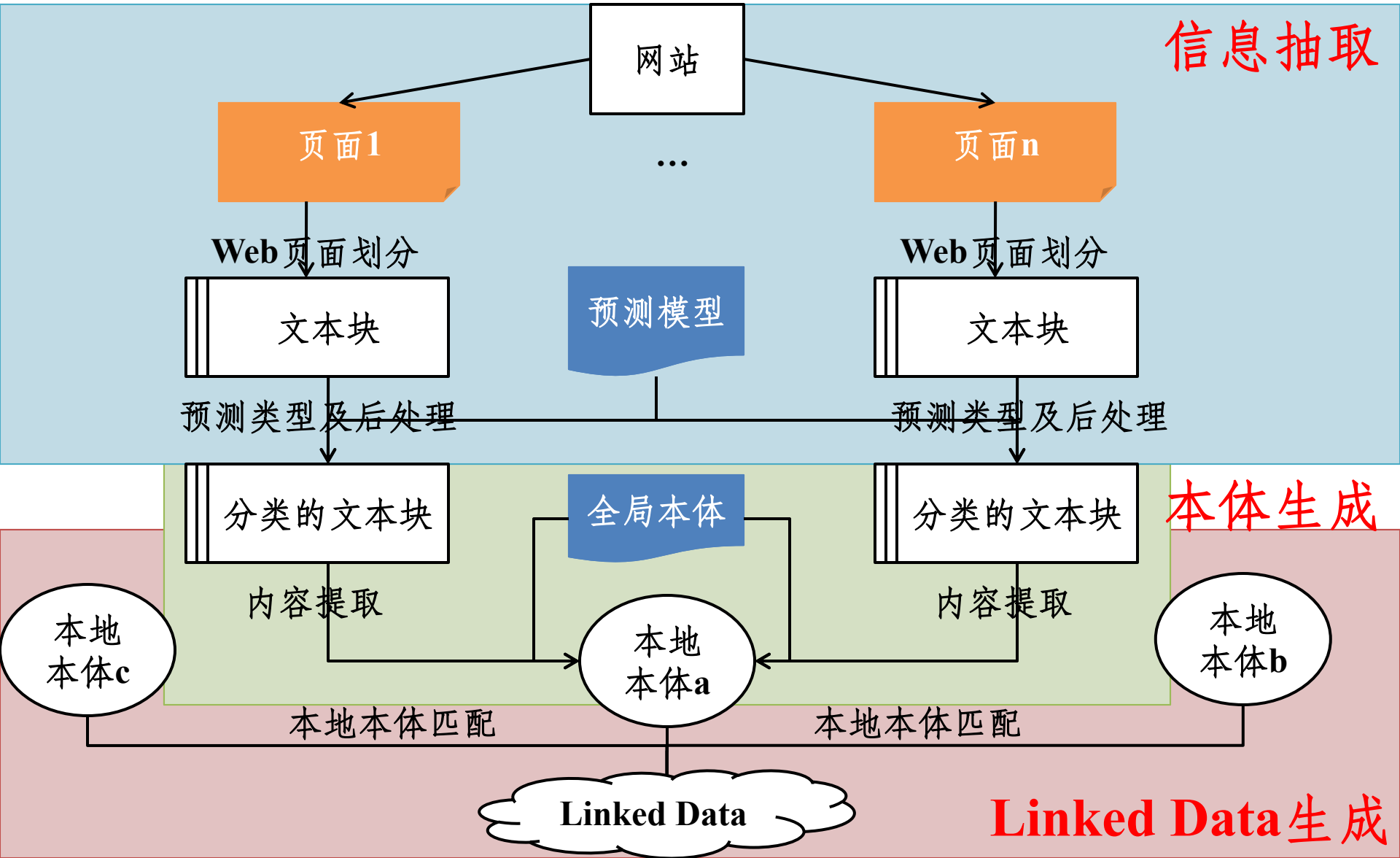
■二、本体生成

3. 基于分类后的文本块生成本体，本体元素取得97%精度和99%召回率

■三、Linked Data生成

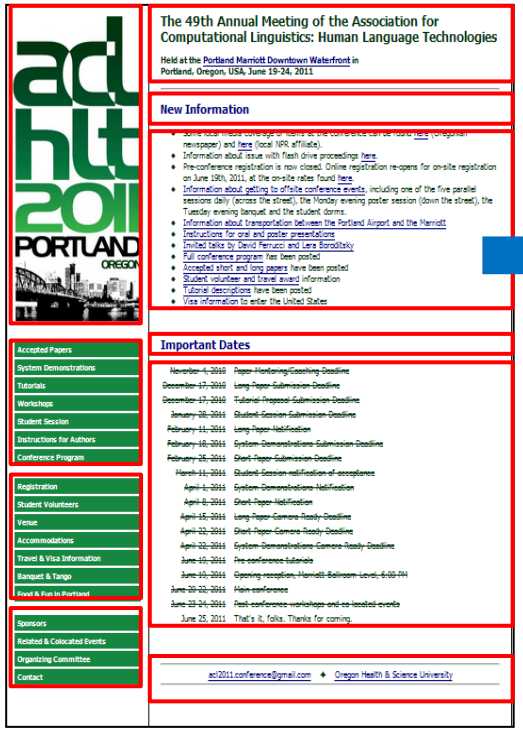
4. 利用本体匹配技术自动将3109个会议集成为学术Linked Data

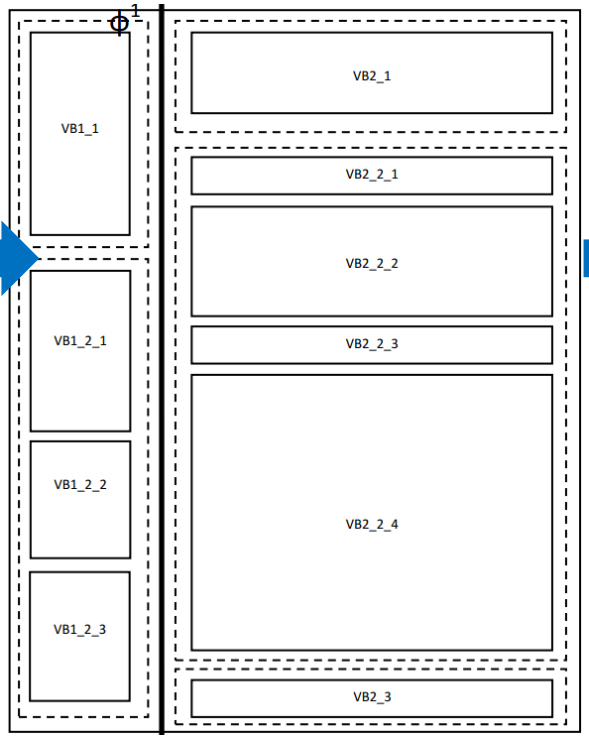
研究路线

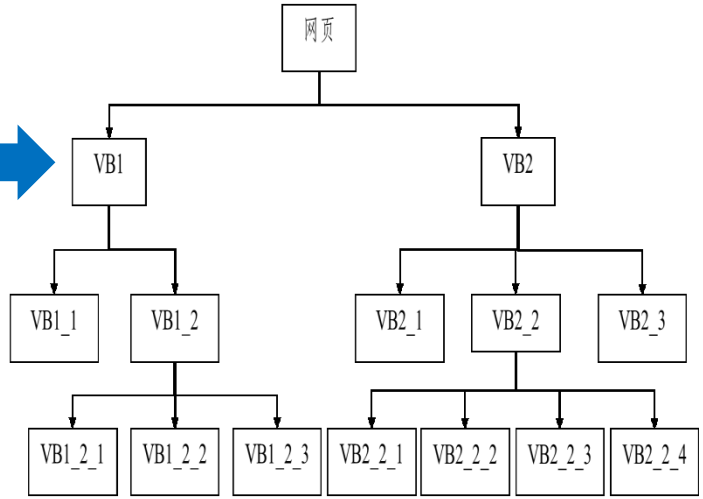

信息抽取
网站
页面1
…
页面n
Web页面划分
Web页面划分
文本块
预测模型
文本块
预测类型及后处理
预测类型及后处理
本体生成
分类的文本块
全局本体
分类的文本块
内容提取
内容提取
本地
本体c
本地
本体a
本地
本体b
本地本体匹配
本地本体匹配
Linked Data
Linked Data生成

Web页面划分——VIPS算法

VIPS算法：基于视觉的网页分块算法 (Deng Cai, 2003)







Web页面划分——VIPS算法缺点

遗漏大量
重要信息

Important Dates

Paper submission deadline:

- Submissions for papers will be admitted for review as a batch on the 1st day of each month at 5:00 AM Pacific Time, April 2010 – March 2011. US daylight savings time will be observed.
- The last submission date for 2011 is March 1, 2011.
- Papers submitted after Jan. 1, 2011 will not receive revision requests.

