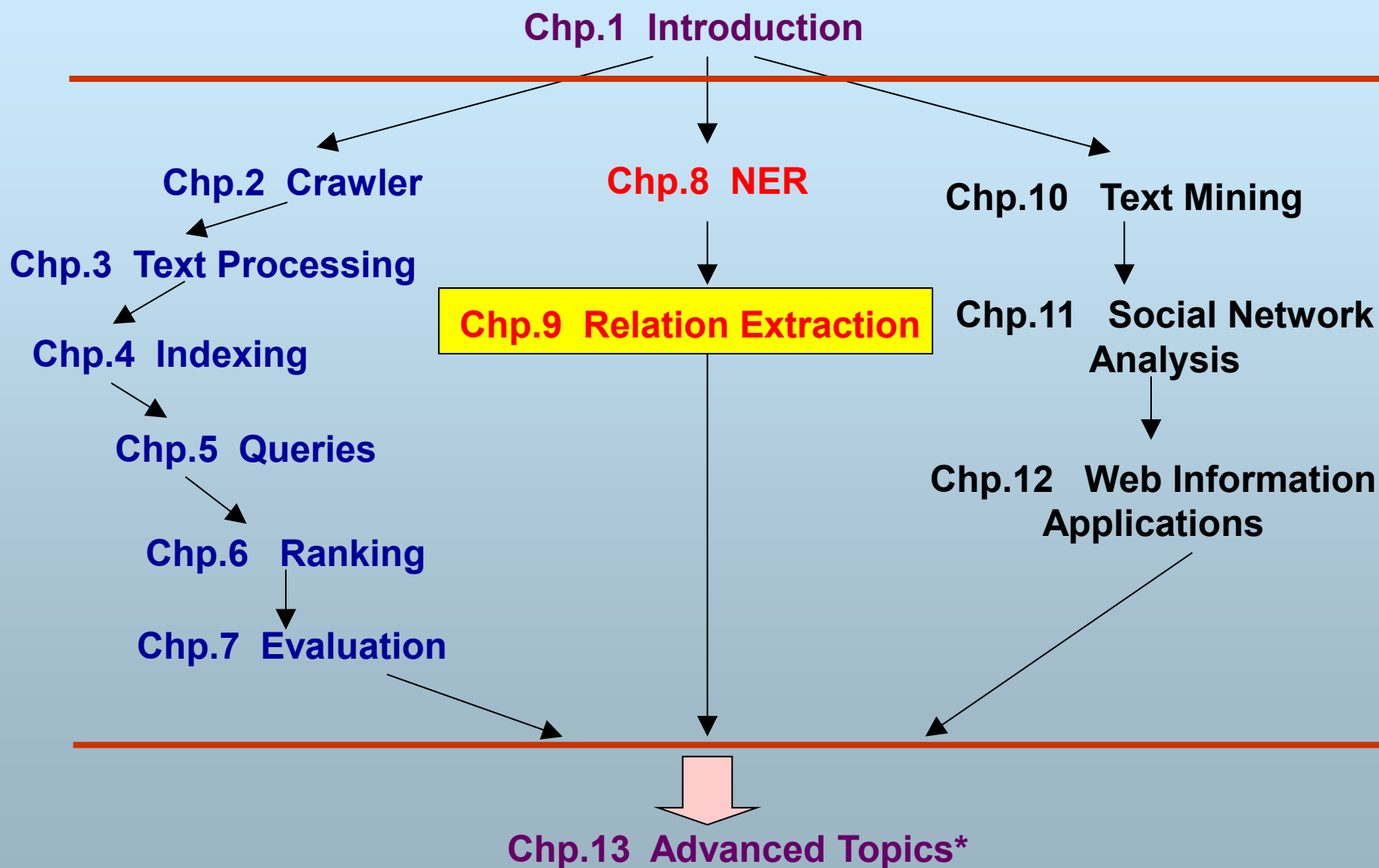


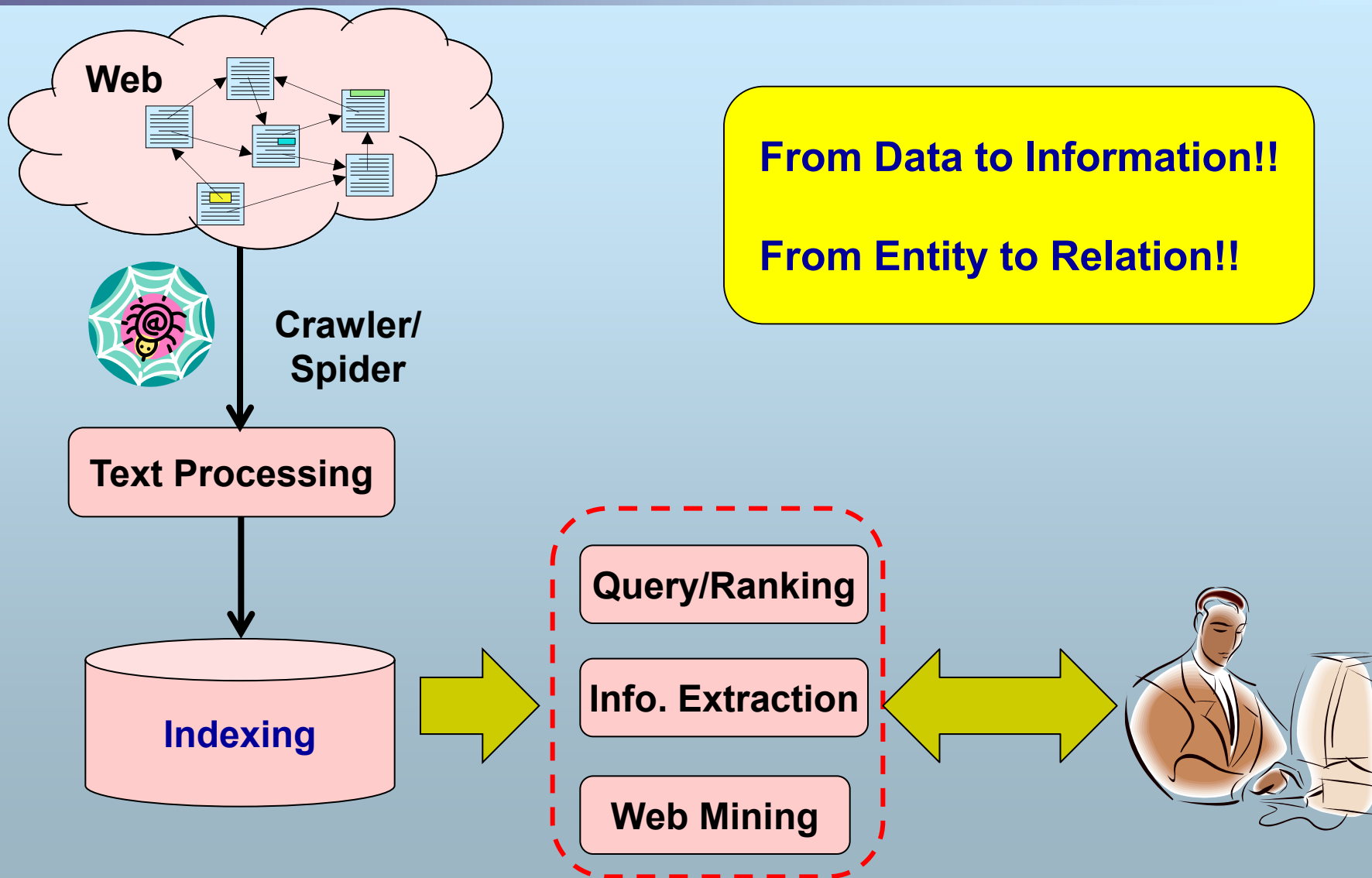
Relation Extraction



课程知识结构



本章讨论的问题



本章讨论的问题

■ 例如，给定下面的文档：

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY \$6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PERSON Tim Wagner] said. [ORG United Airlines] an unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

■ 命名实体识别之后我们能再抽取什么？



Extracting Relations

本章讨论的问题

■ 续上例

Domain

United, UAL, American Airlines, AMR

Tim Wagner

Chicago, Dallas, Denver, and San Francisco

$$\mathcal{D} = \{a, b, c, d, e, f, g, h, i\}$$
$$a, b, c, d$$
$$e$$
$$f, g, h, i$$

Classes

United, UAL, American, and AMR are organizations

Tim Wagner is a person

Chicago, Dallas, Denver, and San Francisco are places

$$Org = \{a, b, c, d\}$$
$$Pers = \{e\}$$
$$Loc = \{f, g, h, i\}$$

Relations

United is a unit of UAL

American is a unit of AMR

Tim Wagner works for American Airlines

United serves Chicago, Dallas, Denver, and San Francisco

$$PartOf = \{\langle a, b \rangle, \langle c, d \rangle\}$$
$$OrgAff = \{\langle c, e \rangle\}$$
$$Serves = \{\langle a, f \rangle, \langle a, g \rangle, \langle a, h \rangle, \langle a, i \rangle\}$$

