NLP Techniques in Knowledge Graph

Shiqi Zhao

Outline

- **Baidu Knowledge Graph**
- Knowledge Mining
- Semantic Computation

Zhixin for Baidu PC Search

- Knowledge graph

Zhixin for Baidu PC Search

- Exact answers

Zhixin for Baidu PC Search

- List recommendation

Zhixin for Baidu Mobile Search

- Knowledge graph

Zhixin for Baidu Mobile Search

- Exact answers

Zhixin for Baidu Mobile Search

- List recommendation

Outline

- Baidu Knowledge Graph
- **Knowledge Mining**
- Semantic Computation

Knowledge to Mine

Named entity mi

→ 中国合伙人

AVP mining

→

中文名：	中国合伙人	类型：	剧情，喜剧，文艺，青春
外文名：	American dreams in China	主演：	黄晓明，邓超，佟大为
其它译名：	三个中国先生/中国先生	片长：	112分钟
出品时间：	2013年	上映时间：	2013年5月17日（中国）
出品公司：	中影集团、我们制作有限公司	分级：	Hong Kong：IIA
制片地区：	中国	对白语言：	汉语普通话
导演：	陈可辛	色彩：	彩色
编剧：	周智勇、张冀	发行公司：	北京光线影业有限公司

Hyponymy learning

→ 电影->励志电影->中国合伙人

Related entity mining

→

Mining Named Entities

- Traditional NE categories
  - person, location, organization
- Many more new categories useful for web applications
  - Movie, TV series, music, book, software, computer game
- More fine-grained taxonomy
  - Organization -> {school, hospital, government, company,…}
  - Computer game -> {net game, webpage game,…}
- Characteristics of NEs on the web
  - New NEs emerge rapidly, especially for software, games, and novels
  - Names of NEs on the web are informal

Learning NEs from Query Logs

- Query logs contain a large volume of named entities
  - About 70% search queries contain Nes (Pasca, 2007)
- NEs can be recognized using context features

Baidu suggestions

Useful context words for NE recognition:

电影 | 在线观看 | 百度影音 | 下载 | 完整版 | 经典台词 | 影评 | 插曲

Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In CIKM.

Learning NEs from Query Logs

- Bootstrapping approach:
  - Given a hand of seed NEs of a category *C*:
    - Learning context features of the seeds from queries
    - Extracting new seed entities of category *C* using the learnt context features
    - Expanding context features using the expanded seed set
    - ……
- Advantage of query log based method
  - It can cover newly emerging NEs
- Disadvantage of query log based method
  - Old or unpopular NEs are likely to be missed

Learning NEs from Plain Texts

- Text wrappers are widely used for extracting NEs from plain texts
  - Wrapper example: “电影《[X]》”, “影片[X]，导演”
  - [X] is a placeholder that can be filled with movie names

中国合伙人 在线观看_高清视频完整版_电影网
2012年7月23日 - “除了票房,这个世界上还没有另外的标准更适合衡量一部电影”,照此来看,陈可辛的确成功了,《中国合伙人》的火爆票房表明了他对内地影市的精准拿捏。...
www.m1905.com/mdb/film/22121... 2012-7-23 - 百度快照 687

《中国合伙人》百度影音_高清在线观看_电影_琪琪影院
中国合伙人百度影音高清在线观看,剧情片中国合伙人演员黄晓明 邓超 佟大为 杜鹃/中国先生/三个中国先生,中国合伙人剧情介绍:电影《中国合伙人》是一部根据真人真事...
www.77vcd.com/Drama/zhongguohehuor... 2013-5-25 - 百度快照 643