Workshop proposal deadline:
January 31, 2011 (11:59pm PST)

Industrial track:
Paper submission deadline: March 31, 2011 (10:00pm PDT)
Author notification: May 19, 2011

Tutorial submission deadline:
March 31, 2011 (7:00am PDT)

Demonstration track:
Submission deadline: April 1, 2011 (10:00am PDT)
Author notification: May 27, 2011 (5:00pm PDT)

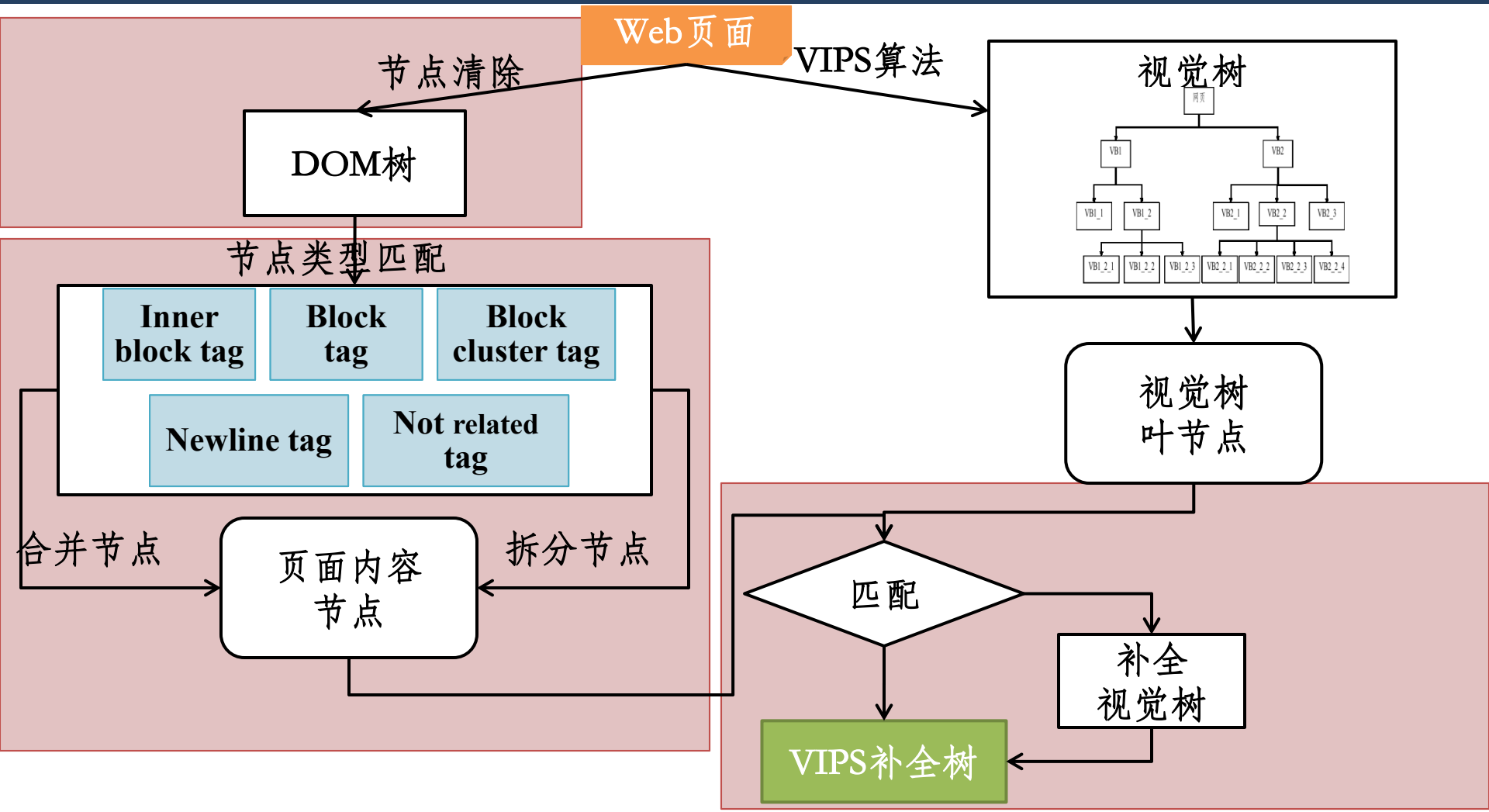
Challenges and Vision track:
Paper submission deadline: April 1, 2011 (11:59pm PDT)
Author notification: May 15, 2011

PhD workshop submission deadline:
April 1, 2011 (11:59pm PDT)

Camera-ready papers due (all tracks):
June 19, 2011

Conference Dates:
August 29 - September 3, 2011

Web页面划分——构建VIPS补全树



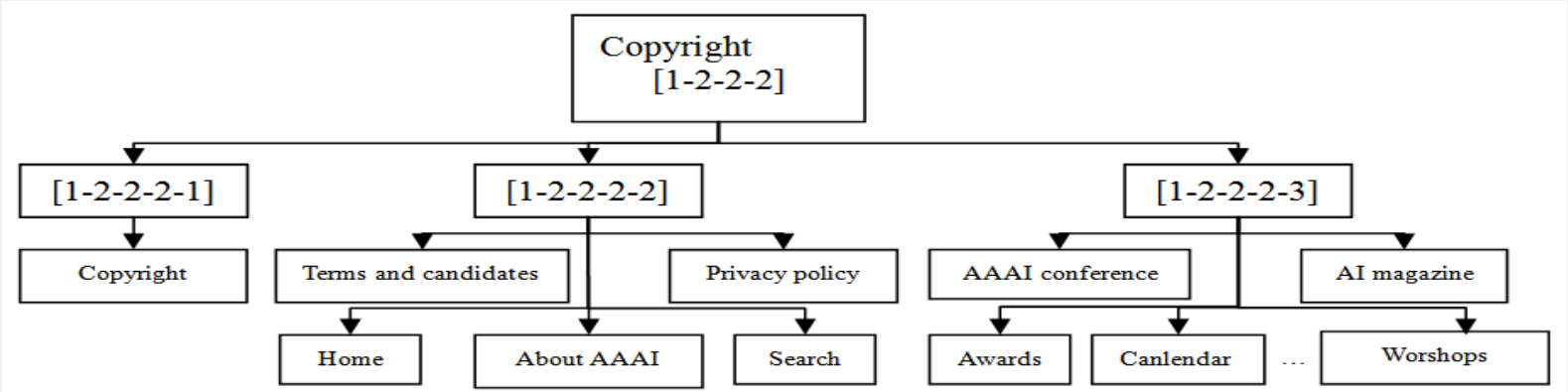
Web页面划分——VIPS补全树例子

This site is protected by copyright and trademark laws under US and International law. All rights reserved. Copyright © 1995–2011 Association for the Advancement of Artificial Intelligence.

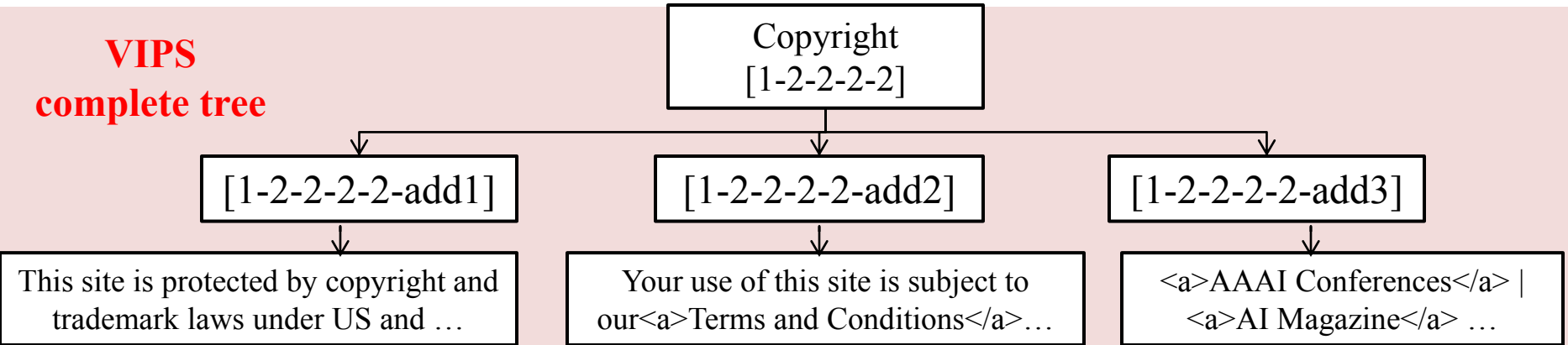
Your use of this site is subject to our Terms and Conditions and Privacy Policy | Home | About AAAI | Search | Contact AAAI

AAAI Conferences | AI Magazine | AI Topics | Awards | Calendar | Digital Library | Jobs | Meetings | Membership | Press | Press Room | Publications | Resources | Symposia | Workshops

VIPS vison tree



VIPS complete tree



Web页面划分——VIPS补全树去噪

- 噪音区域：版权、导航、广告等
- 提出了一种根据特征二维标准去噪方法
  1. 噪音区域的特征：视觉、关键字、文本
  2. 特征的二维标准：影响因素和数量
  3. 两种去噪处理方法：
     分支法(Branch)　权重法(Weight)
- 去噪结果:学术信息的网页VIPS补全树

- 页面划分总结
  1. 将初步的文本抽取转化为页面划分问题进行处理
  2. 获得高质量的学术信息文本块

文本块分类——文本块类型

有效类型

1. DI(dateItem)
2. PI(placeItem)
3. PO(position)
4. PE(peopleItem)
5. AR(area)
6. TO(topic)
7. PA(paper)
8. CO(collection)

9. N(notRelated)
10. R(related)

无效类型

Author notification: May 19, 2011 — 时间信息

Palm Springs, California — 地点信息

Program committee

Berthold Reinwald, IBM Almaden Research Center (chair)
Phil Bohannon, Yahoo! Research (chair)

参与者信息

Call for Papers: Industrial and Applications Track

- Big Data
- Cloud Computing

主题信息

- *DoublePlay: Parallelizing sequential logging and replay* by Kaushik Veeraraghavan, Dongyoon Lee, Benjamin Wester, Jessica Ouyang, Peter Chen, Jason Flinn, Satish Narayanasamy

论文信息

Submissions must be made by the deadline of May 6, 2011, through the website: http://www.easychair.org/conferences/?conf=ccs2011 The review process will be carried out in two phases and authors will have an opportunity to comment on the first-phase reviews. Authors will be notified of the first-phase reviews on Monday, June 20, 2011 and can send back their comments by Thursday, June 23, 2011. [CO]

Become a Sponsor of AAAI-12

Your participation in the AAAI-12 Sponsor Program will give you instant visibility to a diverse group of AI professionals. More...

