清華大學
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

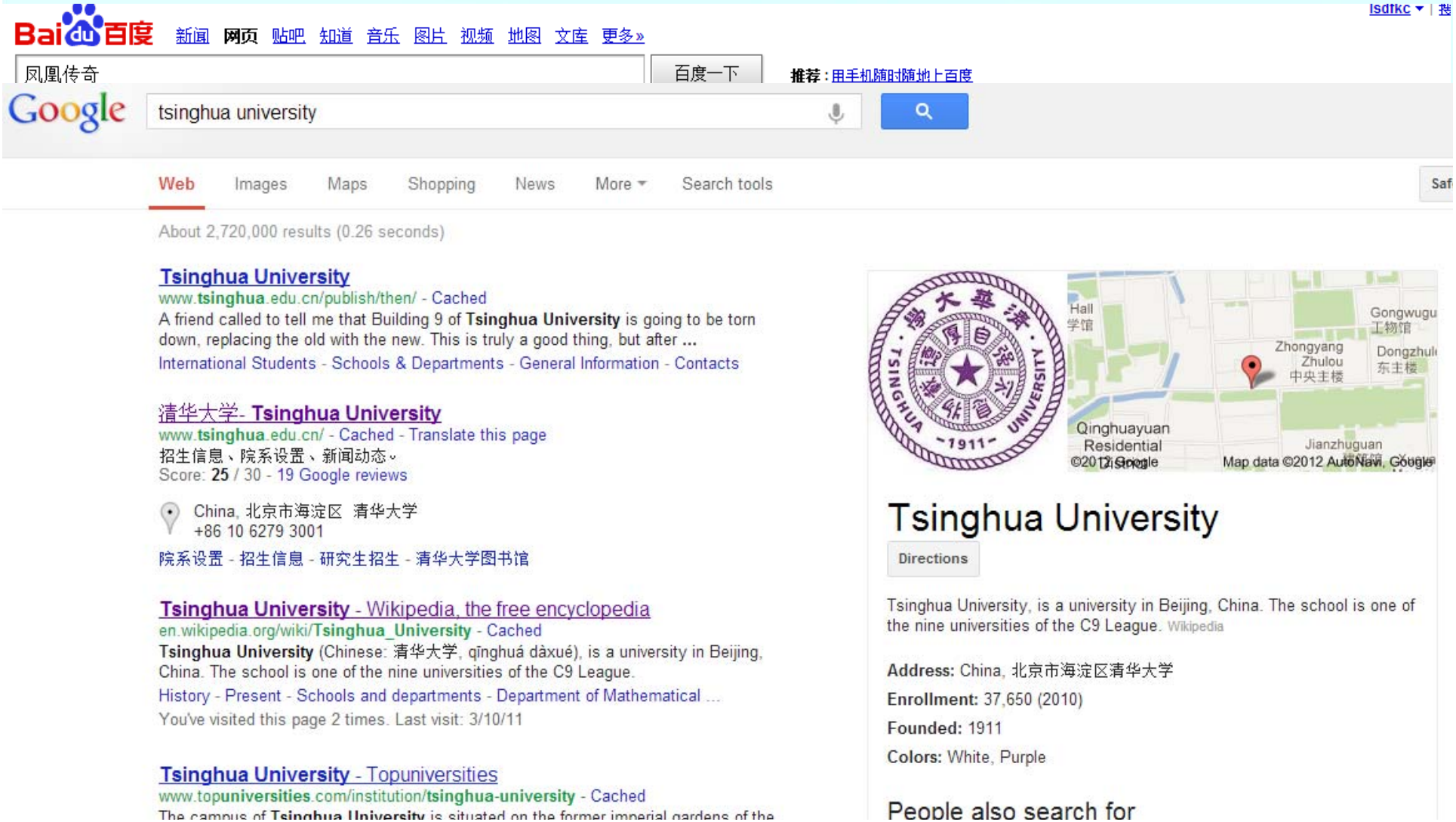
Access the Internet for getting news, URL, etc.