Learning NEs using Url-text Hybrid Patterns

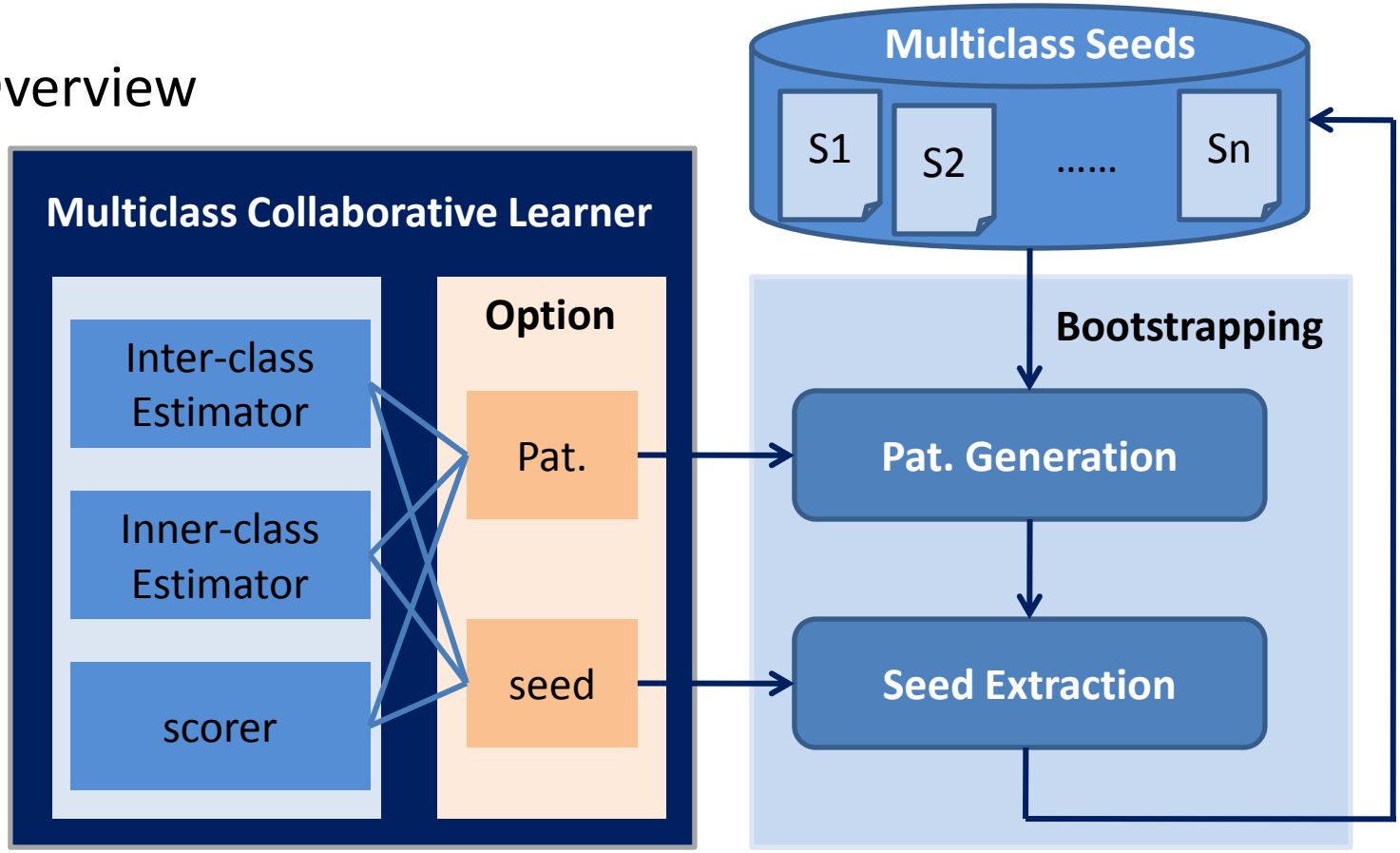
- Is it possible to extract NEs from webpage titles only?
  - Yes! 99% NEs can be found in some webpage titles
- Url-text hybrid patterns
  - Url constraints should be taken into consideration
    - Simple text patterns are enough for credible url (website)
    - Complicated text patterns are needed for low-quality url
- Url-text hybrid pattern learning
  - utp = (up, tp, c, f)
  - Example:

up	http://www.imdb.com/name/nm\d+/
tp	^ (.+?)\- imdb
c	Star
f	0.9

Zhang et al. 2013. Bootstrapping Large-scale Named Entities using URL-Text Hybrid Patterns. To appear in IJCNLP.