本章主要内容

- 关系抽取概念
- 关系抽取方法

一、关系抽取概念

■ 关系抽取

- 从文本中识别出两个实体或多个实体之间存在的事实上的关系。

…在文本中检测实体之间语义关系的一种技术…

摘录自 维基百科

- 例如

Bill Gates works at Microsoft.



beEmployee(Bill Gates, Microsoft)

- 1997年，MUC-7上首次引入了关系抽取任务（Template Relation）

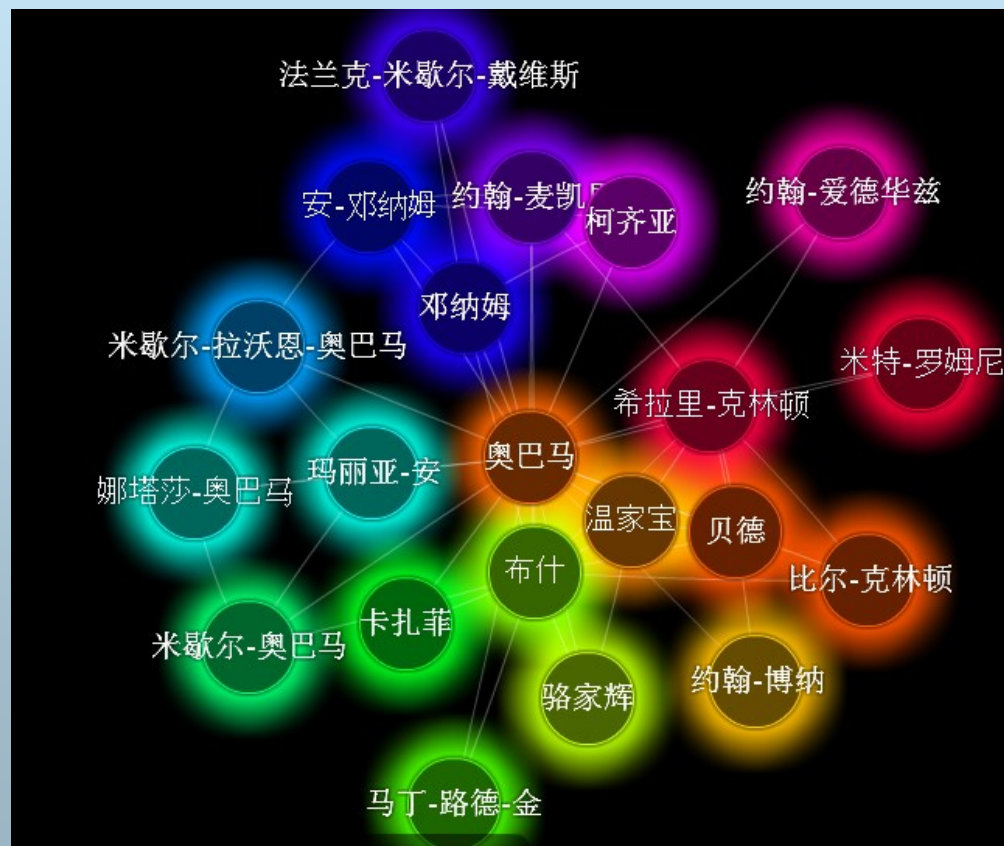
一、关系抽取概念

■ 关系抽取有什么意义？

- 作用1：提高搜索引擎发现知识的能力



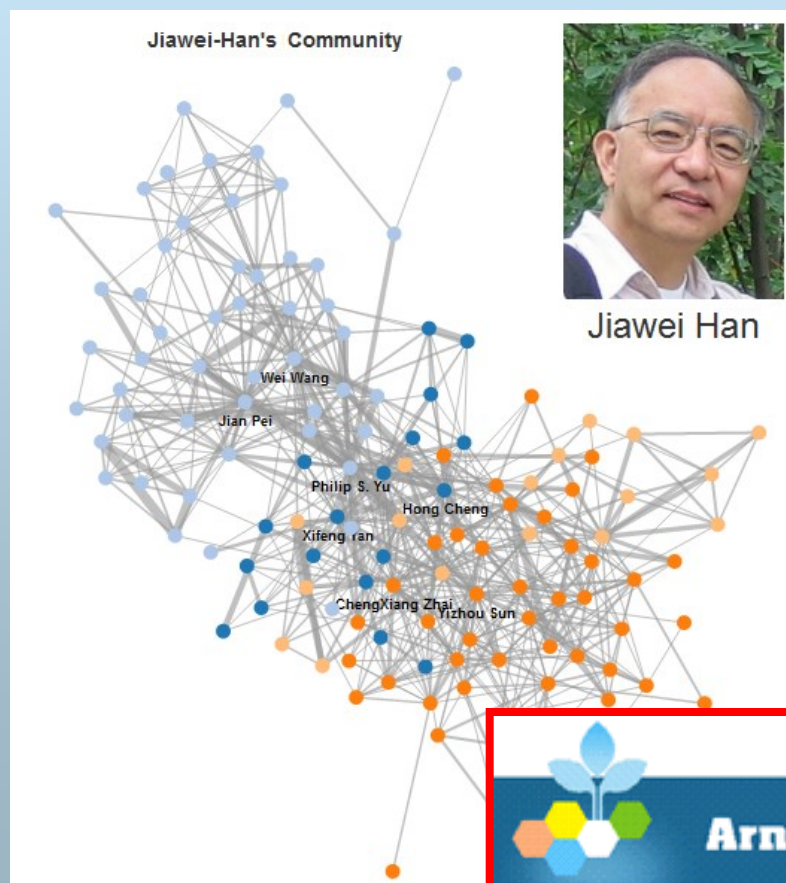
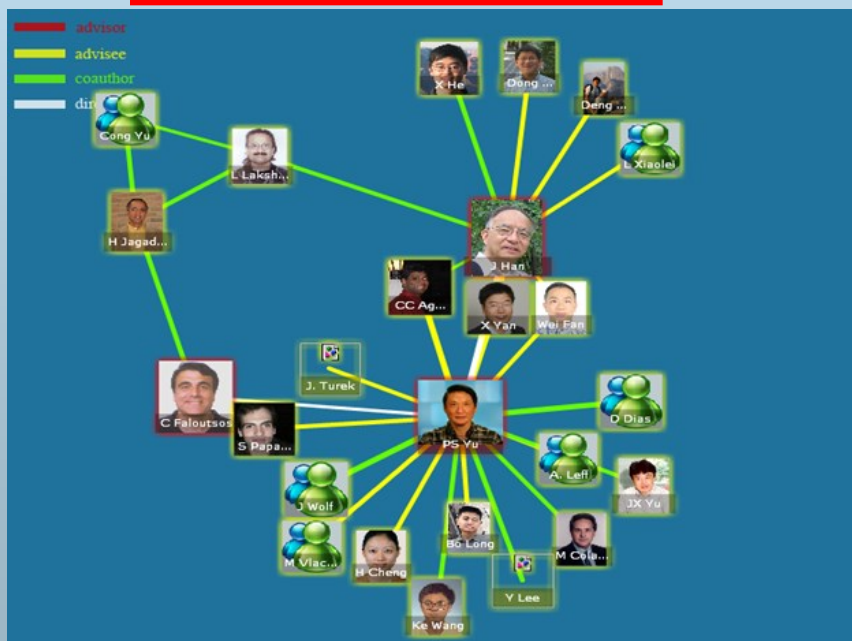
<http://renlifang.msra.cn/>



