



清华大学
Tsinghua University

信息获取与知识图谱

清华大学计算机系
朱小燕

zxy-dcs@tsinghua.edu.cn, @朱小燕THU

2013.10.12



Content



1

- Information & Knowledge

2

- Challenges

3

- Achievements



Information & Knowledge



What do we use the internet for?

INFORMATION AQUISITION

Access the Internet
for getting news,
URL, etc.



INFORMATION SERVICE

**Build knowledge
graph for
generating answers**



Information Services



- Baidu Box Computing

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

凤凰传奇 百度一下 推荐: 用手机随时随地地上百度

Google tsinghua university

Web Images Maps Shopping News More Search tools

About 2,720,000 results (0.26 seconds)

[Tsinghua University](#)

[www.tsinghua.edu.cn/publish/then/](#) - Cached

A friend called to tell me that Building 9 of **Tsinghua University** is going to be torn down, replacing the old with the new. This is truly a good thing, but after ...

[International Students](#) - [Schools & Departments](#) - [General Information](#) - [Contacts](#)

[清华大学- Tsinghua University](#)

[www.tsinghua.edu.cn/](#) - Cached - [Translate this page](#)

招生信息、院系设置、新闻动态。

Score: 25 / 30 - 19 Google reviews



China, 北京市海淀区 清华大学
+86 10 6279 3001

[院系设置](#) - [招生信息](#) - [研究生招生](#) - [清华大学图书馆](#)

[Tsinghua University - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Tsinghua_University](#) - Cached

Tsinghua University (Chinese: 清华大学, qīnghuá dàxué), is a university in Beijing, China. The school is one of the nine universities of the C9 League.

[History - Present](#) - [Schools and departments](#) - [Department of Mathematical ...](#)

You've visited this page 2 times. Last visit: 3/10/11

[Tsinghua University - Topuniversities](#)

[www.topuniversities.com/institution/tsinghua-university](#) - Cached

The campus of **Tsinghua University** is situated on the former imperial gardens of the



Tsinghua University

[Directions](#)

Tsinghua University, is a university in Beijing, China. The school is one of the nine universities of the C9 League. [Wikipedia](#)

Address: China, 北京市海淀区清华大学

Enrollment: 37,650 (2010)

Founded: 1911

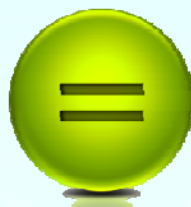
Colors: White, Purple

People also search for

Knowledge

- What is knowledge?
 - Entity + Rule
 - Triple (predicate, semantic network, rule, framework,)

