

NLP Techniques in Knowledge Graph

Shiqi Zhao

Outline



Baidu Knowledge Graph



Knowledge Mining



Semantic Computation

Zhixin for Baidu PC Search

- Knowledge graph

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

泰囧

百度一下

推荐: [用手机随时随地地上百度](#)

Named entities

猜您喜欢

人再囧途之泰囧在线观看 百度视频

2012年上映 | 豆瓣评分: 7.6

地区: 内地

类型: 喜剧 | 剧情 | 动作

人再囧途之泰囧

人再囧途之泰囧

人再囧途之泰囧

人再囧途之泰囧

人再囧途之泰囧

人再囧途之泰囧

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

火龙果

百度一下

推荐: [用手机随时随地地上百度](#)

Normal entities

火龙果 百度百科

火龙果, 英文Pitaya, 本名青龙果、红龙果。原产于中美洲热带。火龙果营养丰富、功能独特, 它含有一般植物少有的植物性白蛋白及花青素, 丰富的维生素和可溶性膳食纤维。火龙果树为仙人掌科的三角...

基本介绍 - 物种分布 - 营养价值 - 形态特征 - 分类学 - 更多>>

[baike.baidu.com/](#) 2013-08-03

火龙果 百度图片 - 举报图片

火龙果种植 盆栽火龙果种植图解 火龙果的切法 红肉火龙果

[image.baidu.com](#) - 查看全部266,000张图片

其他人还搜

山竹 红心火龙果 猕猴桃 榴莲 火龙果花

Zhixin for Baidu PC Search

- Exact answers

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

法国首都

百度一下



巴黎
法国,首都

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

中国人口

百度一下

1,353,821,000人 (2012年估计)

中国,人口

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

勿忘我花语

百度一下



永恒的爱、浓情厚谊、永不变的心
勿忘我,花语

来自百度百科 | 报错

Zhixin for Baidu PC Search

List recommendation

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

聪明的狗

百度一下

您要找的是不是以下结果:



边境牧羊犬

请问边牧会看家吗 - 边境牧羊犬俱乐部 - 狗民论坛 - 狗民网

狗民论坛 边境牧羊犬俱乐部 请问边牧会看家吗 上一主题 | 下一主题 4 3 ... 会, 我家的最近地盘意识见长, 见生人都会狂吠。而且我...

[边境牧羊犬_百度百科](#) | [更多相关搜索结果>>](#)



贵宾犬

贵宾犬训练视频世界最聪明的犬种排行_贵宾专题网

据美国哥伦比亚大学心理学教授STANLEYCOREN结合208位各地训狗专家, 63名小型动物兽医, 及14名研究警戒犬与护卫狗的专家对各著名犬...

[贵宾犬_百度百科](#) | [更多相关搜索结果>>](#)

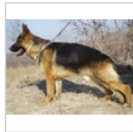


拉布拉多

最聪明的狗是拉布拉多吗? - 精华知识 - 搜搜问问

拉布拉多是非常聪明的狗狗, 广泛应用于救生、搜救、搜爆、缉毒、导盲等领域, 它的适用性和实用性是边境牧羊所无法比的。 好:6 不好:...

[拉布拉多_百度百科](#) | [更多相关搜索结果>>](#)



德国牧羊犬

阿拉斯加雪橇犬德国牧羊犬哈士奇哪种狗最适合看家护院_百度知道

您好, 我觉得还是德国牧羊犬比较适合看家护院, 以上内容仅供参考。 ... 当然是德...但是他们三个都不是专门看家的狗, 德牧智商最高, ...

[德国牧羊犬_百度百科](#) | [更多相关搜索结果>>](#)

展开更多 聪明的狗

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

适合放在卧室的植物

百度一下

您要找的是不是以下结果:



吊兰

吊兰, 适合放到卧室吗? - 卧室-养花-百科-天涯问答

吊兰, 适合放到卧室吗? chdaozh2009-10-29 14:14:22 发布 卧室 养花 百科 ...冬季不要放在阳台、窗边, 以免冻着。 添加评论 评...