一、关系抽取概念

■ 关系抽取有什么意义？

- 作用1：提高搜索引擎发现知识的能力



一、关系抽取概念

■ 关系抽取有什么意义？

- 作用2：广泛应用于各种知识库的构建



<http://dbpedia.org/>

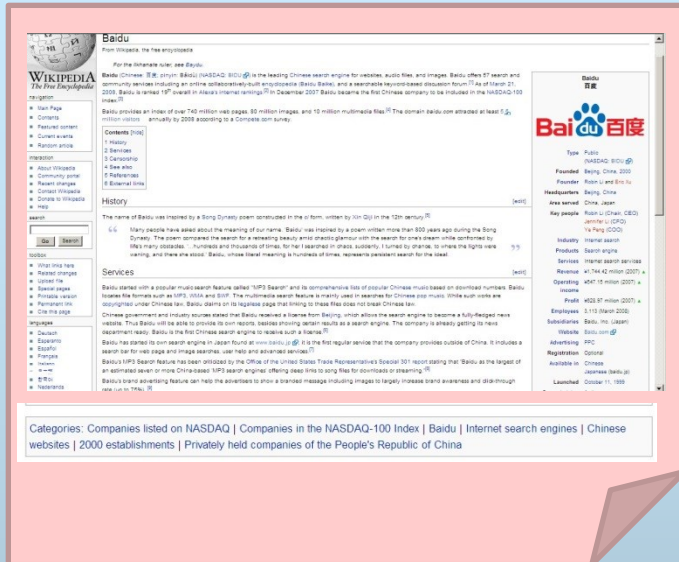
超过三百四十万个实体
以及十亿个实体关系



<http://www.mpi-inf.mpg.de/yago-naga/yago/>

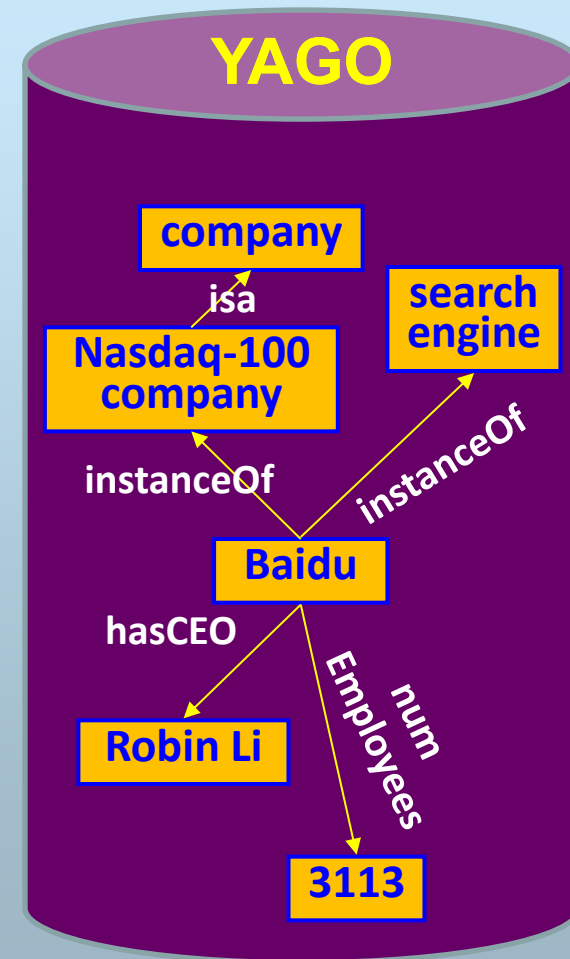
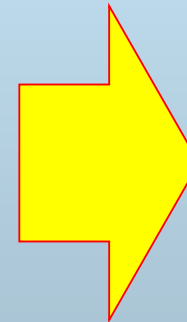
超过两百万个实体
以及两千万实体关系

YAGO



The screenshot shows the Baidu homepage with search results for 'Baidu'. The results include a Wikipedia entry and a list of categories: Companies listed on NASDAQ, Companies in the NASDAQ-100 Index, Baidu, Internet search engines, Chinese websites, 2000 establishments, Privately held companies of the People's Republic of China.

Baidu 百度	
Type	Public (NASDAQ: BIDU)
Founded	Beijing, China, 2000
Founder	Robin Li and Eric Xu
Headquarters	Beijing, China
Area served	China, Japan
Key people	Robin Li (Chair, CEO) Jenniffer Li (CFO) Ye Peng (COO)
Industry	Internet search
Products	Search engine
Services	Internet search services
Revenue	¥1,744.42 million (2007) ▲
Operating income	¥547.15 million (2007) ▲
Profit	¥828.97 million (2007) ▲
Employees	3,113 (March 2008)
Subsidiaries	Baidu, Inc. (Japan)
Website	Baidu.com
Advertising	PPC
Registration	Optional
Available in	Chinese Japanese (baidu.jp)
Launched	October 11, 1999



Categories: Companies listed on NASDAQ | Companies in the NASDAQ-100 Index | Baidu | Internet search engines | Chinese websites | 2000 establishments | Privately held companies of the People's Republic of China

一、关系抽取概念

■ 关系抽取有什么意义？

- 作用3：支持知识推理和问答系统研究

PATTY Relation Mining

<https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/>

PATTY Relation Mining
MPI-INF|Databases

Thesaurus Relations Taxonomy

Arg 1 Relation Arg 2

► Join

▼ Vldb12 Demo Queries

Find Brad Pitt costars (Semantic-relatedness query)

1-19 of 19

Arg1	Arg2	Pattern
Brad Pitt	Juliette Lewis	be reunited with;
Brad Pitt	Anthony Hopkins	also cast as;
Brad Pitt	Heinrich Harrer	led [[adj]];
Brad Pitt	Anthony Hopkins	won [[det]] chance to showcase [[adj]] skills to gain [[det]] younger set of was cast [[adj]];