Structured
Information



KNOWLEDGE



Abstracted Information

Coded Information

**Drawback: Lack of ability to
communicate with machine
DIRECTLY!**

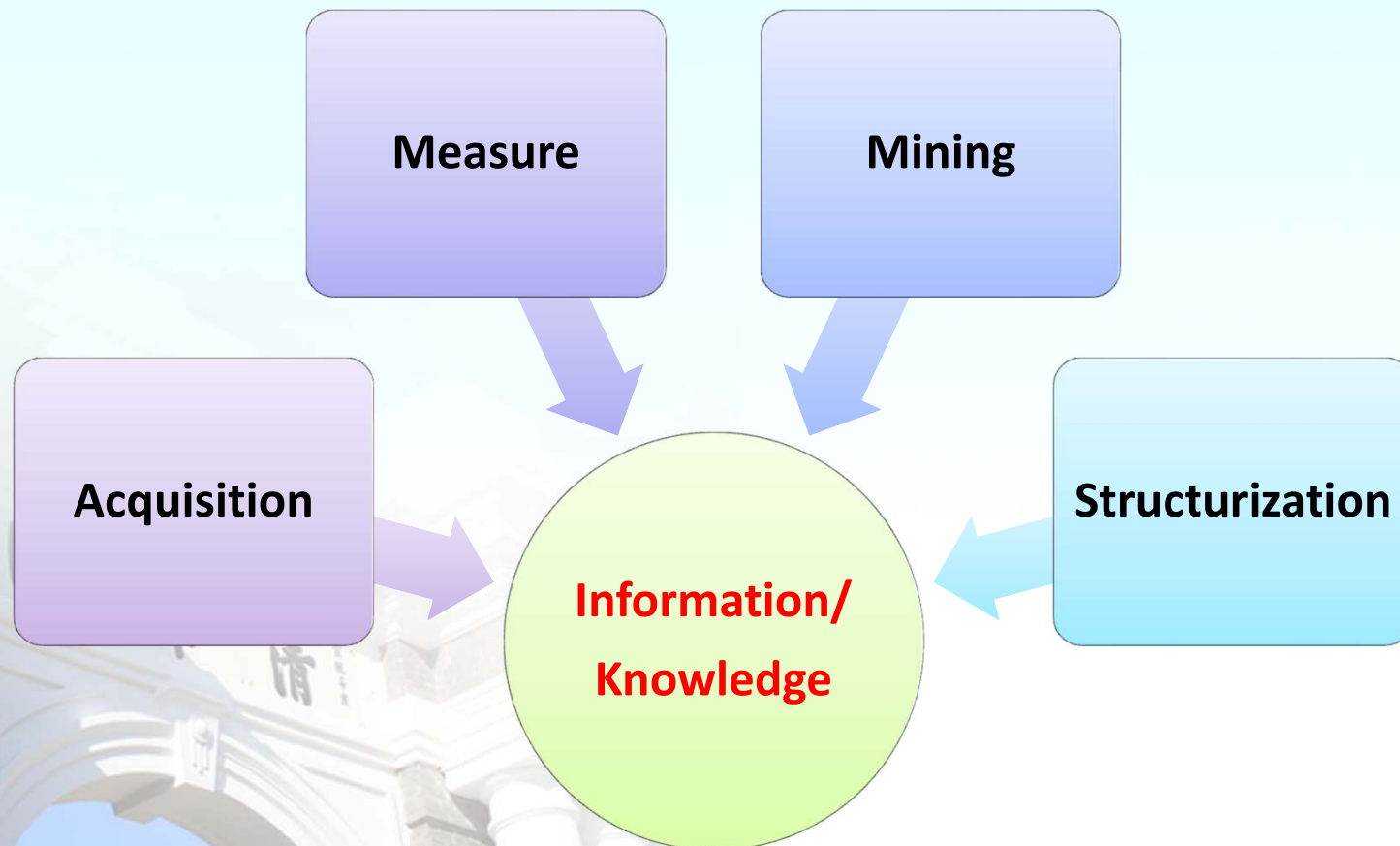
知识的作用：帮助信息计算、理解、评价

Challenges



Scientific Perspective

- **Our Goal**
 - **Information Measurable**
 - **Knowledge Computable**



Application Perspective

Information acquisition

Knowledge graph

- How to build knowledge bases
 - To build from original materials
 - To extend and update
 - To merge difference databases
 - To verify the knowledge
- How to use knowledge bases
 - For answer generation
 - For answer re-ranking
 - For inference
 - For vertical search
 -

Achievements



Achievements (1)