[吊兰_百度百科](#) | [“适合放在卧室的植物_吊兰”的更多结果>>](#)

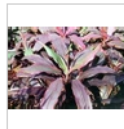


绿萝

绿萝花可以放在卧室里吗_百度知道

放心吧不会影响健康没有那么大的影响力 建议晚上就搬到阳台白天搬回来, 绿叶植物会在白天大量消耗二氧化碳制造氧气, 但是到了晚上不...

[绿萝_百度百科](#) | [“适合放在卧室的植物_绿萝”的更多结果>>](#)



千年木

适宜放在卧室里的植物 - 花卉百科 - 藏花阁花卉论坛 - Powered by ...

适宜放在卧室里的植物千年木只要对它稍加关心, 它就能长时间生长, 并带来优质的空气。在抑制有害物质方面其他植物很难与千年木相提...

[千年木_百度百科](#) | [“适合放在卧室的植物_千年木”的更多结果>>](#)



虎尾兰

虎尾兰适合放在卧室里吗?_百度知道

可以, 虎尾兰在卧室内还可以有效地吸收房间内的有害气体, 如甲醛等。 ...很适合室内养 ...适合 它夜间放氧。且吸收有害气体强 ...

[虎尾兰_百度百科](#) | [“适合放在卧室的植物_虎尾兰”的更多结果>>](#)

展开更多 适合放在卧室的植物

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



致我们终将逝去的青春

厨子戏子痞子

不二神探

速度与激情5

西游·降魔篇

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



**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

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

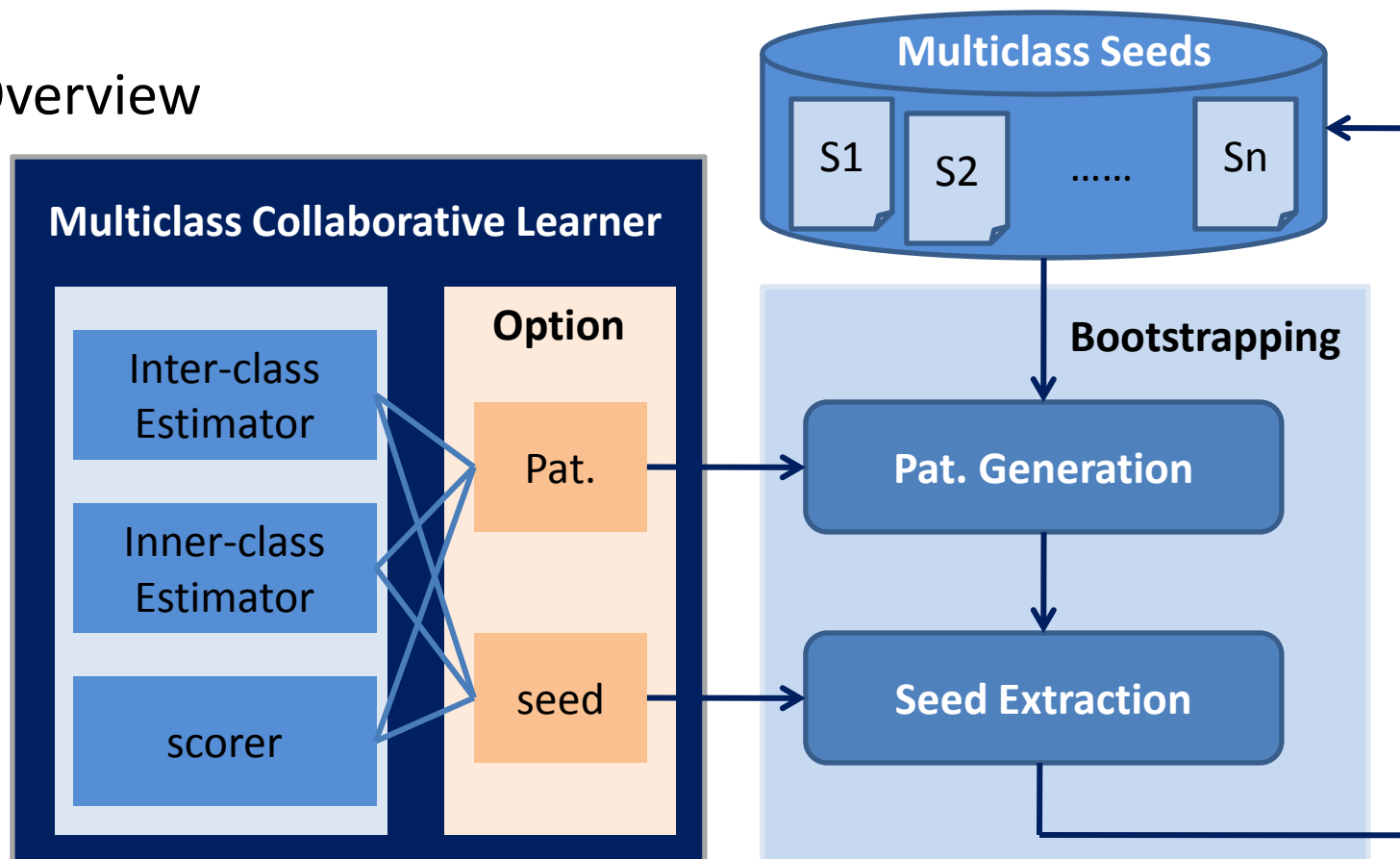
Learning NEs using Url-text Hybrid Patterns