[R]
[N]

文本块分类——特征向量和分类模型

视觉	关键字	文本
• 标题 • 位置 • 表格 • 列表	• 标题类型 • 时间 • 地点 • 人物 • 机构 • 职位 • 论文	• 文本长度 • 单词数 • 大写单词数 • 超链接长度 • 字体大小 • 字体粗细

- 分类模型

  贝叶斯网络(Bayesian Network)分类器

文本块分类——初步分类结果和上下文信息

- 文本块初步分类结果
  精度:0.75　召回率:0.67　F1:0.6

- 分类改进：引入上下文

上下文信息的发现

- 区域 —— 算法6:GetChunkAreaType 算法7:GetTopicArea 算法8:VerifyChunkAreaType
- 上条记录 —— 算法9:GetPreChunkType 算法10:GetVisuallySame
- 标题列表 —— 算法11:GetTitles

文本块分类——分类后处理

- 分类错误

  问题描述：无效类型识别为有效类型

  解决依据：具有类型X的特征，不是X类型
  是Y类型，不包含Y的特征

- 遗漏分类

  问题描述：有效类型识别为无效类型

  解决依据：视觉、文本、关键字、上下文

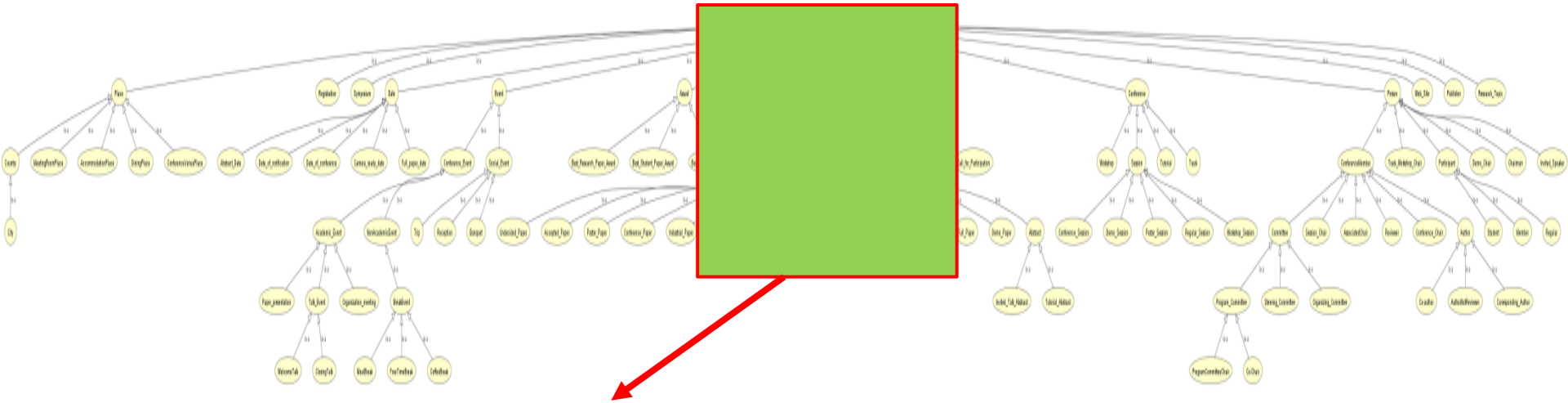
- 分类后处理结果

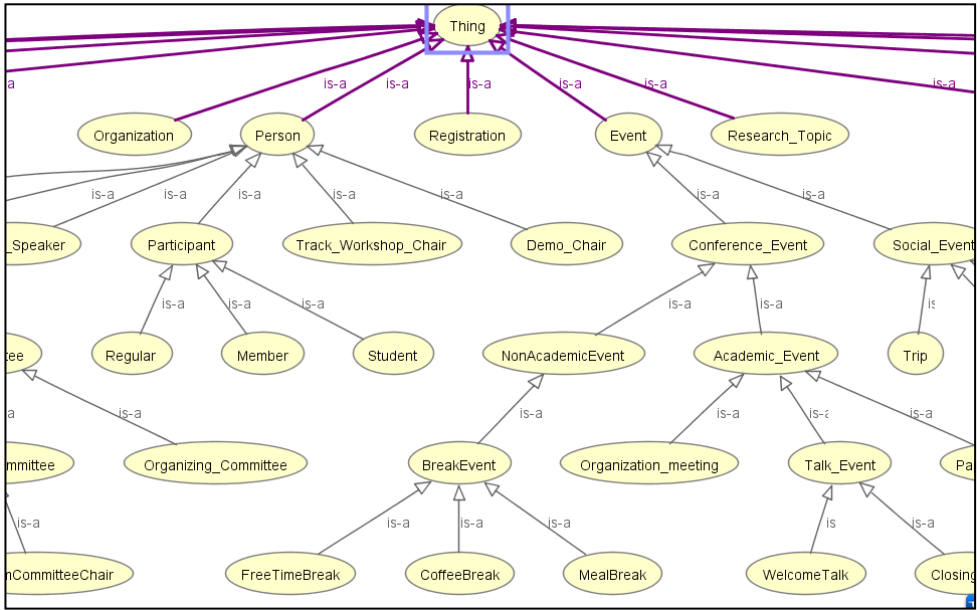
  精度:0.96　召回率:0.98　F1:0.97

- 文本块分类总结

  1. 获得高质量的学术信息文本块分类
  2. 分类后的文本即为抽取的学术信息

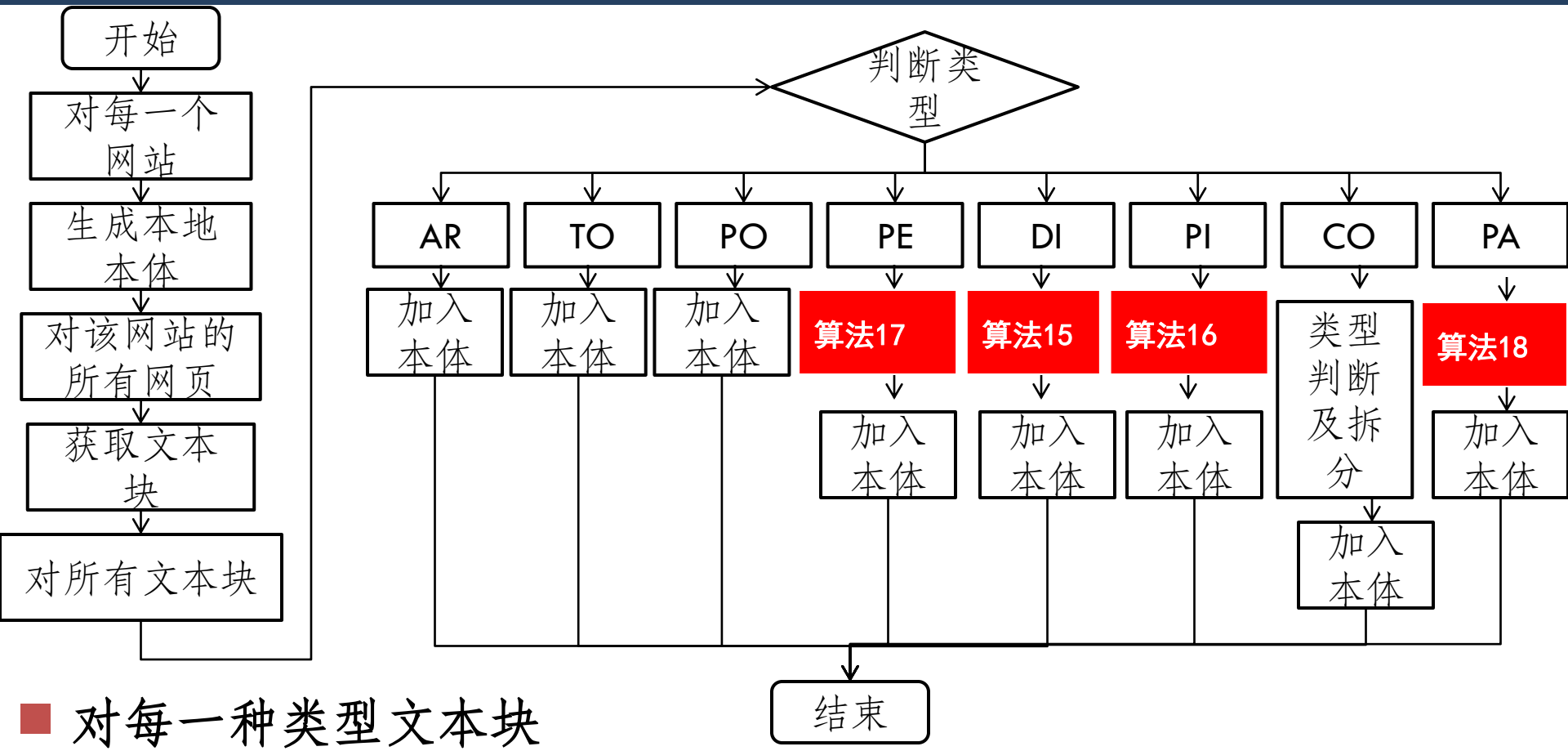
本体生成——全局本体构造





- 构造方式
  人工构造
- 全局本体信息
  97个类 27个属性

本体生成——本地本体生成



- 对每一种类型文本块
- 文本抽取→加入本体
- 平均精度为**0.97**，平均召回率为**0.99**，平均F值为**0.98**

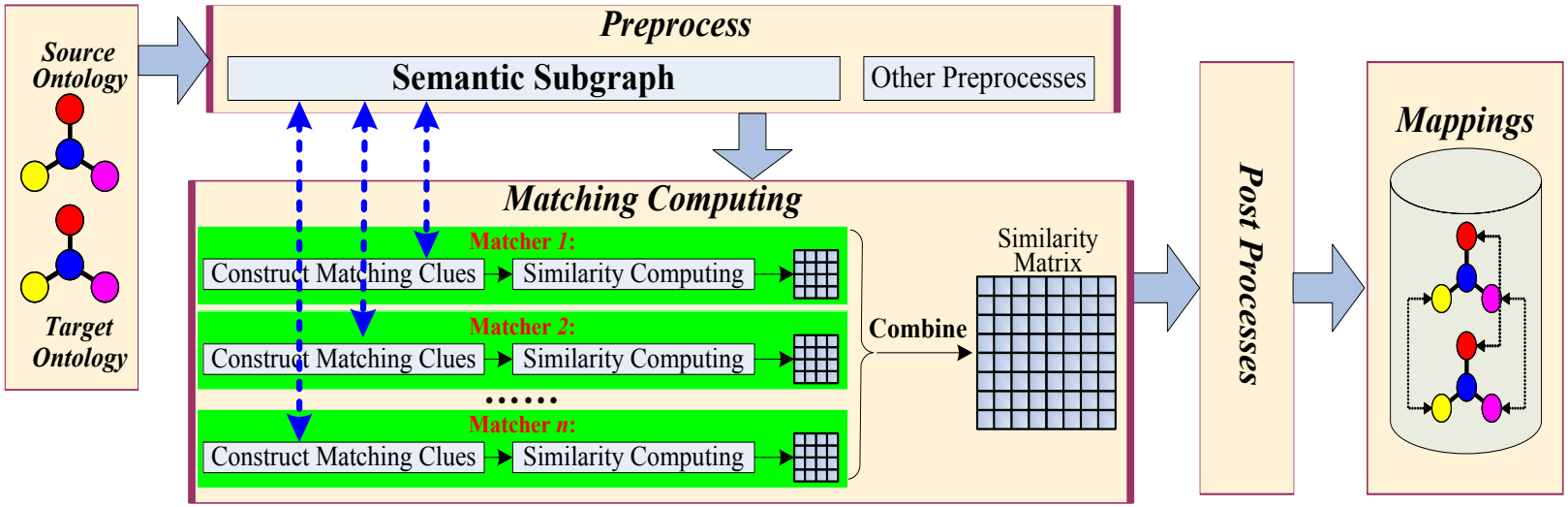
Linked Data数据生成——本体匹配技术

■Liked Data生成技术：本体匹配

通过本体匹配发现本地本体间的关联，从而建立Linked Data

■本体匹配技术：

1. 东南大学研发的通用本体框架工具Lily
2. Lily是本体匹配领域知名的工具，是近年来该领域研究的主要参考工作之一，相关工作得到国际同行的大量引用



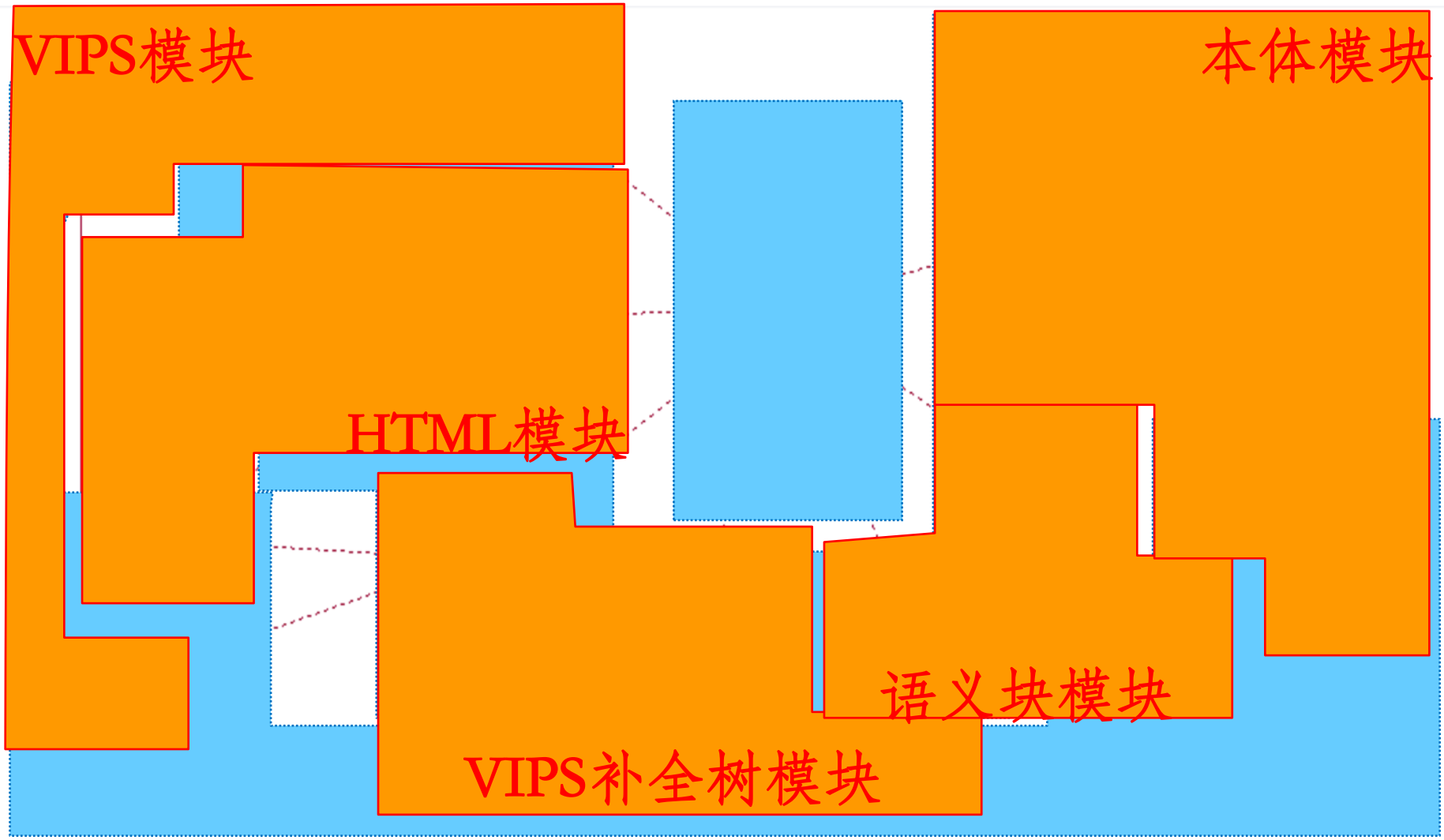
Linked Data数据生成——本体匹配例子

AAAI2011本体	IJCAI2011本体	匹配相似度
Author: Wu-Jun Li	**Author:Wu-Jun Li**	**0.98**
Author: Zhi-hua Zhou	**Author: Zhi-hua Zhou**	**0.98**
Author: Qiang Yang	**Author: Qiang Yang**	**0.90**
Topic: Agent-based and Multiagent Systems	Topic: Agent-based and multiagent systems	0.95
Topic: Natural-Language Processing	Topic: Natural language processing	0.95
Topic: Planning and Scheduling	Topic: Planning and scheduling	0.94
Topic: Knowledge Representation, Reasoning and Logic	Topic: Knowledge representation and reasoning	0.89
Topic: Constraints, Satisfiability, and Search	Topic: Constraint satisfaction, search, and optimization	0.85
Topic: Machine Learning	Topic: Machine learning and data mining	0.80
Topic: Robotics and Vision	Topic: Robotics	0.80
Topic: Web and Knowledge-based Information Systems	Topic: Web and information systems	0.78
Program_Committee: Shaul Markovitch	Program_Committee: Shaul Markovitch	0.98