- Information acquisition
 - Information metrics and its applications
 -



Search with Key Words

Key Words



URL



Refine Key Words



Inspect



Information



Information Acquisition Platform



Application Layer

Complex
QA

Vertical
Search

Enterprise
Search

Computational
Advertisement

Semantic Layer

Content
Understanding

User
Understanding

Sentiment
Understanding

Analysis Layer

Concept
extension

Semantic
relatedness

Semantic tagging

User interest modeling

Emotion
analysis

Opinion
extraction

Similarity
metric

Question type
classification

Focus
extraction

Answer
typing

Authority/Expert modeling

Opinion
summarization

Sentiment
classification



EXTERNAL SOURCES

Chunking

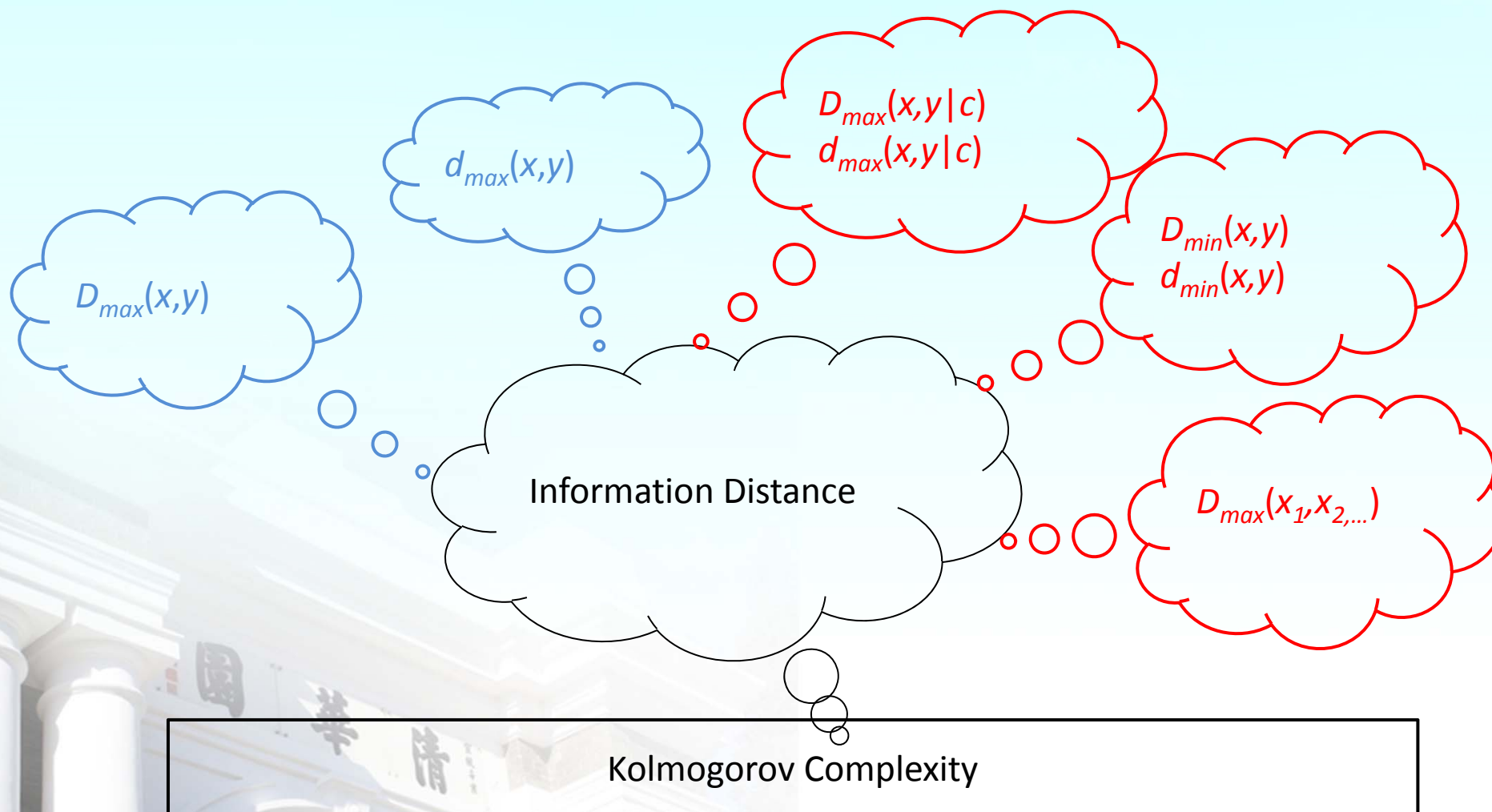
POS tagging

Tokenization

Parsing

NLP TOOLS

Information Distance



Main publications

- Answer re-ranking
 - KDD 2007
- Concepts measure, relatedness evaluation
 - COLING 2010 **BEST PAPER**, IJCAI 2011
- Question answer pairs similarity measure
 - ACL 2012 **BEST STUDENT PAPER**
- Multiple document summarization
 - CIKM 2008, ICDM 2009
- 开放领域问答系统平台 – 趣答

Achievements (2)

- Knowledge base construction
 - Chinese Knowledge Base Construction
 - Information Organization for Multi-source UGCs via Topic Hierarchy Construction



- Semi-structured text
 - Tables, info-boxes, etc.
- Free text

- Open knowledge bases
 - Freebase, Yago, DBPedia, etc.
- Inter-language transfer
 - Wikipedia Multi-language linker
 - Google translation



A portrait of Barack Obama, smiling, with the word "Topic" written above him.

```

/m/02mjmr /common/name Barack Obama↵
/m/02mjmr /common/notable type /government/us_president↵
/m/02mjmr /biology/animal_owner/animals_owned↵
/m/05t073s↵
/m/02mjmr /biology/animal_owner/animals_owned↵
/m/02xv_y1↵
/m/02mjmr /people/person/nationality /m/09c720↵
/m/02mjmr /people/person/gender /m/05zppz↵
/m/02mjmr /people/person/place_of_birth /m/02hrh0 ↵
/m/09c720 /common/name United States of America↵
/m/09c720 /location/dated_location/date_founded↵
7/4/1776↵

```



From Wikipedia, the free encyclopedia

贝拉克·奥巴马

维基百科，自由的百科全书

Knowledge Base Construction

- **Knowledge Base**

- Knowledge structure

- Triple

- entity
 - relation

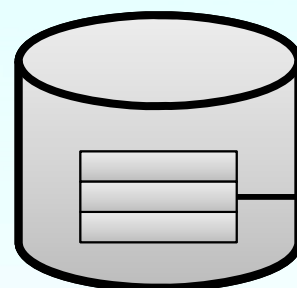
- Size of depository

- Resources

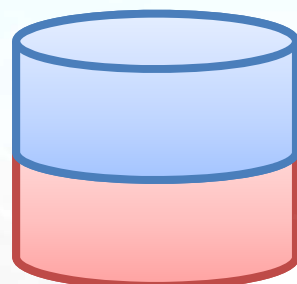
- Baidu Baike
 - Freebase
 - Wikipedia

- Contains:

- **250,000** entities
 - **1660,000** triples



Subject	Relation	object
.....		
巴拉克 奥巴马	拥有国籍	美利坚合众国
.....		