Learning NEs using Url-text Hybrid Patterns

- Overview



Zhang et al. 2013. Bootstrapping Large-scale Named Entities using URL-Text Hybrid Patterns. To appear in IJCNLP.

Learning NEs using Url-text Hybrid Patterns

- Multiclass Collaborative Learning (MCL)
  - NEs of multiple classes are extracted simultaneously
  - Bootstrapping NEs and url-text hybrid patterns iteratively
    - A small set of seeds is required for each class
  - Inter-class and intra-class scoring approaches are used for controlling the quality of NEs and patterns yielded in each iteration
    - **Inter-class scoring**: A correct NE of a class should not be extracted by patterns of other classes; A correct pattern of a class should not extract seeds from other classes.
    - **Intra-class scoring**: A correct NE of a class should not be extracted by only one pattern of the class; A correct pattern of a class should not yield a lot of NEs that cannot be extracted by other patterns of the class.

Zhang et al. 2013. Bootstrapping Large-scale Named Entities using URL-Text Hybrid Patterns. To appear in IJCNLP.

AVP Mining

- AVP: Attribute Value Pairs
- Where do AVP data come from?
  - Online encyclopedia
    - Baidu Baike, wikipedia
  - Vertical websites
    - IMDB, douban for videos
  - Plain web documents
    - Automatically mining AVP knowledge from the structured, semi-structured, and unstructured texts

AVP Mining from Online Encyclopedia

Not accurate enough

Semi-struc
information, wh
to be detected and
extracted automatically

Accurate, but not perfect

structured infobox
can be directly

柯震东

百科名片

柯震东（Kai Ko）

中文名：	柯震东	职业：	学生，演员，歌手
外文名：	Kai Ko	经纪公司：	可米瑞智国际艺能有限公司
别名：	凯凯	代表作品：	《那些年，我们一起追的女孩》，音乐专辑《有话直说》
国籍：	中国		
民族：	汉族	主要成就：	第48届台湾电影金马奖最佳新演员
出生地：	台湾澎湖	唱片公司：	索尼音乐
出生日期：	1991年6月18日		