INFORMATION SERVICE

Build knowledge graph for generating answers

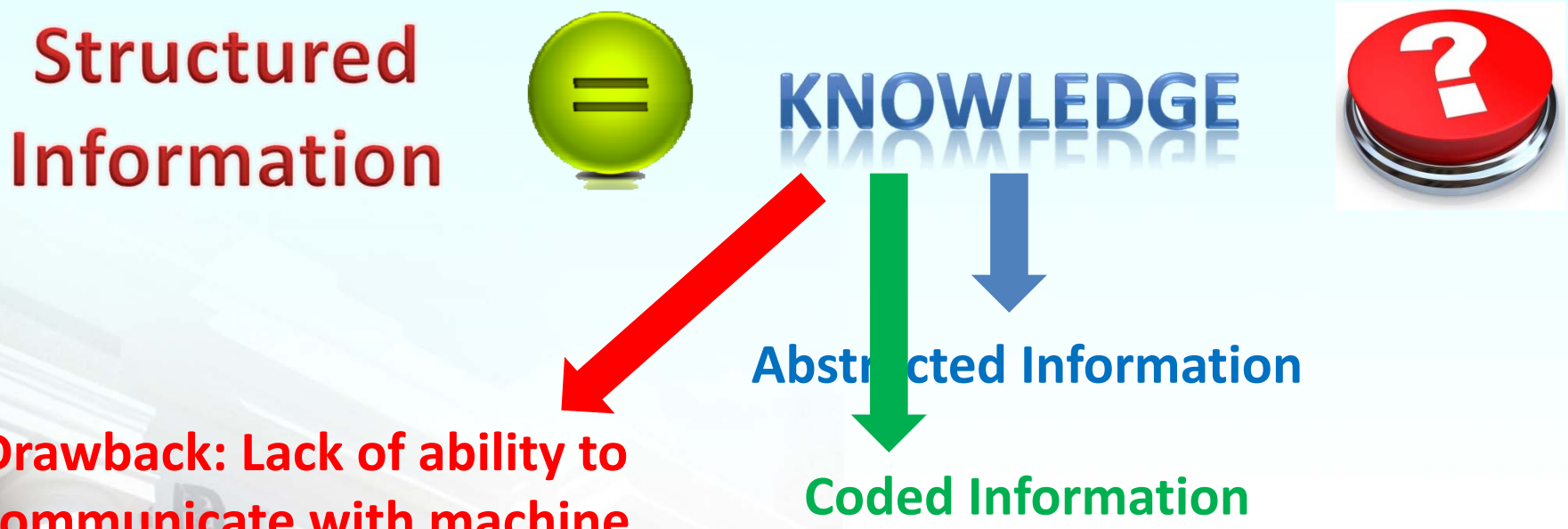
Information Services

- Baidu Box Computing



Knowledge

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



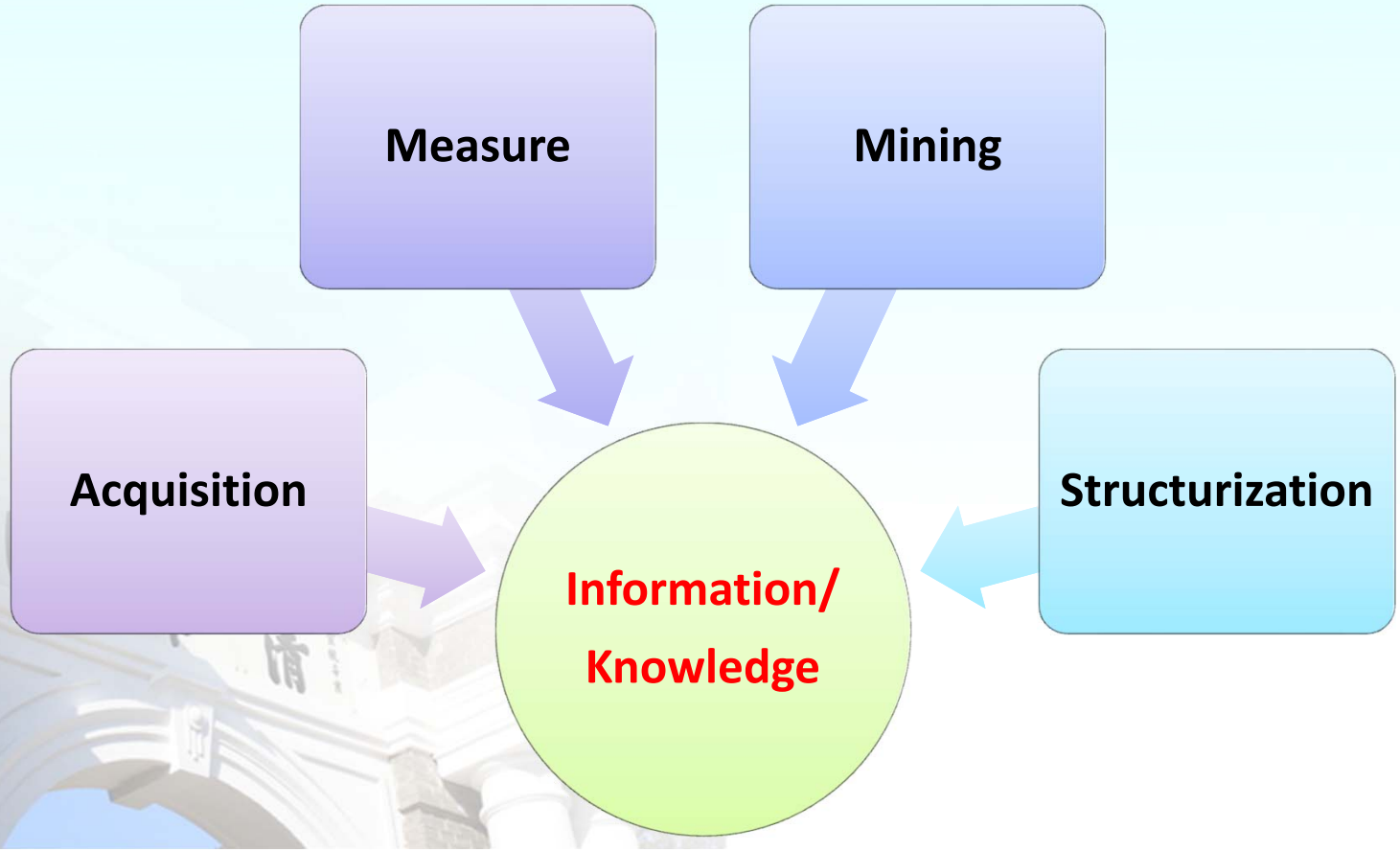
知识的作用：帮助信息计算、理解、评价

清華大學