Extracted From Baidu Baike:
120, 000 entities, **650,000** triples

Transfer From Freebase:
130, 000 entities, **1000,000** triples

Information Organization for Multi-source UGCs via Topic Hierarchy Construction

- **Multi-source UGCs**

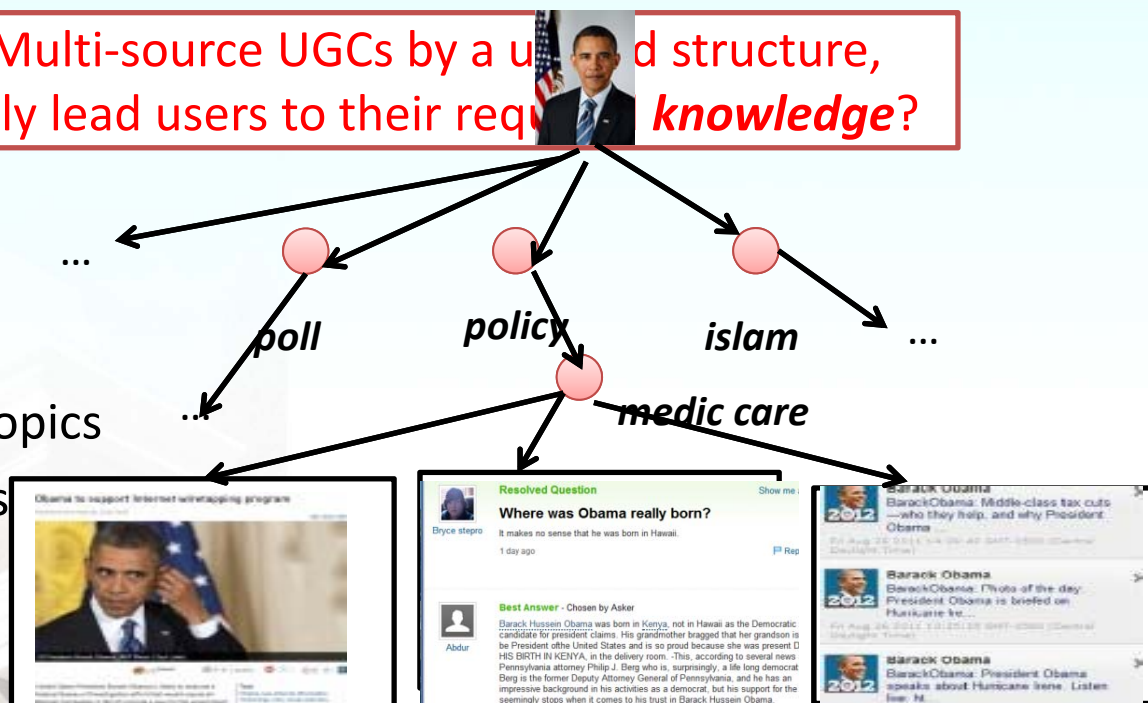
- Quality of Contents
- Power of Statistics
- Authority, timeliness, etc..



Can we organize Multi-source UGCs by a unified structure, which can effectively lead users to their required **knowledge**?

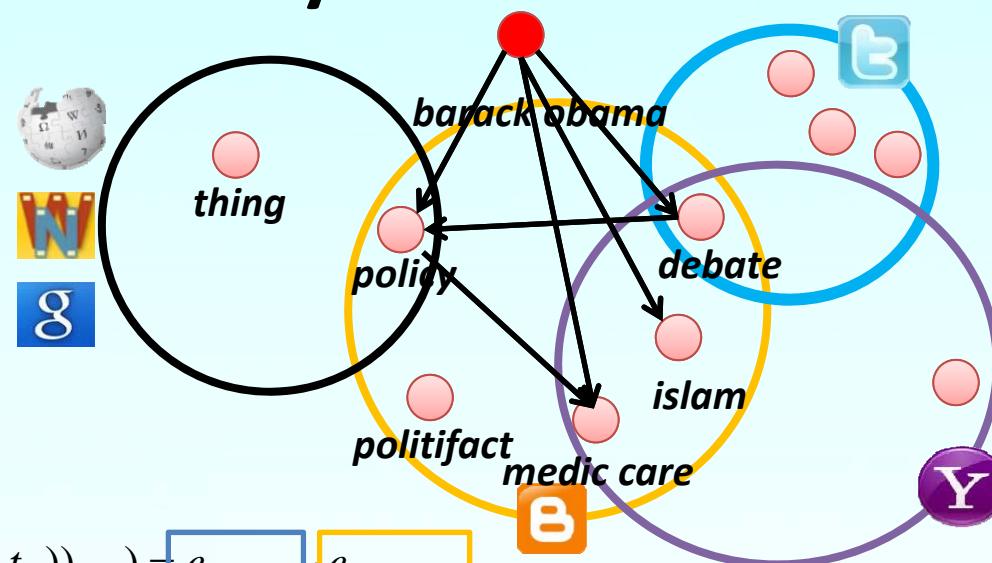
- **Topic Hierarchy**

- root node: user topic
- non-root node : sub topics
- leaf node: link to UGCs



Information Organization for Multi-source UGCs via Topic Hierarchy Construction

- **Topic extraction**
 - Keyword extraction
 - Hyponym mining
- **Sub-topic relation**