Program_Committee: Min-Ling Zhang	**Program_Committee: Min-Ling Zhang**	**0.98**
Program_Committee: Jie Tang	**Program_Committee: Jie Tang**	**0.98**
Program_Committee: Jérôme Euzenat	Program_Committee: Jérôme Euzenat	0.98
Program_Committee: Jeff Heflin	Program_Committee: Jeff Heflin	0.98
……	……	……

系统实现——总览

实现语言	**Java**
源程序规模	**5个包，24个类，5100余行代码**
开发平台	**Eclipse 3.5.2**
运行环境	**Windows XP; Java JDK 1.6**
开源工具	**Dom4j, Cobra, Weka, Protege**

系统实现——类、模块关系图

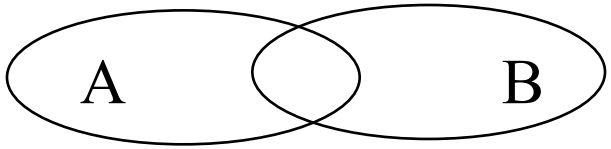


实验数据集准备

- **Data Set:**
  30个2010-2011年计算机科学领域学术会议网站
  136个页面
  3509个有效抽取对象
- **Training:**
  10个随机选取的网站
  62个页面
- **Running Data:**
  20个剩余的会议网站用于系统性能验证

62小时爬获取的3109个学术会议网站作为Linked Data生成数据源

评估标准



设**A**为系统对测试**Web**页面分块文本的自动标注结果，**B**为手工标注的参考结果，则：

$$\Pr ecision = \frac{A \cap}{A}$$

$$\operatorname{Re} call = \frac{A \cap}{B}$$

$$F - Measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}$$

编号	网站	视觉树(VIPS分析结果)		补全树(VIPS+SB分析结果)	
		节点数	叶节点数	节点数	叶节点数
1	AAAI-10	106	82	109	85
2	CIKM-11	109	80	119	90
3	ICDE-10	94	69	101	76
4	ICDM-10	131	98	143	110
5	ICSE-11	145	99	153	107
6	INFOCOM-10	143	93	143	93
7	SIGIR-11	133	86	133	86
8	SIGMOD-10	86	59	86	59
9	SOSP-11	147	101	147	101
10	VLDB-10	117	81	123	87
11	AAAI-11	153	108	157	112
12	ACL-11	178	137	193	152
13	AIDE-11	170	121	174	125
14	ASAP-10	50	31	59	40
15	ASPLOS-11	60	39	76	55
平均补全节点个数					6

VIPS 补全树分析

实验结论

1. 完整地保持VIPS补全树的框架
2. 遗漏文本块能找到正确的位置

编号	网站	去除噪音前		去除噪音后		去除个数/去除前个数	
		节点数	叶节点数	节点数	叶节点数	节点数	叶节点数
1	AAAI-10	109	85	45	30	0.59	0.65
2	CIKM-11	119	90	64	32	0.46	0.64
3	ICDE-10	101	76	46	22	0.54	0.71
4	ICDM-10	143	110	105	70	0.27	0.36
5	ICSE-11	153	107	80	37	0.48	0.65
6	INFOCOM-10	143	93	103	69	0.28	0.26
7	SIGIR-11	133	86	75	42	0.44	0.51
8	SIGMOD-10	86	59	49	27	0.43	0.54
9	SOSP-11	147	101	125	85	0.15	0.16
10	VLDB-10	123	87	42	17	0.66	0.80
11	AAAI-11	157	112	102	68	0.35	0.39
12	ACL-11	193	152	158	110	0.18	0.28
13	AIDE-11	174	125	108	71	0.38	0.43
14	ASAP-10	59	40	46	17	0.22	0.58
15	ASPLOS-11	76	55	39	17	0.49	0.69
平均值						**0.39**	**0.51**