一、关系抽取概念

■ 关系如何表示？

● 二元组 <subject, objects>

- ◆ 适合特定领域关系抽取，例如企业收购关系
- ◆ <Microsoft, Nokia>

● 三元组 <subject, predicate, object>

- ◆ 适合多类型关系抽取，例如企业之间的商业关系抽取
- ◆ <Microsoft, acquisition, Nokia>
- ◆ <Microsoft, cooperation, Intel>

● 多元组，例如 <subject, predicate, object, time>

- ◆ 目前在时态关系抽取上研究较多，例如
- ◆ <Clinton, as-president, USA, [2001, 2008]>
- ◆ <Trump, as-president, USA, [2016, NOW]>

一、关系抽取概念

■ 常用的关系抽取数据源

- Wikipedia
- Yellow pages

 <p>【三门】【厂家低价直销】 ¥470.00/5个起批 苏州市相城区欣乐办公家具</p>	 <p>隆重推荐!【汇绿品牌】 ¥100.00/50个起批 广东汇绿实验室设备科技有限</p>	 <p>供应文件柜(图) 价格面议 宁波市镇海铭祺五金制品厂</p>
 <p>特价 文件柜 钢制文件柜 ¥300.00/5个起批 苏州东兴办公家具有限公司</p>	 <p>供应文件柜、保险柜、更 ¥350.00/2件起批 南京市建邺区单国成五金经营</p>	 <p>厂家直销: 乐宝钢制分体 ¥650.00/10套起批 北京欣硕形科贸有限公司</p>

中国科学技术大学
University of Science and Technology of China



校训 红专并进 理实交融
创建时间 1958年
学校类型 公立大学、研究型大学
校长 侯建国
教师 3164
学生 15500多人, 其中博士生1900多人, 硕士生6200多人, 本科生7400多人[1]
本科生 7473
研究生 8100
校址 中国安徽省合肥市
校园环境 分为东、西、南、北四个校区(规划新建中校区)
隶属于 中国科学院
网站 <http://www.ustc.edu.cn>

二、关系抽取方法

- 基于规则的方法
- 基于模式的方法
- 基于机器学习的方法

1、基于规则的方法

- 根据欲抽取关系的特点预先手工设定一些词法、句法和语义模式规则，然后再从自由文本中寻找相匹配的关系实例

1、基于规则的方法

■ 例如：<X, IS_A, Y>关系抽取（同义词关系）

- Naproxen sodium is a nonsteroidal anti-inflammatory drug (NSAID). [wiki]

■ 抽取规则

● Rule 1: “Y such as X”

- ◆ Universities such as MIT and CMU

● Rule 2: “X or other Y”

- ◆ Apples or other fruits

● Rule 3: “Y including X”

- ◆ Typical machine learning methods including SVM and CRF
.....

● Rule 4: “Y, especially X”

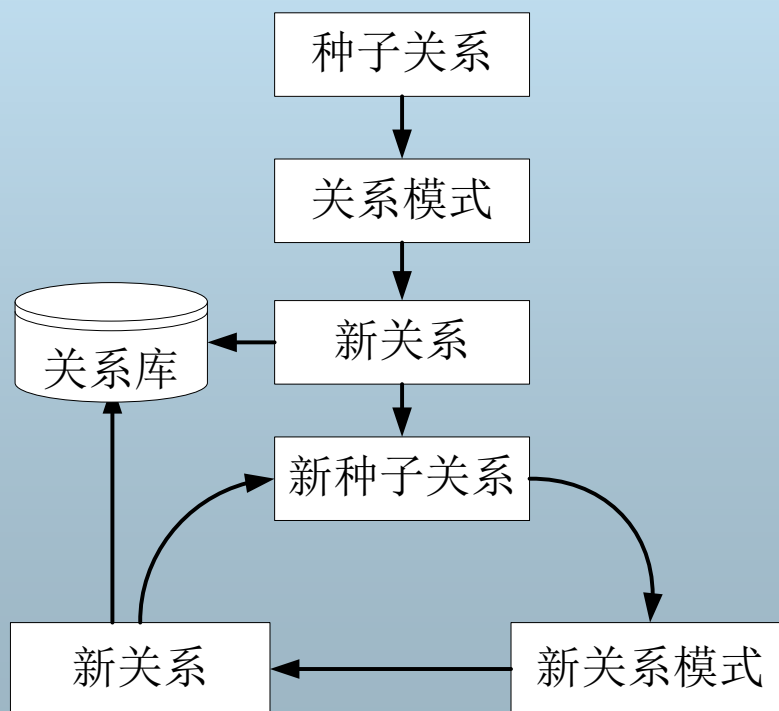
- ◆ Most students, especially Ph.D. candidates