Tsinghua University

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

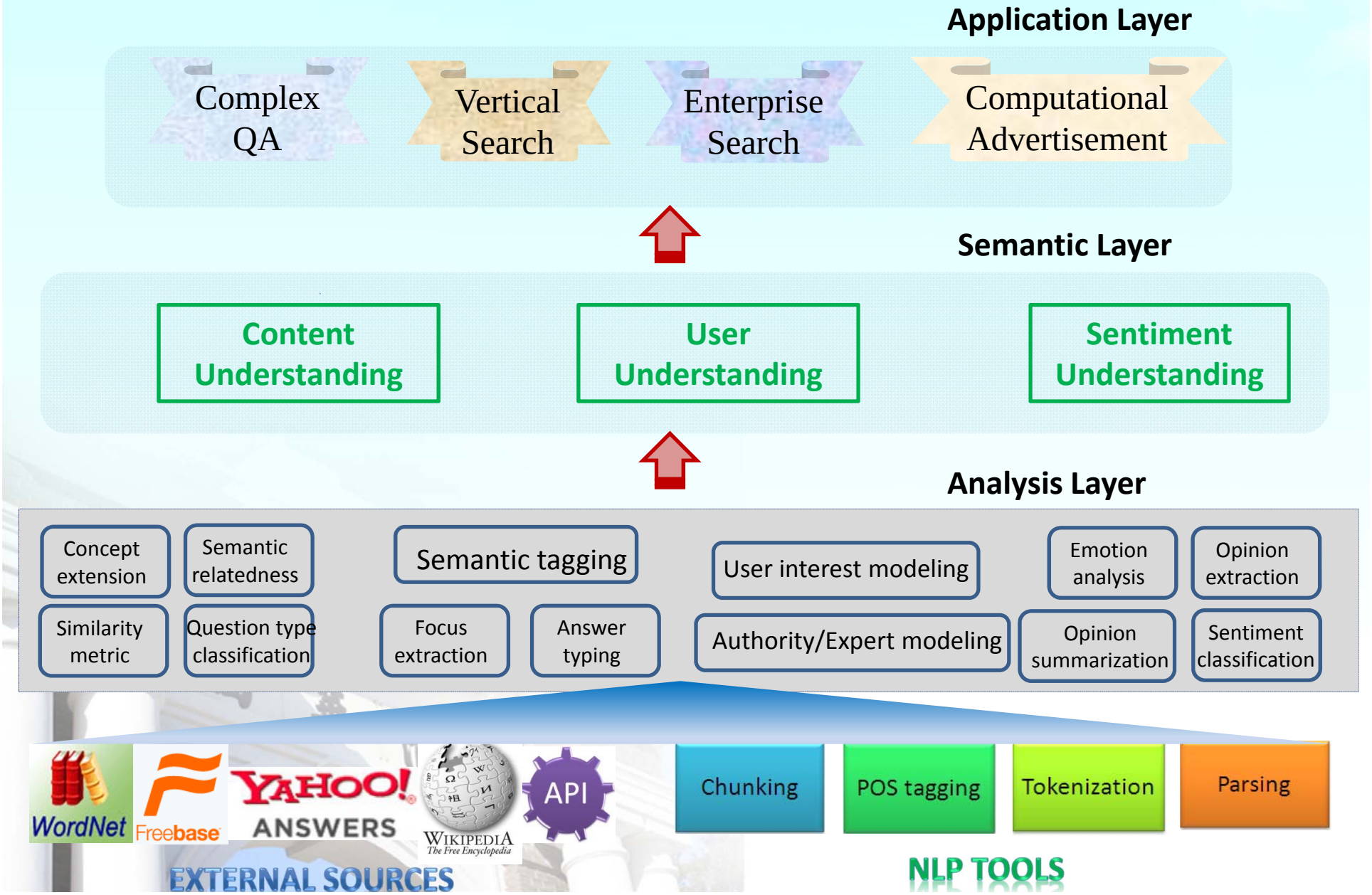
URL

Information

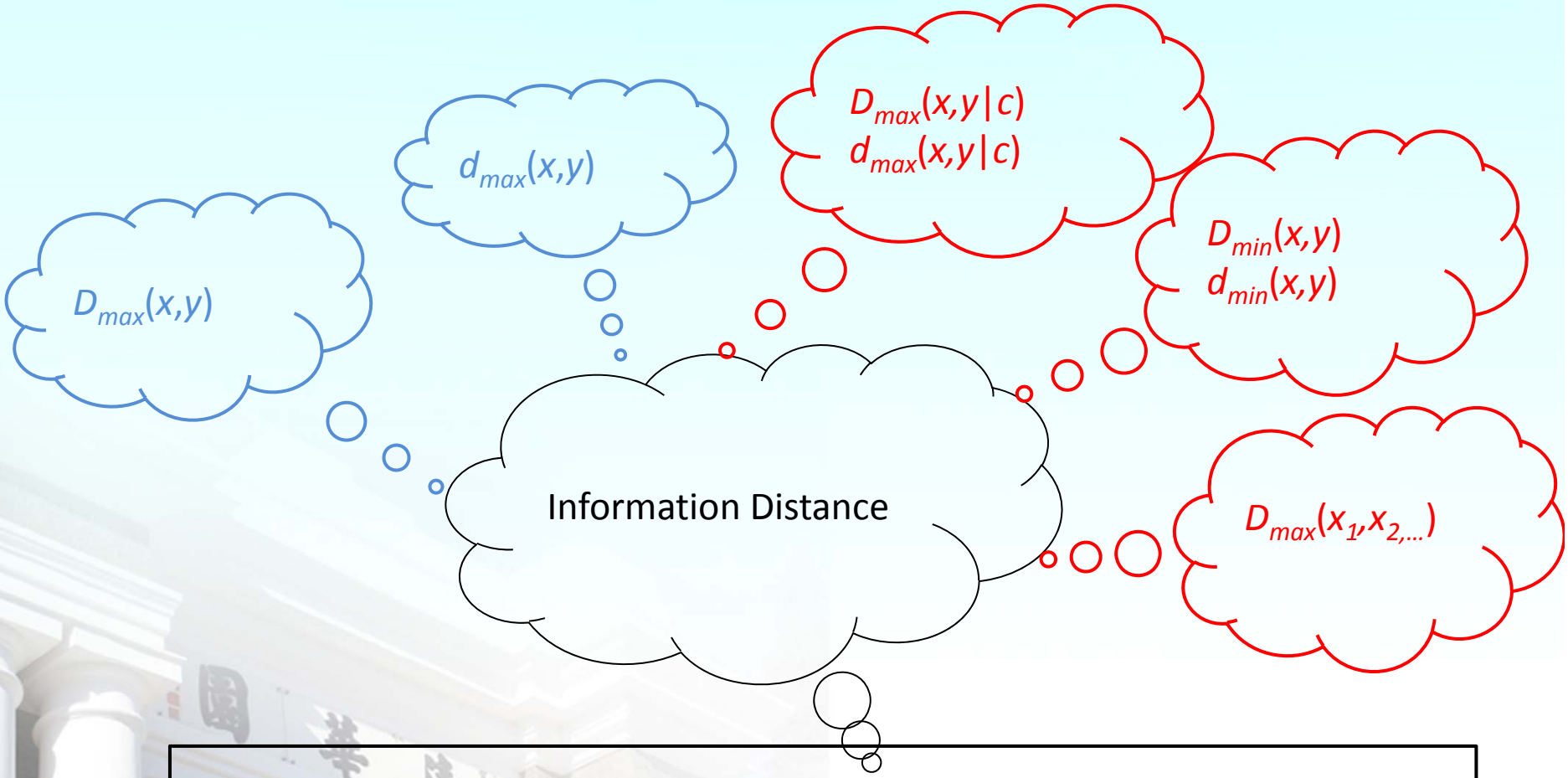
Inspect

Refine Key Words

Information Acquisition Platform



Information Distance



Kolmogorov Complexity