清除噪音结果分析

实验结论

1. 正确识别并处理大部分噪音区域
2. 减少VIPS补全树50%的节点数

编号	网站	去除文本块个数	文本块来自			
			左侧导航区	抬头导航区	版权说明区	主体区
1	AAAI-10	55	42	10	3	0
2	CIKM-11	58	57	0	1	0
3	ICDE-10	54	49	4	1	0
4	ICDM-10	40	40	0	0	0
5	ICSE-11	70	60	9	1	0
6	INFOCOM-10	24	24	0	0	0
7	SIGIR-11	44	43	0	1	0
8	SIGMOD-10	32	32	0	0	0
9	SOSP-11	16	16	0	0	0
10	VLDB-10	70	69	0	1	0
11	AAAI-11	44	32	10	0	2
12	ACL-11	42	36	0	0	6
13	AIDE-11	54	47	4	1	2
14	ASAP-10	23	22	0	1	0
15	ASPLOS-11	38	38	0	0	0
清除噪音节点的Precision						98.5%

清除噪音区域精度分析

实验结论

1. 达到98.5%的精确度

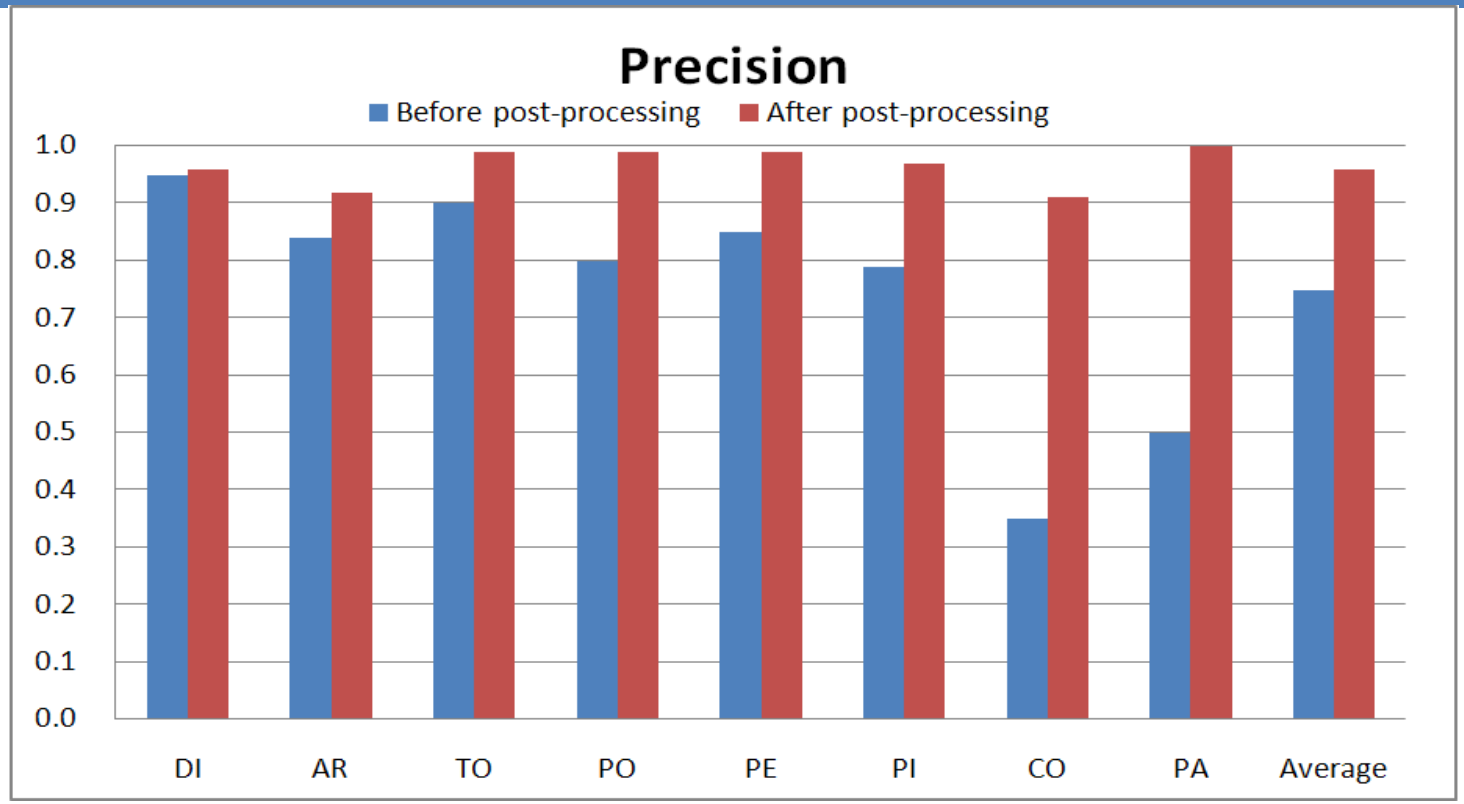
类型	后处理前			后处理后		
	Precision	Recall	F-measure	Precision	Recall	F-measure
DI	0.95	0.75	0.84	0.96	0.99	0.98
AR	0.84	0.91	0.87	0.92	0.98	0.95
TO	0.90	0.41	0.57	0.99	0.99	0.99
PO	0.80	0.69	0.74	0.99	0.98	0.99
PE	0.85	0.60	0.71	0.99	0.99	0.99
PI	0.79	0.72	0.75	0.97	0.97	0.97
CO	0.35	0.80	0.49	0.91	0.95	0.93
PA	0.5	0.5	0.5	1	1	1
平均值	**0.75**	**0.67**	**0.6**	**0.96**	**0.98**	**0.97**