1、基于规则的方法

- 通常针对特定领域的特定关系抽取任务，可以根据想抽取的关系的特点设计针对性的规则
- 基于手工规则的方法需要领域专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动，因此这种方法的代价很大。
- 知识库构建完成后，对于特定的领域的抽取具有较好的准确率，但移植到其他领域十分困难，效果往往较差。

2、基于模式的方法

- 首先由种子关系生成关系模式，然后基于关系模式抽取新的关系，得到新关系后，从中选择可信度高的关系作为新种子，再寻找新的模式和新的关系，如此不断迭代，直到没有新的关系或新的模式产生。



2、基于模式的方法

■ 代表性系统

- **DIPRE (Dual Iterative Pattern Relation Extraction)**
(Sergey Brin, 1998)
- **Snowball** (Eugene Agichtein, 2000)

DIPRE

- **Tuple**: Instance / occurrence of a relation
 - <Foundation, Isaac Asimov> --- <Title, Author>
- **Pattern**: consists of constants and variables
 - ?x , by ?y
- **Assumptions**
 - No single source contains all the tuples
 - Each tuple appears on many web pages
 - Components of tuple appear “close” together
 - There are repeated patterns in the way tuples are represented on web pages

DIPRE

■ Naïve Approach

- Study a few websites and come up with a set of patterns, e.g., regular expressions

`letter = [A-Za-z.]`

`title = letter{5,40}`

`author = letter{10,30}`

`(title) by (author)`

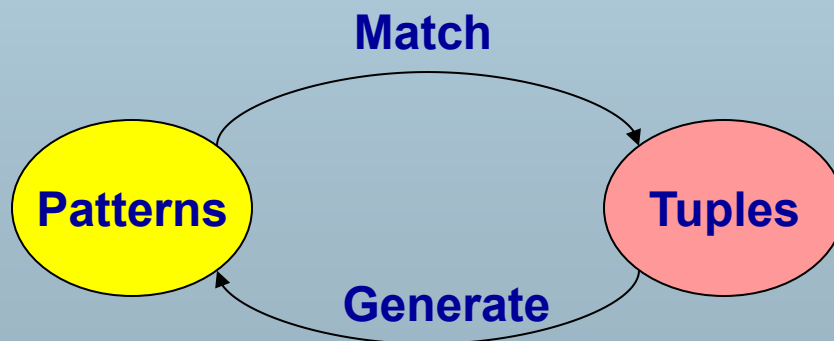
DIPRE

■ Problems with naïve approach

- A pattern that works on one web page might produce nonsense when applied to another
 - ◆ So patterns need to be page-specific, or at least site-specific
- Impossible for a human to exhaustively enumerate patterns for every relevant website
 - ◆ Will result in low coverage

DIPRE

- **Better Approach: 考虑patterns和tuples之间的双重影响关系**
 - Find tuples that match a set of patterns
 - Find patterns that match a lot of tuples
 - DIPRE (Dual Iterative Pattern Relation Extraction)



DIPRE

算法过程 (from Ullman's slides, Infolab@Stanford)

1. $R = \text{SampleTuples}$

- e.g., a small set of *<title, author>* pairs

2. $O = \text{FindOccurrences}(R)$

- Occurrences of tuples on web pages
- Keep some surrounding context

3. $P = \text{GenPatterns}(O)$

- Look for patterns in the way tuples occur

4. $R = \text{MatchingTuples}(P)$

5. Return or go back to Step 2

Occurrences

- e.g., Titles and authors
- Restrict to cases where author and title appear in close proximity on web page

** Foundation by Isaac Asimov (1951)**

- **url** = <http://www.scifi.org/bydecade/1950.html>
- **order** = [title,author] (or [author,title])
 - denote as 0 or 1
- **prefix** = “ ” (limit to e.g., 10 characters)
- **middle** = “ by ”
- **suffix** = “(1951) ”
- **occurrence** =
(‘Foundation’, ‘Isaac Asimov’, url, order, prefix, middle, suffix)

Patterns

` Foundation by Isaac Asimov (1951)`

`<p> Nightfall by Isaac Asimov (1941)`

- `order = [title, author] (say 0)`
- `shared prefix = `
- `shared middle = by`
- `shared suffix = (19`
- `pattern = (order, shared prefix, shared middle, shared suffix)`

URL Prefix

- Patterns may be specific to a website
 - Or even parts of it
- Add *urlprefix* component to pattern

<http://www.scifi.org/bydecade/1950.html> occurrence:

 Foundation by Isaac Asimov (1951)