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

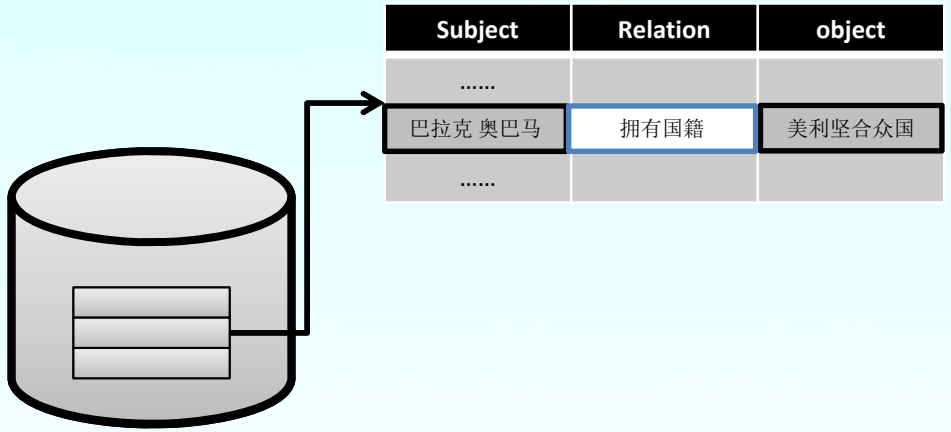
Knowledge Base Construction

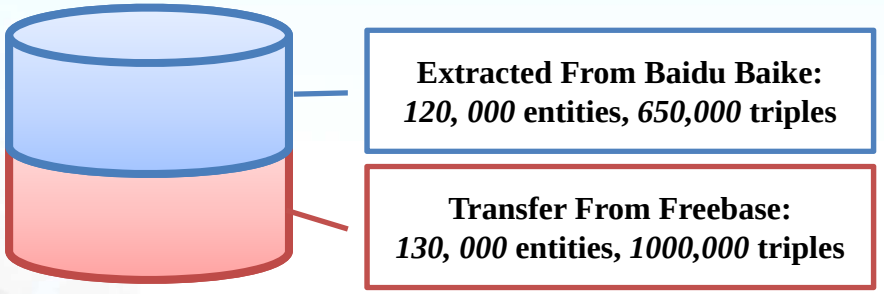
- **Knowledge Extraction**
  - Semi-structured text
    - Tables, info-boxes, etc.
  - Free text

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