$$p(r(t_A, t_B)) \propto F(e_1(r(t_A, t_B)), e_2(r(t_A, t_B)), \dots) = e_{\text{direction}} \cdot e_{\text{relatedness}}$$

$$e_{\text{direction}}(r(t_A, t_B)) = \sum_{e_k \in E_{\text{direct}}} w_k \cdot e_k(t_A, t_B)$$

$$e_{\text{relatedness}}(r(t_A, t_B)) = \prod_{e_s \in E_{\text{undirect}}} e_s(t_A, t_B)$$

directed-evidences	Source
$e_{\text{pattern0}} \sim e_{\text{pattern5}}$	Search engine
$e_{\text{wiki_title}}$	Wikipedia
$e_{\text{wiki_cate}}$	Wikipedia
e_{wnet}	WordNet

undirected-evidences	Source
$e_{\text{dis_doc}}$	crawled UGCs
$e_{\text{dis_sen}}$	crawled UGCs
$e_{\text{wiki_pmi}}$	Wikipedia

Information Organization for Multi-source UGCs via Topic Hierarchy Construction

- **Topic Organization**

- Depth vs. relatedness
- Real-time update

- **Topic Hierarchy Construction**

- Via iteration
- Each iteration:

- Add a new topic t to the current hierarchy H :

$$t = \arg \max_{t_s \in T - T_{i-1}} \sum_{t_k \in T_{i-1}} (w(r(t_k, t_s)) + w(r(t_s, t_k)))$$

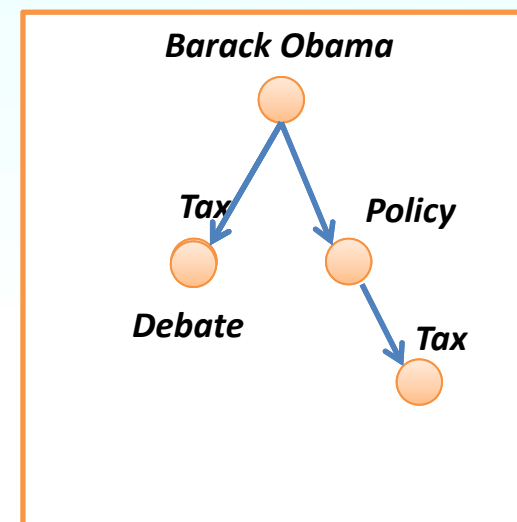
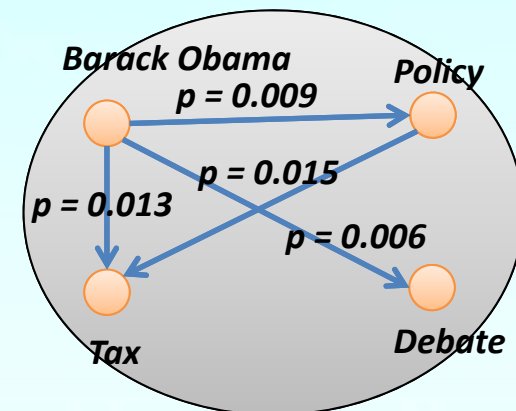
- Update the weights of nodes and edges on H :

$$w_t(t_k) = \sum_{t_g \in T_G} w(r(t_{root}, t_g)) \cdot w(r(t_k, t_g))$$

$$w_r(t_s \rightarrow t_k) = \max_{\substack{L \text{ ends with } t_s \rightarrow t_k}} \sum_{u=0}^{|L|-1} w_t(t_u) \cdot w(r(t_u, t_{u+1}))$$