<http://www.scifi.org/bydecade/1940.html> occurrence:

<p> Nightfall by Isaac Asimov (1941)

shared *urlprefix* = http://www.scifi.org/bydecade/19

pattern = (urlprefix,order,prefix,middle,suffix)

Generating Patterns

1. Group occurrences by *order* and *middle*
2. Let **O** = set of occurrences with the same *order* and *middle*
 - `pattern.order = O.order`
 - `pattern.middle = O.middle`
 - `pattern.urlprefix =` longest common prefix of all urls in O
 - `pattern.prefix =` longest common prefix of occurrences in O
 - `pattern.suffix =` longest common suffix of occurrences in O

Example

<http://www.scifi.org/bydecade/1950.html> occurrence:

` Foundation by Isaac Asimov (1951)`

<http://www.scifi.org/bydecade/1940.html> occurrence:

`<p> Nightfall by Isaac Asimov (1941)`

- `order = [title, author]`
- `middle = " by "`
- `urlprefix = http://www.scifi.org/bydecade/19`
- `prefix = " "`
- `suffix = " (19"`

Example

<http://www.scifi.org/bydecade/1950.html> occurrence:
Foundation, by Isaac Asimov, has been hailed...

<http://www.scifi.org/bydecade/1940.html> occurrence:
Nightfall, by Isaac Asimov, tells the tale of...

- **order = [title, author]**
- **middle = ", by "**
- **urlprefix = <http://www.scifi.org/bydecade/19>**
- **prefix = ""**
- **suffix = ", "**

Finding occurrences and matches

■ Finding occurrences

- Use inverted index on web pages
- Examine resulting pages to extract occurrences

■ Finding matches

- Use *urlprefix* to restrict set of pages to examine
- Scan each page using regular expressions constructed from pattern

Snowball

- Snowball (*Agichtein and Gravano, 2000*)
- Improvement of DIPRE
 - Trust only tuples that match many patterns
 - Trust only patterns with high “*support*” and “*confidence*”

Snowball

■ Pattern Support

- Eliminate patterns not supported by at least n_{min} known good tuples
 - ◆ either seed tuples or tuples generated in a prior iteration

■ Pattern Confidence

- Suppose tuple t matches pattern p
- the confidence of pattern p
 - ◆ the probability that tuple t is valid

2、基于模式的方法

- 不同算法的差异主要在于模式生成方法和匹配方法
- 适合某种特定的具体关系的抽取，如校长关系、首都关系
- 基于字面的匹配，没有引入更深层次的信息，如词性、句法、语义信息等。
- 移值性差，必须为每一个具体的关系生成自己的识别模式。

3、基于机器学习的方法

- 采用机器学习方法关系抽取模型，然后，先通过标注语料库训练得到一个再利用该模型对自由文本中出现的关系实例进行识别。
- 往往将关系抽取问题变换为一个分类问题，然后采用机器学习中常用的分类器来解决。

3、基于机器学习的方法

■ 分类

● 基于特征向量的方法

- ◆ 基于特征向量，然后使用SVM、最大熵（ME）等进行分类
- ◆ 关键在于特征集的确定而不是机器学习方法
- ◆ 目前的难点在于如何找出适合关系抽取的新的有效的词汇、句法或语义特征

● 基于核函数的方法（Zelenko, 2003）

- ◆ 不需要构建特征向量，而是利用结构特性进行抽取
- ◆ 直接将对关系实例表示为某种结构的树（例如语法树），并通过计算某种核函数的值来计算实例之间的相似度(例如基于相同子树的数目、相同路径的长度等)，再使用支持核函数的分类器进行关系抽取
- ◆ 近年来的研究热点
 - 浅层树核（Zelenko, 2003）、依存树核（Culotta, 2004）、最短依存树核（Bunescu, 2005）、卷积树核（Zhou, 2007）

Zelenko D., Aone C., Richardella A., Kernel methods for relation extraction, Journal of Machine Learning Research, 2003, 3: 1083-1106.

本章小结

■ 关系抽取的概念

- 实体之间的关系，应用领域广泛，是事件抽取的基础

■ 关系抽取的方法

● 基于规则的方法

- ◆ 简单，适合特定领域的特定关系

● 基于模式的方法

- ◆ 半监督学习的方法，自我修正

● 基于机器学习的方法

- ◆ 基于特征向量的方法：如SVM、ME等，需要构建特征向量

- ◆ 基于核函数的方法

- 不需要构建特征向量，在学习和分类过程中基于结构化树的核函数来度量相似性
- 可以同时抽取多类型关系