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



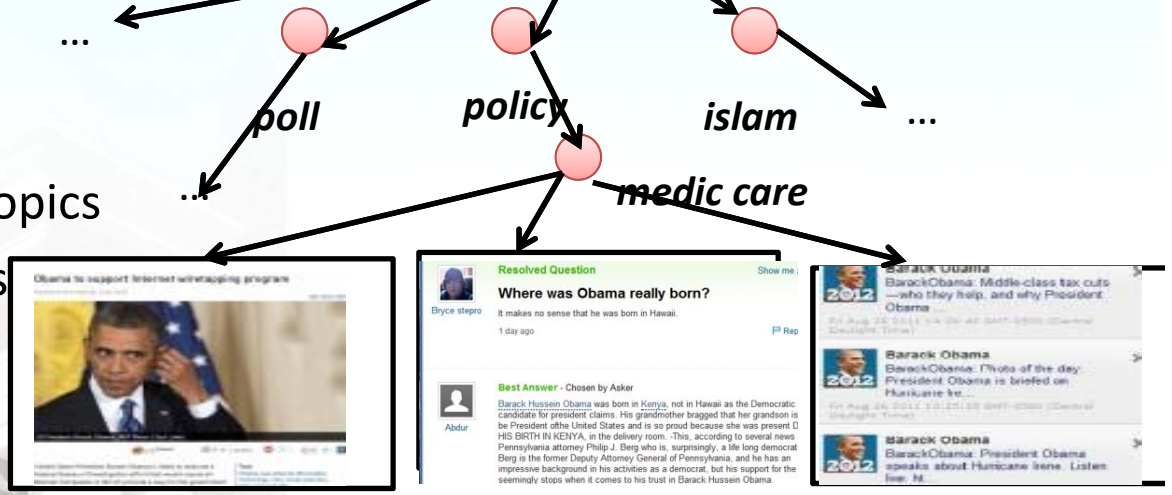


Information Organization for Multi-source UGCs via Topic Hierarchy Construction

- **Multi-source UGCs**
  - Quality of Contents
  - Power of Statistics
  - Authority, timeliness, etc..