文本块修正结果分析

实验结论

1. **分类结果修正效果明显**
2. **修正后精度提高21%，召回率提高31%，F值提高37%**

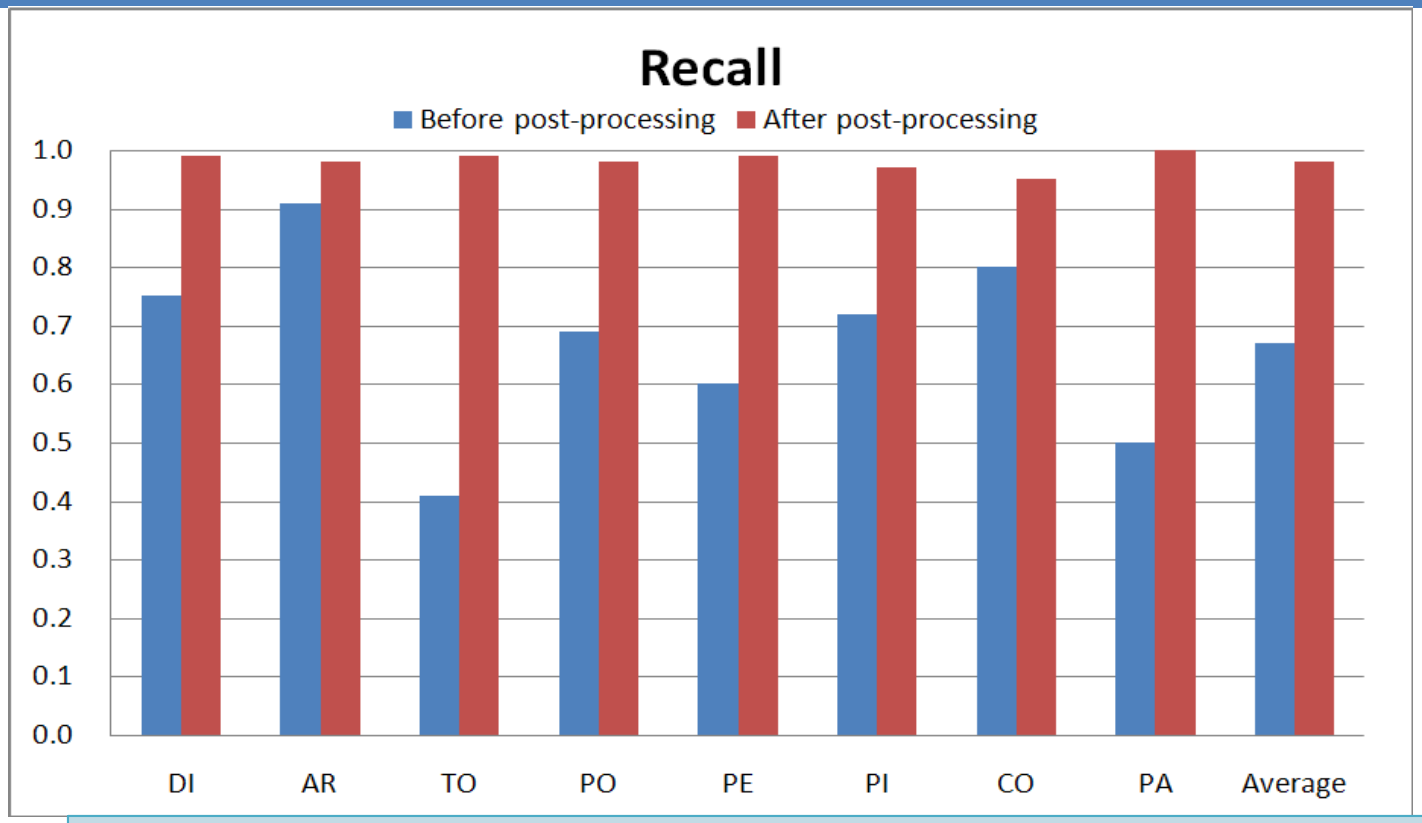
后处理前后分类结果的比较：Precision



实验结论

1. 文本块分类的Precision明显提高
2. CO类型甚至提高2倍以上
3. Precision均大于0.90

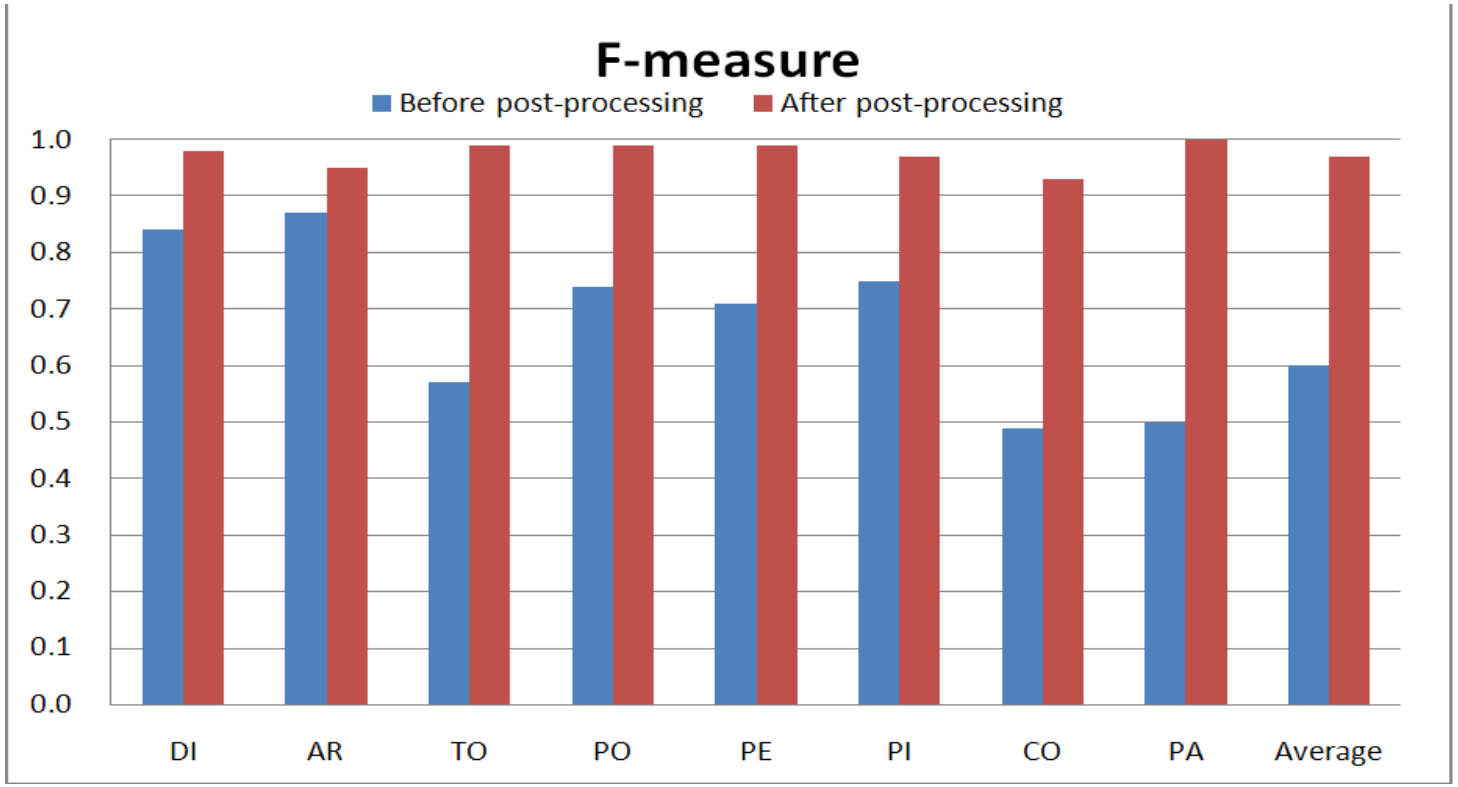
后处理前后分类结果的比较：Recall



实验结论

1. 文本块分类的Recall有大幅提高
2. TO类型提高2倍以上
3. Recall均大于0.95

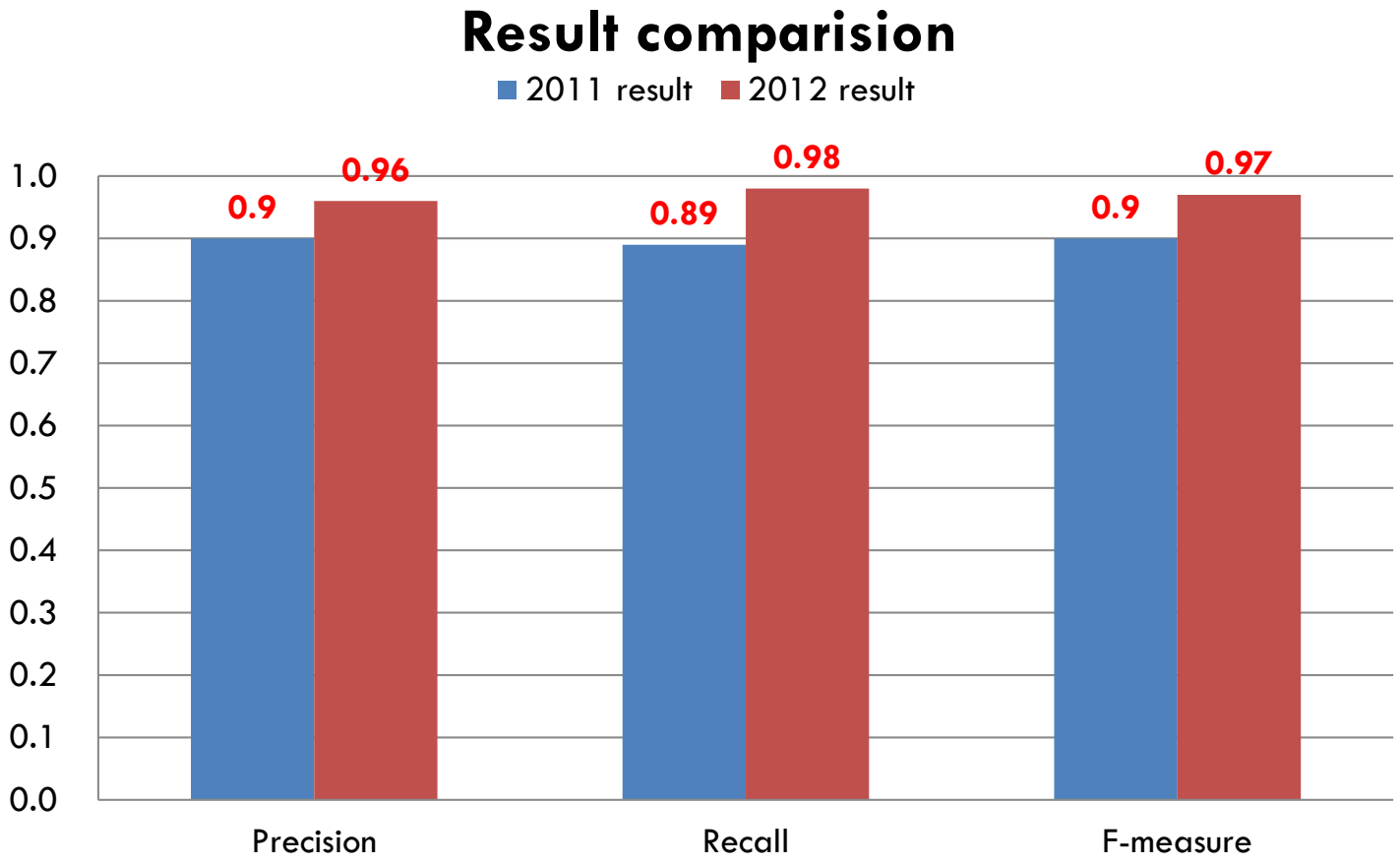
后处理前后分类结果的比较：F-Measure



实验结论

1. 文本块分类的F-measure大幅提高
2. F-measure均大于0.92，有的类型甚至达到0.99

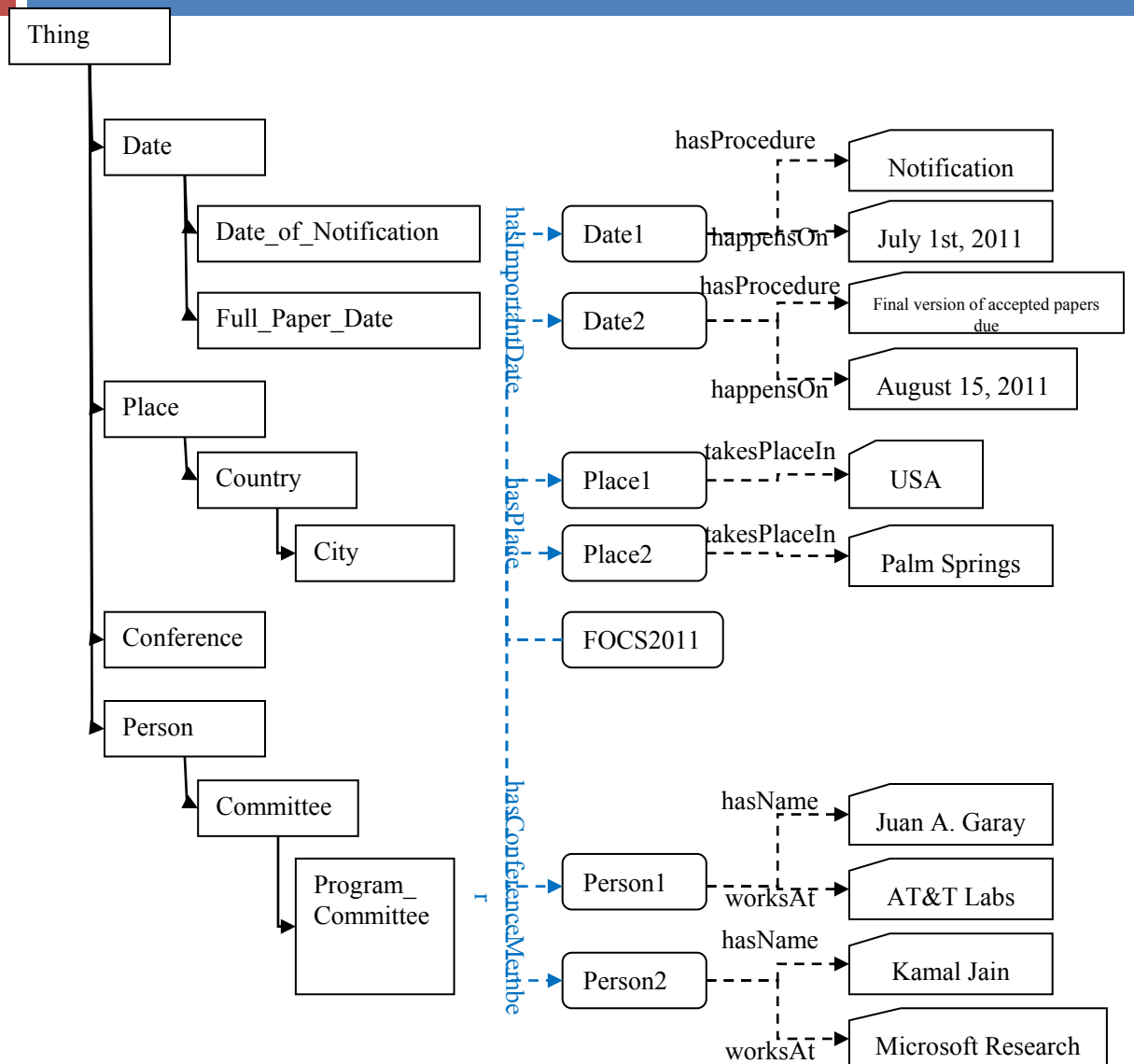
本文结果和2011年结果对比



实验结论

1. 算法较2011年结果有了很大改进
2. Precision, Recall和F-measure的值都在0.95以上

FOCS-2011本地本体构建实例



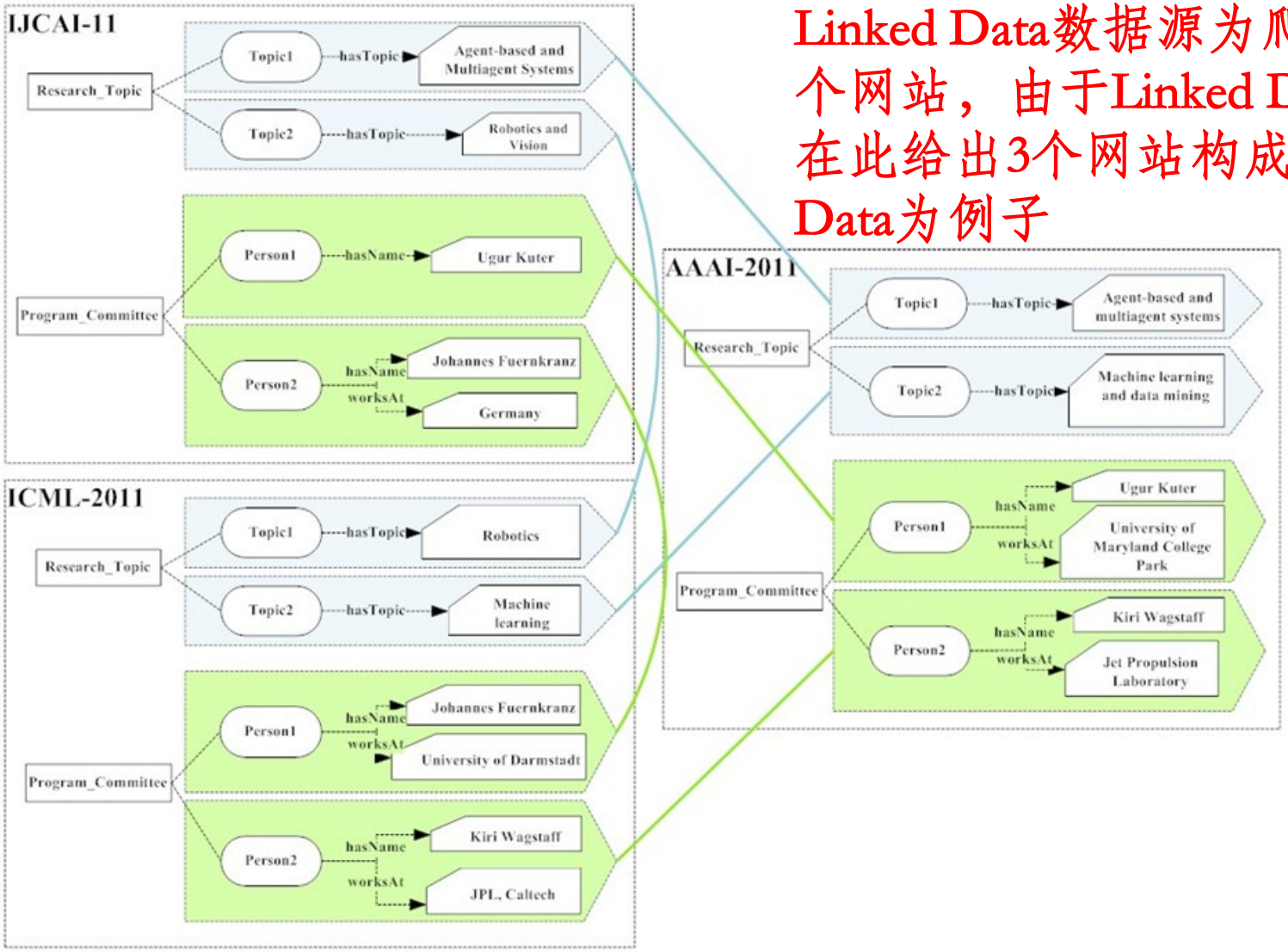
	会议	时间	地点	人物	主题	文档
正确个数	12	141	10	150	126	2
多出个数	1	0	1	1	0	0
缺失个数	0	3	0	9	0	0
Precision	0.92	1	0.91	0.99	1	1
Recall	1	0.98	1	0.94	1	1
F-measure	0.96	0.99	0.95	0.97	1	1

本体生成中对文本块抽取分析

实验结论

1. 平均精度为**0.97**，召回率为**0.99**，**F**值为**0.98**

Linked Data数据实例