- 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:

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

Learning NEs using Url-text Hybrid Patterns



- Overview



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.

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

柯震东

百科名片



柯震东 (Kai Ko)



structured infobox
can be directly



Semi-structured
information, which needs
to be detected and
extracted automatically

Not
accurate
enough



Accurate,
but not
perfect

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](#)

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	主要成就:	第48届 台湾电影金马奖 最佳新演员
民族:	汉族	唱片公司:	索尼音乐
出生地:	台湾澎湖		
出生日期:	1991年6月18日		

AVP Mining from Vertical Websites

豆瓣电影 电影 影人、影院、电视剧

影讯&购票 分类 影评 预告片

惊天动地 White House Down (2013)

导演: 罗兰·艾默里奇
编剧: 詹姆斯·范德比尔特
主演: 查宁·塔图姆 / 杰米·福克斯 / 玛吉·吉伦哈尔 / 杰森·克拉科 / 理查·詹金斯 / 乔伊·金 / 詹姆斯·伍兹 / 尼古拉斯·怀特 / 吉米·辛普森 / 迈克尔·墨菲 / 蕾切尔·勒费夫尔 / 兰斯·莱迪克
类型: 剧情 / 动作 / 灾难
官方网站: www.whitehousedown.com
制片国家/地区: 美国
语言: 英语
上映日期: 2013-07-22(中国大陆) / 13(美国)
片长: 132分钟(中国大陆) / 132分钟(美国)
又名: 白宫末日(台) / 白宫坠落
IMDb链接: t.cn/R334879

★★★★★ 7.6
(32720人评价)
★★★★★ 15.4%
★★★★ 51.8%
★★★ 29.5%
★★ 2.6%
★ 0.6%

movie

商品参数 本单详情 产品实拍 用户口碑 聚美优势

INTRODUCTION

蜂毒面膜听上去是有点吓人，但是蜂毒面膜猛到不行，好莱坞名媛，英国皇室，贝嫂啊现在都在用蜂毒面膜，蜂毒不是毒素而是提取蜜蜂身上毒液。这些毒液接触并轻猛皮肤后，会刺激皮肤产生天然的胶原蛋白和弹性，从而产生紧致和平滑效果。国内最近也非常流行呢，快来抢购吧！