人物档案

姓名：柯震东

罗马拼音：Ko Chen Tung

英文名：Kai Ko

私下昵称：凯凯

身高：183cm

体重：75kg

血型：B型

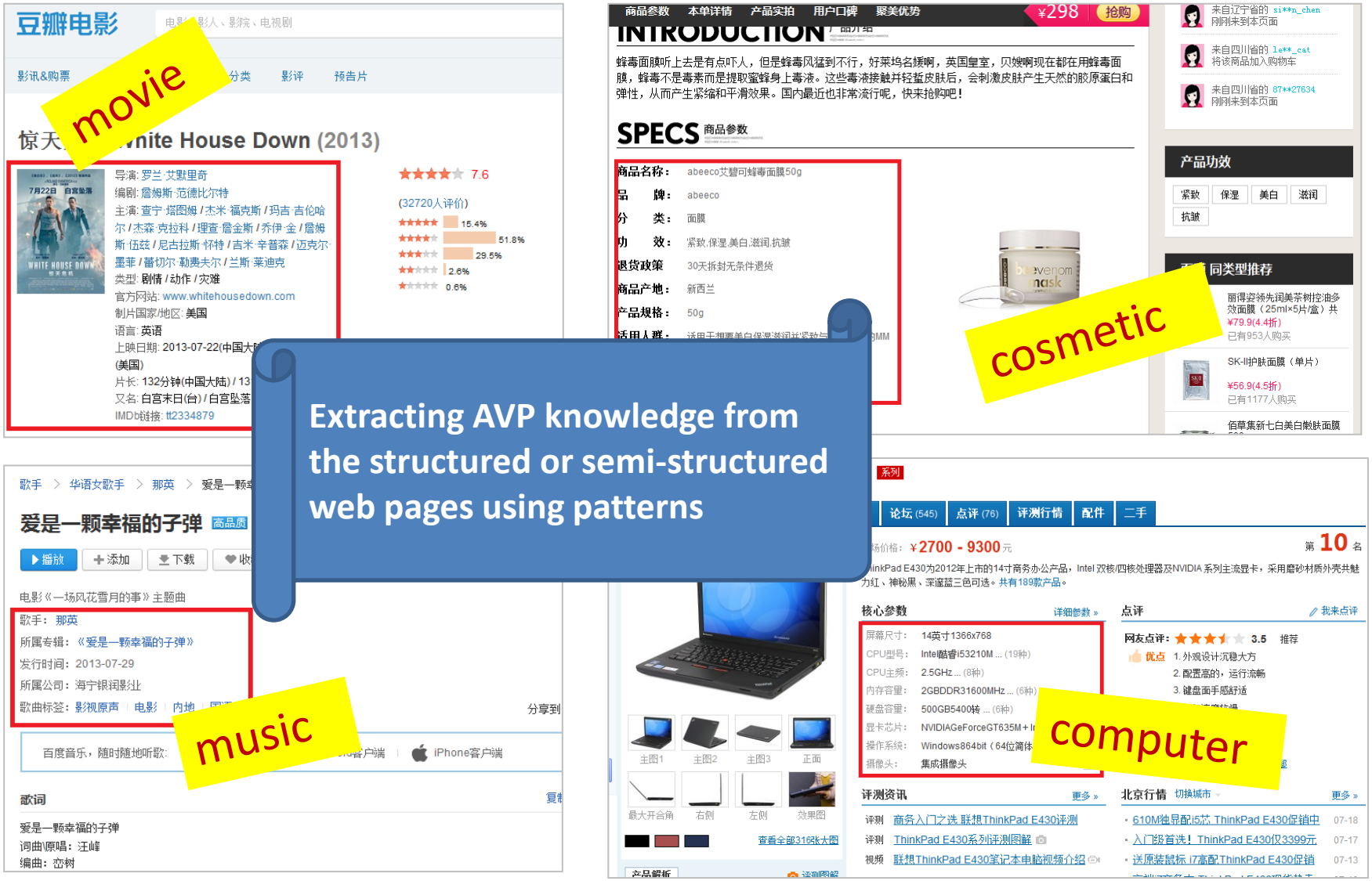
语言：普通话、闽南话

专长：田径、篮球

祖籍：浙江省宁波市象山县

最喜欢的造型师：高源泽

AVP Mining from Vertical Websites



AVP Mining from Vertical Websites

- Two problems
  - How to find the vertical websites?
    - It is easy to find the websites for large and popular domains
      - E.g., movie, music, novel
    - It is hard to find such websites for long tail domains
      - E.g., cosmetic, magazine
  - How to generate extraction patterns?
    - Different websites cannot share identical patterns☹

AVP Mining from Vertical Websites

- How to find the vertical websites
  - Prepare some seed queries for the desired category
  - Extract websites getting the most clicks from query logs
  - Bootstrapping is fine, but not necessary
  - Manual assessment should be done to select the ones of high quality

It is much easier than manually collecting all the websites

AVP Mining from Vertical Websites

- How to generate extraction patterns
  - Hand-crafted patterns to guarantee high accuracy

  - But we have tools that can help us edit patterns conveniently

AVP Mining from Vertical Websites

- AVP knowledge is accumulated on a day to day basis
  - Different categories are updated in different time intervals
  - New websites are added once they are identified
  - Disordered or run-down websites are automatically detected and manually processed