- Remove the potential cycles on the hierarchy:

$$H = \text{Optimum_Branching}(H')$$



Resultant Hierarchy

Contribution

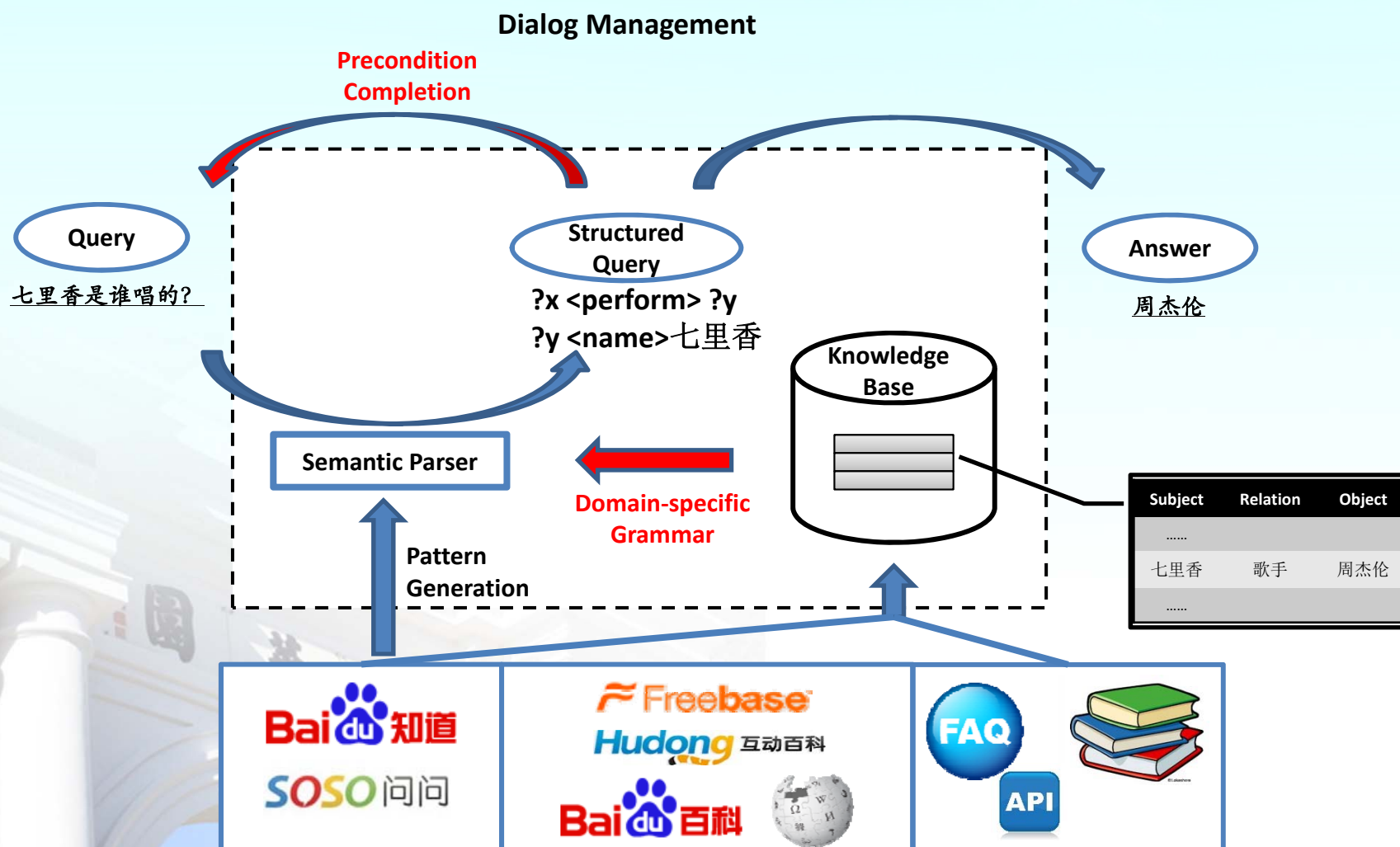
- Internet information organization by topic trees
- Propose the algorithm of topic hierarchy construction, outperform the state-of-art algorithms
- Public in SIGIR 2013



Achievements (3)

- Vertical Search Platform
 - **Unified Knowledge Representation**: Triples (compatible with Freebase)
 - **General Information Processing Pipeline**
 - High reusability of most modules
 - Easily and rapidly portable to different vertical domains if provided enough domain data
 - **Accommodating Heterogeneous Source of Data**
 - Knowledge Base
 - CQA
 - FAQs
 - Query logs (scripts from the mobile company)
 - Encyclopedia (Wikipedia, Baidu Baike)
 - Free texts, books
 - APIs

Vertical Search Platform



Applications

微信公共账号

- Public health care (Database + CQA)

公众健康问答

- Music search (Database)

音乐问答

- Mobile services (Query log)

业务助手

- Weather (CQA + APIs)