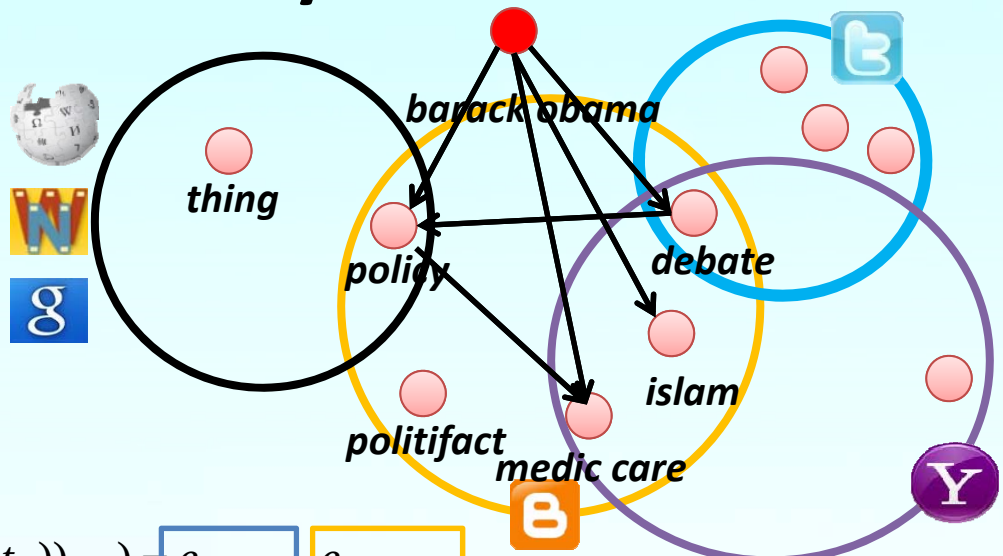
Can we organize Multi-source UGCs by a u d structure, which can effectively lead users to their requ ***knowledge***?

- **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**



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

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

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

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

undirected-evidences	Source
e_{dis_doc}	crawled UGCs
e_{dis_sen}	crawled UGCs
e_{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:

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

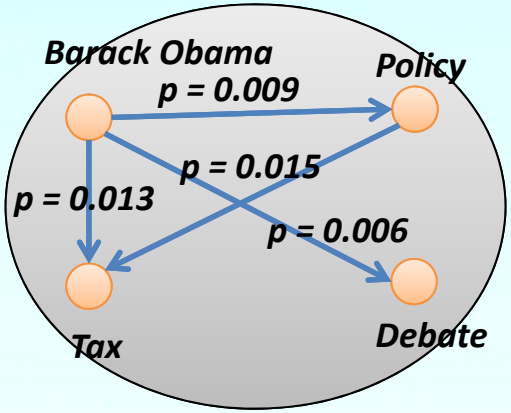
    - Update the weights of nodes and edges on H:

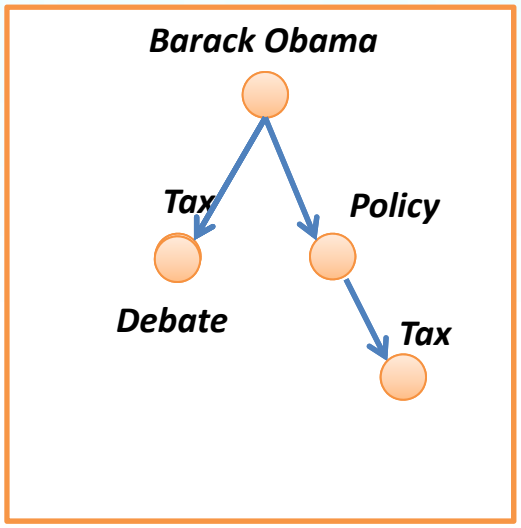
$$w_t(t_k) = \sum_{t_g \in T_G} w(r(t_{root}, t_g)) \cdot w(r(t_k, t_g))$$

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

    - Remove the potential cycles on the hierarchy :