Outline

- Baidu Knowledge Graph
- Knowledge Mining
- **Semantic Computation**

Semantic Computation

- All modules should be optional
  - The input AVP data decides which modules are necessary
  - The dependency among the modules must be obeyed
- The modules are mostly semi-automatic tools
  - Human intervention is needed for supplying seeds, rules, or judges
  - Automatic methods are used for generating candidates before manual labeling

Cleaning

- Detect and clean surface errors
  - Unreadable codes
  - Erroneous Truncation
  - Erroneous attributes
    - Due to mining errors
    - Can usually be detected based on frequency
  - Double byte – single byte replacement
  - English character processing

Value Type Recognition

- Automatically recognize the value types of given attributes based on the AVP data
- Value types include:
  - Number
  - Date / time
  - Entity
  - Enumeration
  - Text (default)
- It can help recognizing illegal attribute values and extracting candidate synonymous attribute names

Value Normalization

- Splitting
  - E.g., *movie_a, movie_b, and movie_c -> movie_a | movie_b | movie_c*
- Generation
  - E.g., *Chinese zodiac / zodiac*: *Tiger / The lion ->*

    Chinese zodiac: *Tiger* and *zodiac*: *The lion*
- Conversion
  - E.g., 2.26m -> 226cm

Attribute Normalization

- Domain-specific problem
  - Some attributes are deemed synonymous only in specific domain or even for two specific knowledge sources
    - E.g., “大小 (size)” and “屏幕 (screen)” are synonymous for some websites on mobile phone, but not open domain paraphrases

Attribute Normalization (cont.)

- Classification model for identifying **candidate** synonymous attributes
  - Features:
    - Attribute surface similarity features
    - Value similarity features
    - Value-type similarity features
    - Entity-value feature
- Raters select correct synonymous attribute pairs from all candidates

Knowledge Fusion

- Fusion of knowledge mined from various data sources
- Key problem:
  - Entity disambiguation
- Solution:
  - Compute the similarity between entities with the same name
    - Some essential attributes can determine the identity of an entity
      - E.g., works of a writer
    - Some other attributes can only be used as similarity features
      - E.g., nationality of a person

Entity Classification

- Why classification is needed?
  - The category information is missing for some entities
  - Not all possible categories of an entity can be mined from the data source
- Solution:
  - Supervised model trained with entities along with their AVPs whose categories are known
  - Both structured data (AVPs) and unstructured data (contextual texts) can be used for exacting classification features

Semantic Computation (cont.)

- Some other semantic computation modules in the knowledge application level
  - Entity disambiguation for reasoning

Attribute	Value
name	Chen Xiaoxu
born	October 29, 1965
television work	Dream of Red Mansions

Link the name to the correct entity ID

Reasoning, e.g., actors who have worked together with Chen Xiaoxu

Semantic Computation (cont.)

- Some other semantic computation modules in the knowledge application level (cont.)
  - Related entity disambiguation

Semantic Computation (cont.)

- Some other semantic computation modules in the knowledge application level (cont.)
  - Search requirement recognition

Semantic Computation (cont.)

- Some other semantic computation modules in the knowledge application level (cont.)
  - Key problem: AVP similarity computation

Attributes should be assigned with different weights

Useful attributes for a tennis player: *Australian Open, French Open, championships, Olympic games, career titles,…*

Useless attributes for a tennis player: *country, residence, born, family,…*

Word mismatch problem should be resolved

Word mismatch problem is especially serious when the AVPs of two entities come from different data sources
e.g.:
silver medal -> second place
fierce forehand -> aggressive forehand

Conclusion

- New trends of web search
  - Knowledge search, semantic search, social search
- Research on semantics is essential for knowledge graph
  - Knowledge base construction and knowledge search both need semantic computation
- Various web resources should be made better use of
  - Web corpora, Query logs, UGC data

Thanks!

Q&A