天气自动问答THU

- Open domain

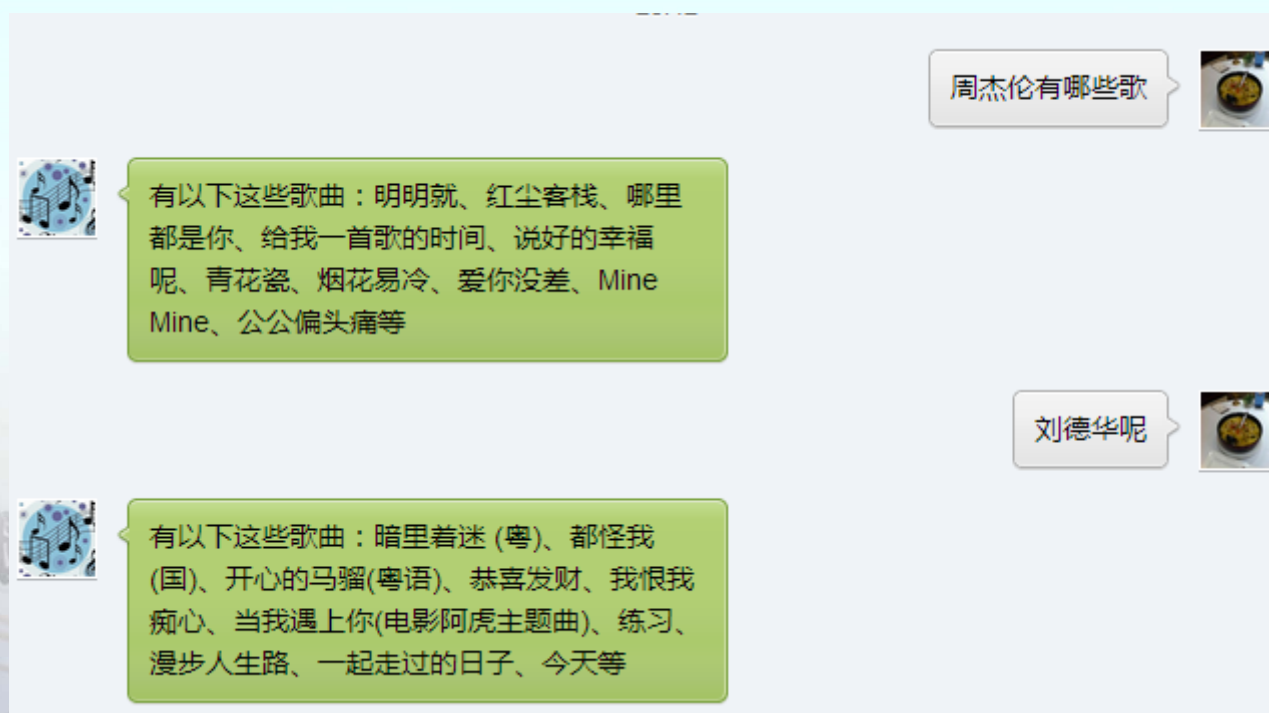
清华小智

- Leaks & Exploits (Free text + Domain knowledge/rules)

- College Recruitment (FAQ)

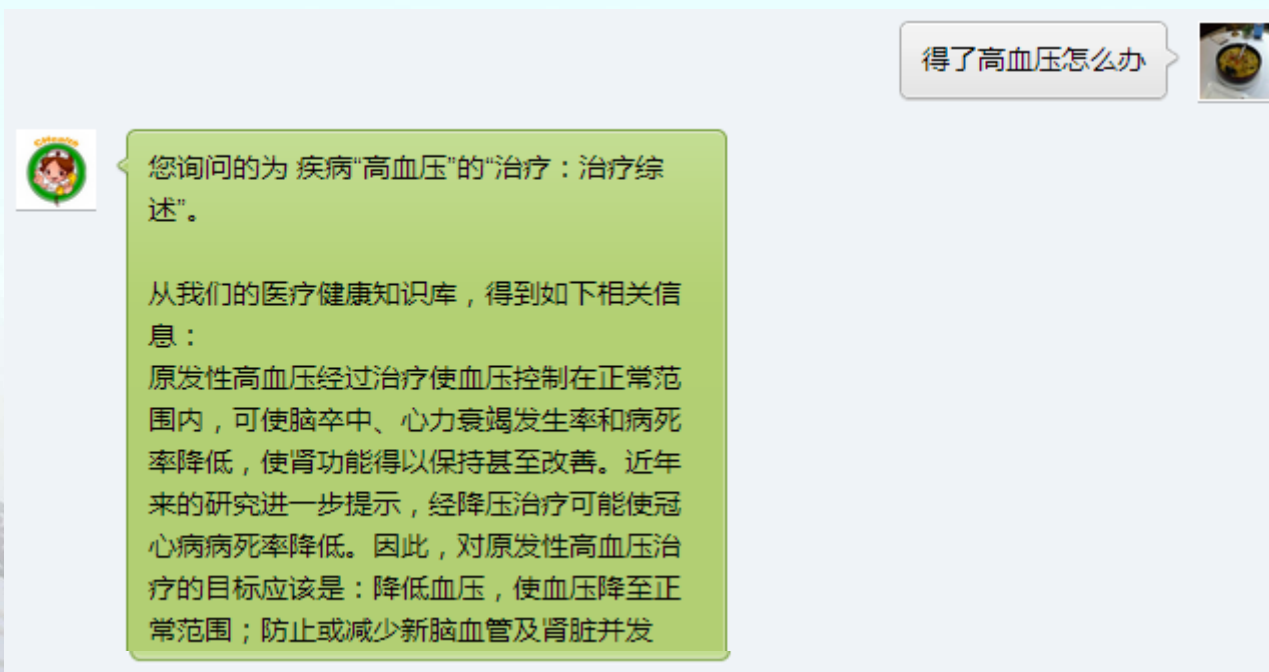
Examples of the Application (1)