$$H = Optimum\_Branching(H')$$





Resultant Hierarchy

Contribution

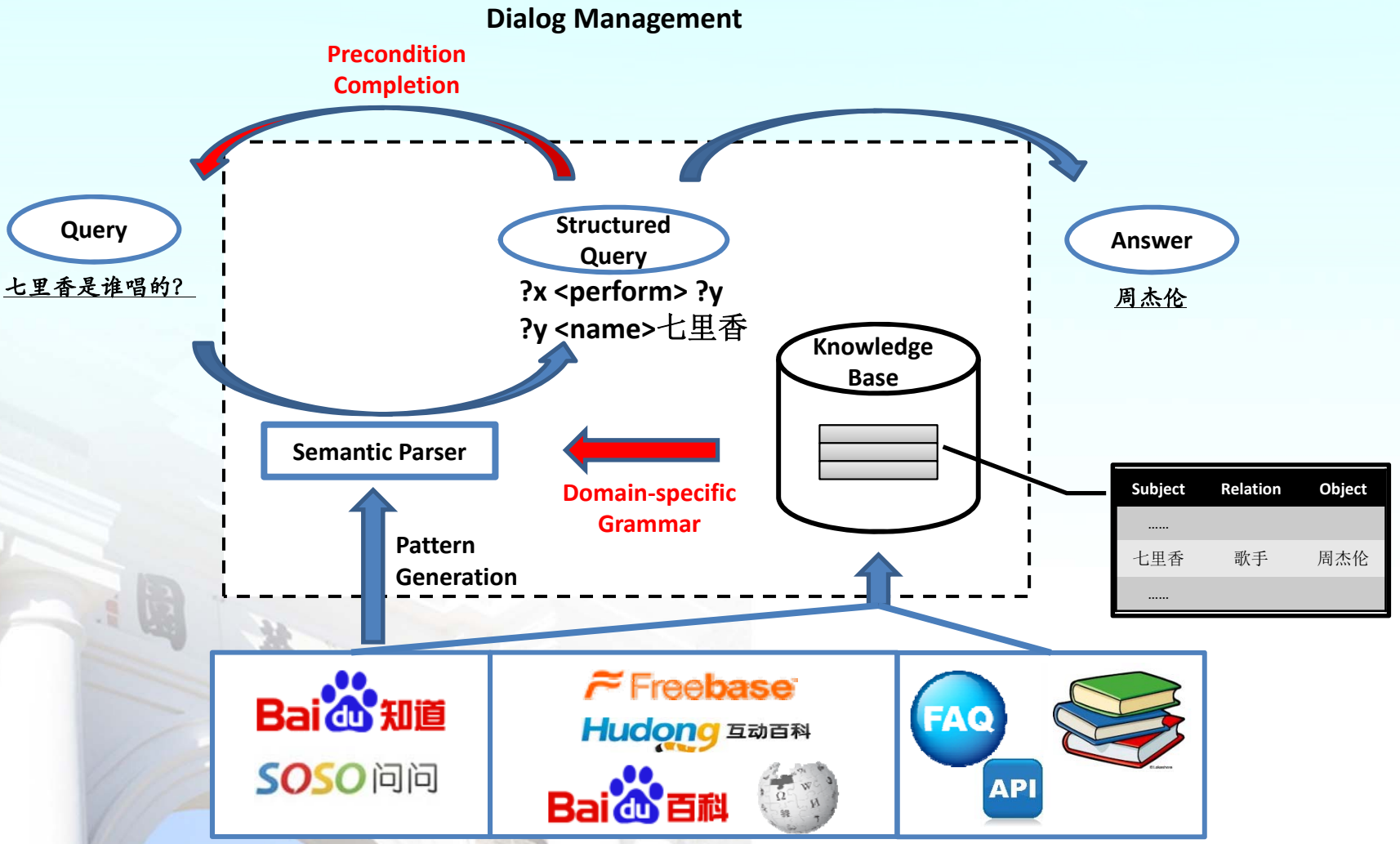
- Internet information organization by topic trees
- Propose the algorithm of topic hierarchy construction, out perform the state-of-art algorithms