Linked Data数据源为爬取出的3109个网站，由于Linked Data非常大，在此给出3个网站构成的局部Linked Data为例子

总结与展望

本文设计并实现了一个面向学术领域的**Linked Data**生成系统：

- Input：　一系列会议网站的页面
- Output：Linked Data

未来的展望：

- **Linked Data**数据质量提高，如共指消解处理
- **Linked Data**如何进一步获得学术社交网络
- 学术社交网络的分析和挖掘

धन्यवाद
Hindi

多謝
Traditional Chinese

ขอบคุณ
Thai

Спасибо
Russian

Gracias
Spanish

Thank You
English

شكراً
Arabic

Obrigado
Brazilian Portuguese

Danke
German

Grazie
Italian

多谢
Simplified Chinese

Merci
French

நன்றி
Tamil

ありがとうございました
Japanese

감사합니다
Korean

Linked Data

- 用语义Web技术组织数据的方式，代表有谷歌的Knowledge Graph
- 其将“网页-链接-网页”或“文档-链接-文档”的传统数据存储方式改为“Data-链接-Data”

本体

- **W3C**对其的定义为：“本体是用来描述与表示各种领域的知识”
- 就是利用本体来架构一个领域知识，并进一步分析此领域中各种概念的关系。
- 为什么使用本体：

Linked Data本身就是由语义**Web**技术构建的，而语义**Web**中的核心就是本体，要获得**Linked Data**，必须先有本体。

本文关注的信息

- 时间信息：记录会议开始时间，结束时间，会议中发生相关事件的时间。
- 地点信息：记录会议的地点。
- 研究信息：记录本次会议注重的方向(**area**)，以及各个方向的主题(**topic**)。
- 参与者信息：包含职位(**speaker, chair…**)，相应的参与者的名字，所在的机构等信息。
- 论文信息：包含了论文的题目，作者。