SPECS 商品参数

商品名称: abeeeco艾碧可蜂毒面膜50g
品牌: abeeeco
分类: 面膜
功效: 紧致、保湿、美白、滋润、抗皱
退货政策: 30天拆封无条件退货
商品产地: 新西兰
产品规格: 50g
适用人群: 适用干性、油性、混合性、敏感性及所有肤质

¥298 抢购

来自辽宁省的 si**n_chen 刚刚来到本页面
来自四川省的 1c**_cat 将该商品加入购物车
来自四川省的 87**27634 刚刚来到本页面

产品功效

紧致 保湿 美白 滋润 抗皱

同类推荐

丽得姿领先词美茶树控油多效面膜 (25ml×5片/盒) 共 ¥79.9(4.4折) 已有953人购买
SK-II护肤面膜 (单片) ¥56.9(4.5折) 已有1177人购买
信草集新七白美白嫩肤面膜

cosmetic

Extracting AVP knowledge from the structured or semi-structured web pages using patterns

歌手 > 华语女歌手 > 那英 > 爱是一颗幸福的子弹

爱是一颗幸福的子弹

高品质

播放 添加 下载 收藏

电影《一场风花雪月的事》主题曲

歌手: 那英
所属专辑: 《爱是一颗幸福的子弹》
发行时间: 2013-07-29
所属公司: 海宁银润影业
歌曲标签: 影视原声 电影 内地 国语

分享到

百度音乐, 随时随地听歌: [PC客户端](#) | [iPhone客户端](#)

歌词

爱是一颗幸福的子弹
词曲演唱: 汪峰
编曲: 峦树

music

系列

论坛(645) 点评(76) 评测行情 配件 二手

参考价格: ¥2700 - 9300元

第10名

ThinkPad E430为2012年上市的14寸商务办公产品, Intel 双核/四核处理器及NVIDIA 系列主流显卡, 采用磨砂材质外壳共魅力红、神秘黑、深邃蓝三色可选。共有189款产品。

核心参数

详细参数 点评 我来点评

网友点评: ★★★★★ 3.5 推荐
优点: 1. 外观设计沉稳大方
2. 配置高的, 运行流畅
3. 键盘手感舒适

屏幕尺寸: 14英寸1366x768
CPU型号: Intel酷睿i53210M ... (19种)
CPU主频: 2.5GHz ... (8种)
内存容量: 2GBDDR31600MHz ... (6种)
硬盘容量: 500GB5400转 ... (6种)
显卡芯片: NVIDIA GeForce GT635M + Intel ...
操作系统: Windows 864bit (64位) 简体中文
摄像头: 集成摄像头

评测资讯 更多

北京行情 切换城市 更多

评测 商务入门之选 联想ThinkPad E430评测
评测 ThinkPad E430系列评测图解
视频 联想ThinkPad E430笔记本电脑视频介绍

610M独显配i5 ThinkPad E430促销中 07-18
入门级首选! ThinkPad E430仅3399元 07-17
送原装鼠标 17高配ThinkPad E430促销 07-13

computer

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	<u>Dream of Red Mansions</u>

Link the name to the correct entity ID



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

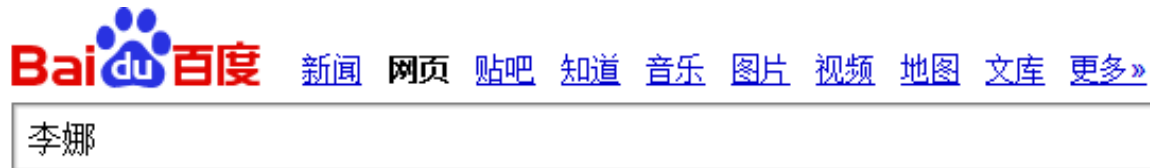
Semantic Computation (cont.)

- 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