- Music QA
 - Resource: Domain Database, CQA



Examples of the Application (2)

- Health QA
 - Resource: Domain Database, CQA, Baidu Baike



Main Works



- **Information distance metrics**, Xian Zhang, Chong Long, Fan Bu, et al, SIGKDD2007, ICDM2009, MI2009, COLING 2010 (best paper) ,
- **Question classification**, Fan Bu, et al, EMNLP2010
- **Question expansion**, Zhicheng Zheng, et al, NAACL 2010
- **Concept relatedness evaluation**, Fan Bu, et al, IJCAI2011
- **Passage retrieval based concept attribute extraction**, Chao Han, et al, CICKing2010
- **Question and answer pair mining**, Shilin Ding, Fan Bu, et al, ACL2008, ACL 2012 (best student paper)
- **Text summarization**, Minlie Huang, et al, ACL2010, AAI2012, CIKM 2008,ICDM 2009
- **Opinion mining**, Fangtao Li, et al, AAI2010, IJCAI2011, COLING2010
- **Information recommendation**, Lijing Qin, Yang Tang, et al, JICAI 2013, SIGIR 2011 workshop on “entertain me”
- **Information extraction**, Xingwei Zhu, et al, SIGIR 2013



Main Publications

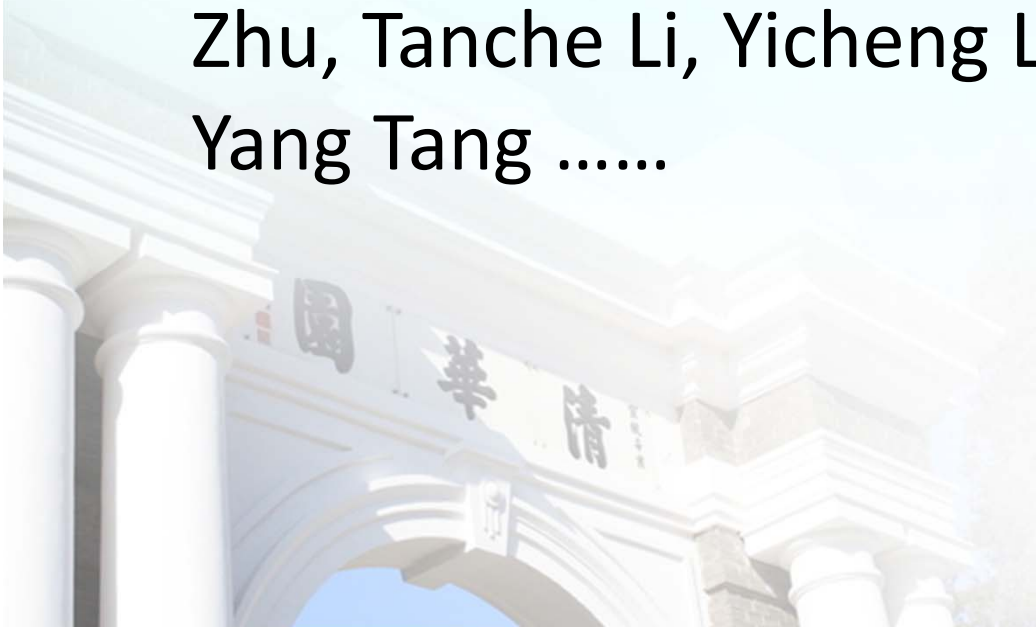


- Natural language processing
 - ACL 2008, 09, 10, 11, 12, **ACL 2012 Best Student Paper Award**, **COLING 2010 Best Paper Award**, EMNLP 2010, NAACL2010,
- Artificial Intelligence
 - AAI 2010, 2012, IJCAI 2011,
- Data Mining
 - SIGKDD 2007, ICDM 2008, 09, 10, PAKDD 2007, WI 2009, CIKM 2006, 08, 12,



Acknowledgement

- Prof. Ming Li and Minlie Huang,
- Dr. Yu Hao,
- Mr. Xian Zhang, Chong Long, Fan Bu, Hao Xiong, Chao Han, Zhicheng Zheng, Xingwei Zhu, Tanche Li, Yicheng Liu, Yipeng Jiang, and Yang Tang



Thanks