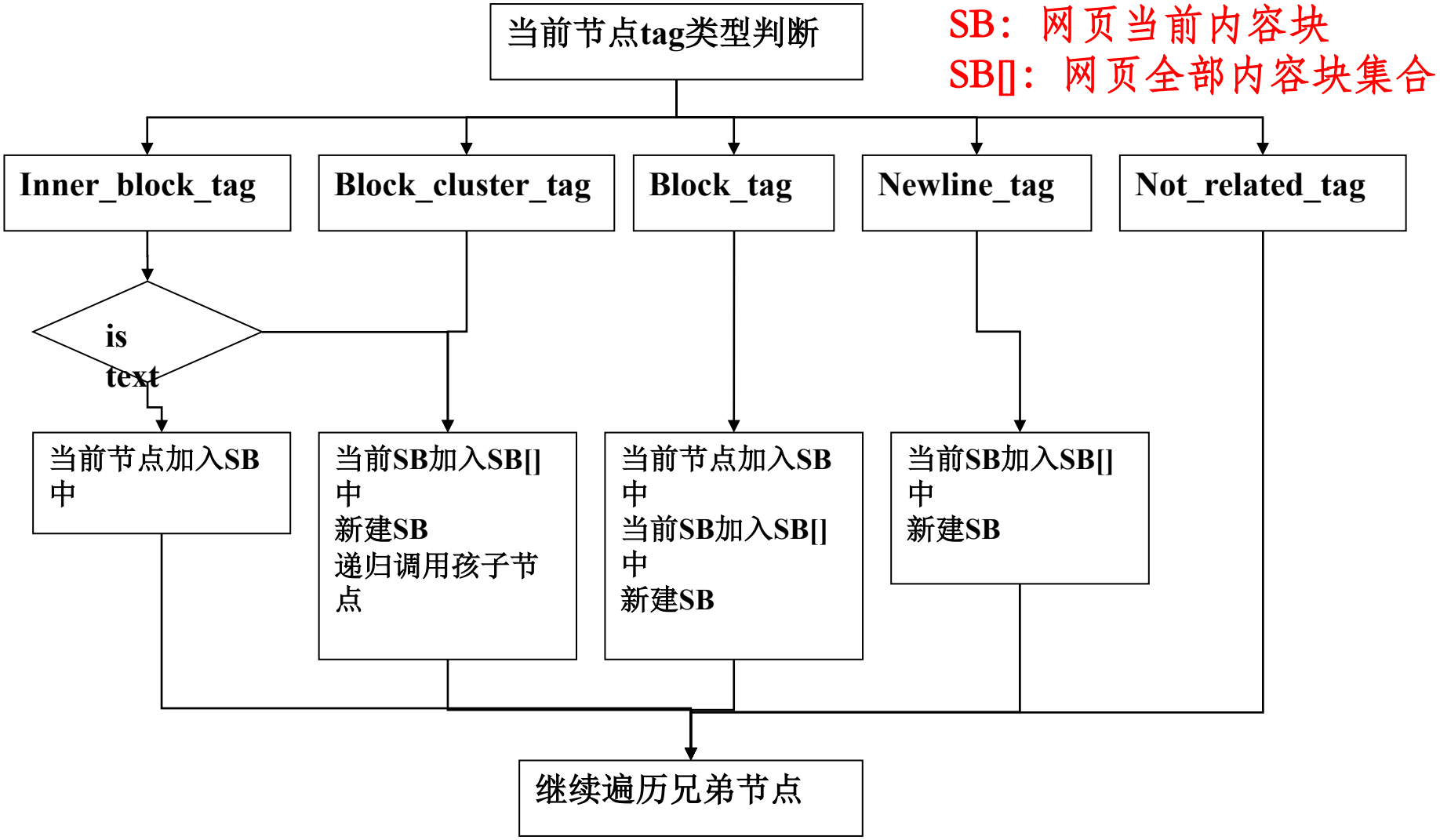
为什么使用VIPS

- VIPS（基于视觉特征的网页分块算法，vision-based page segmentation）相较于其他算法可能会造成信息的疏漏，但是它可以在语义的层面对网页进行分块。VIPS关注了语义块的位置信息，色彩信息，文字视觉信息（如大小，粗细）等信息，并且以ID的形式记录视觉树，对于后续的工作有启发意义，因此我们使用VIPS作为网页结构树的构建依据。

DOM节点分类

Inner block		Block	Block cluster	Newline	Not related		
B	Dfn	Li	Section	Br	Iframe	Colgroup	Isindex
I	Samp	Dd	P	Hr	Frame	Figure	Keygen
U	Acrony	Dt	Article		Frameset	Videa	Output
Em	m	H1	Aside		Noframe	Audio	Meta
Font	Abbr	H2	Div		s	Track	Head
Big	Address	H3	Details		Img	Source	Label
Strong	Q	H4	Span		Ruby	Progress	Fieldset
Small	Cite	H5	Ul		Rt	Meter	Legend
Sup	Ins	H6	Dl		Rp	Canvas	Title
Su	Del	Header	Ol		Link	Command	Base
Pre,	S	Hgroup	Dir		Style	Form	Basefont
Tt	Strike	Footer	Menu		Noscript	Input	Wbr
Kbd	Mark	Center	Table		Object	Textarea	Col
Code	Time	Blockquote	Tbody		Param	Button	Area
Var	Bdi	Caption	Thead		Script	Select	Map
	Bdo	Nav	Tfoot		Applet	Optgroup	Option
Span	A		Th		Embed		
Text			Td				
			Tr				

如何根据DOM节点类型提取内容



如何构建VIPS补全树

算法3：
GetDataNodes

VIPS叶节点LN[]

网页内容节点 SBs[]

- 对SBs[i]，遍历LN[]，寻找匹配
- 若找到，将其作为参数创建DN
- 若没有，为SBs[i]创建LN，加入VIPS视觉树，并创建新的DN

视觉补全树节点DN[]

匹配LN，生成VIPS补全树

算法4：
GenerateStructureTree

噪音特征

- 特征值对结果的影响

是否是噪音区域	包含该特征值	不包含该特征值
特征值1(a1)	可能是	一定不是
特征值2(a2)	一定是	可能不是
特征值3(a3)	可能是	可能不是

- 目标区域含有特征值的数量

  0-1(b1) 0-N(b2)

特征处理方法

- 分支法(Branch)
- 权重法(Weight)
  - **Weight1:**符合条件，加大量权重。
  - **Weight2:**符合条件，加少量权重。
  - **Weight3:**符合条件，加权重，反之，减权重。
- 特征值与处理方法的映射

处理方法	a1	a2	a3
b1		Weight1	Weight2
b2	Branch		Weight3

噪音识别例子

- 例如:
  导航栏有特征[a3,b1]:导航栏宽度/页面宽度 < 0.25
  该特征对应处理方法:Weight2

对于右侧网页
导航栏宽度/页面宽度=0.21<0.25
所以应该加权重

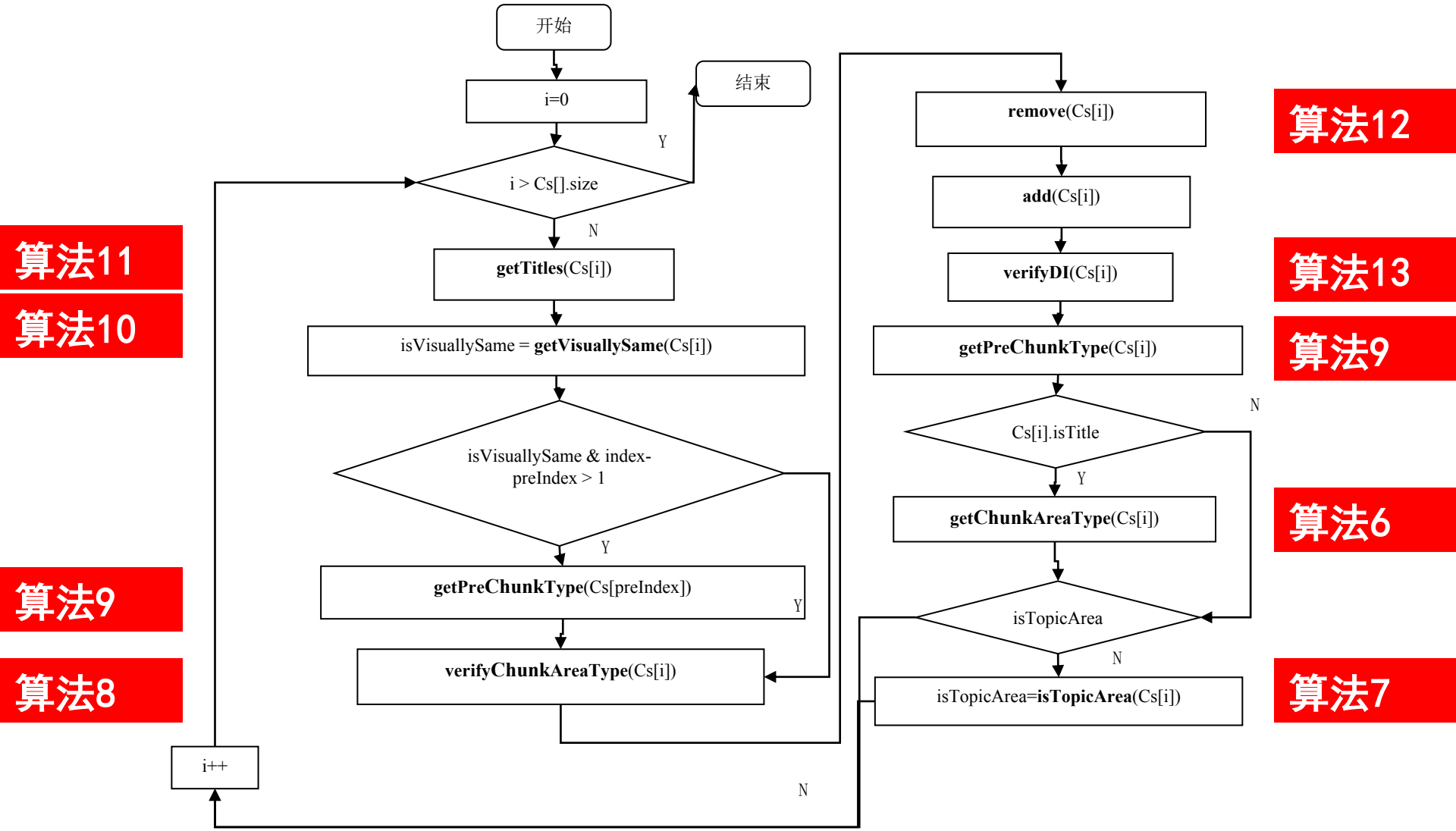
为何使用Bayes Network

- 贝叶斯网络是一种经典的分类模型，我们这里使用这种分类模型是因为我们在2011年已经对其它常用的分类模型，如朴素贝叶斯，决策树等模型进行过实验比较，结果表明，贝叶斯网络在处理该问题上的效果最好。

预测模型如何产生

- 手工对训练数据集网站的文本块进行分类
- 以向量形式存入到.arff文件中
- Java使用Weka的API对文件进行处理，生成预测模型
- 将预测模型存入本体中

何时获取上下文信息
何时进行文本块后处理



文本块类型和本体类的映射

本体类	文本块类型
Conference	-
Date	DI
Place	PI
Conference子类 Event	AR
Research_Topic	TO
作为Person的子类类型	PO
对应的PO所在的子类	PE
Paper Author	PA
视具体情况而定	CO

文本分析

- 时间类型文本块
  - 获取时间字符串
    - 没有：作为时间进程
    - 有：提取时间介词→定位进程字符串

算法15：
GetDate

- 地点类型文本块
  - 查找地点关键字
  - 获取关键字附近大写开头的单词

算法16：
GetPlace

文本分析

- 参与者类型文本块
  - 根据分隔符划分字符串
  - 根据关键字判定子字符串描述内容
    - 人名
    - 机构名
    - 职位

算法17：
GetPeople

- 论文类型文本块
  - 根据关键字划分字符串
  - 根据文本、关键字判定子字符串描述内容
    - 论文标题
    - 作者
    - 摘要

算法18：
GetPaper