- 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



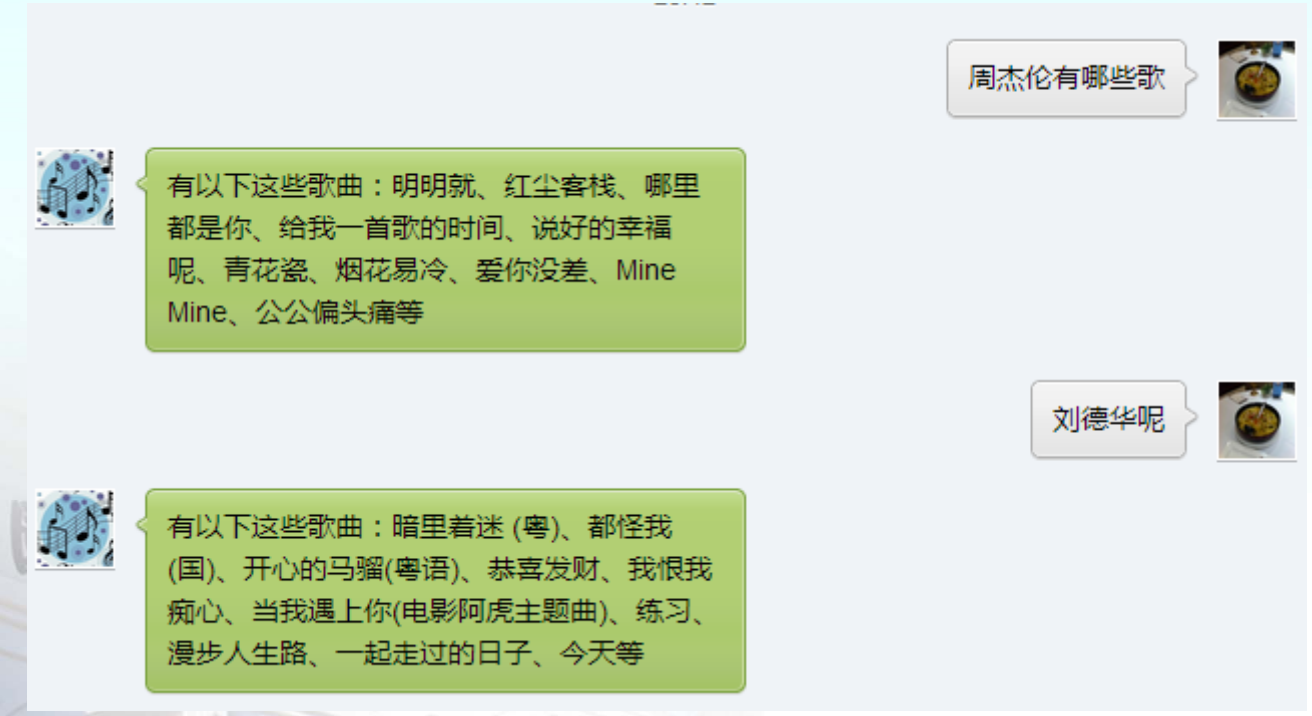
Applications

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）

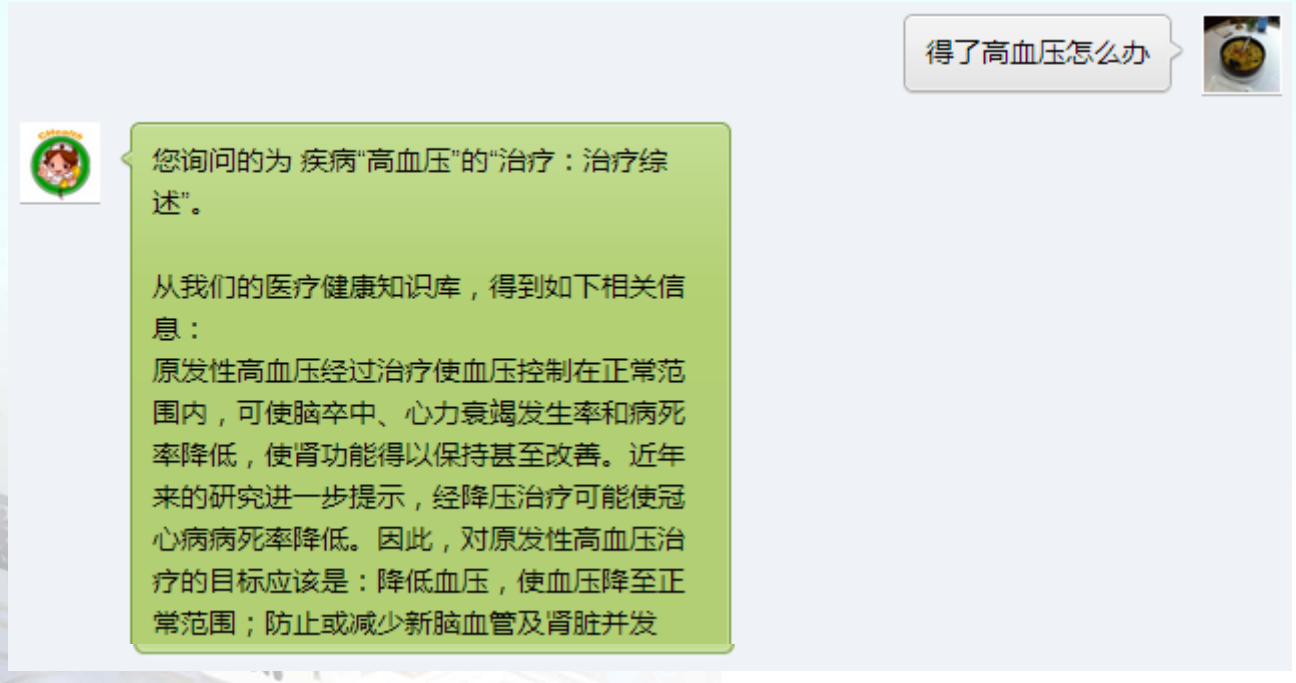
Examples of the Application （1）

- 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, CICLing2010
- Question and answer pair mining, Shilin Ding, Fan Bu, et al, ACL2008, ACL 2012（best student paper）
- Text summarization, Minlie Huang, et al, ACL2010, AAAI2012, CIKM 2008,ICDM 2009
- Opinion mining, Fangtao Li, et al, AAAI2010, 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
  - AAAI 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 ……

清華大學
Tsinghua University

Thanks