



# 基于低维表示的大规模实体关系挖掘技术

Large-scale **entity relation extraction** based on *low-dimensional representations*

范淼

直博三年级

语音和语言技术中心

清华大学计算机系

指导教师：郑方、周强

[fanmiao.cs@thuis.com](mailto:fanmiao.cs@thuis.com)

# 目录

|   |        |
|---|--------|
| 1 | • 领域综述 |
| 2 | • 研究动机 |
| 3 | • 科学问题 |
| 4 | • 研究意义 |
| 5 | • 研究计划 |
| 6 | • 研究进度 |
| 7 | • 参考文献 |



# 1. 领域综述

- **信息抽取 (Information Extraction)** 在自然语言处理领域有长达近20年的研究历史，始终致力于将无结构化的文本转换为结构化的信息，而有便于给诸如：**问答系统 (Question-Answering System)**、**信息检索 (Information Retrieval)** 等其他应用领域提供更加便利的知识表示。

Google 清华大学

网页 图片 地图 新闻 视频 更多 搜索工具

找到约 4,470,000 条结果 (用时 0.47 秒)

**清华大学 - Tsinghua University**  
www.tsinghua.edu.cn/

6月16-20日，清华大学-哈佛大学燕京学社比较文学和世界文学研讨班在清华大学举办。研讨班的主要论题包括生态批评研究，全球健康问题，文学与全球化城市。翻译...

4.6 ★★★★★ 83条 Google 评价 写评价

中国北京市海淀区双清路30号  
+86 10 6279 3001

清华大学的新闻搜索结果

双胞胎姐妹学霸将一起入读清华大学(图)  
新浪网 · 9小时前  
傅浩怡说，她报考了清华大学社会学专业，妹妹傅浩悦则报考了清华大学人文科学专业，姐妹俩从上学以来一直在同一个班，这次希望分开一下了。

广东惠州双胞胎姐妹学霸入读清华大学无悬念  
新浪网 · 13小时前

双胞胎姐妹学霸将一起入读清华大学  
新华网 · 14小时前

更多关于“清华大学”的新闻

清华大学 - 维基百科，自由的百科全书  
zh.wikipedia.org/zh/清华大学

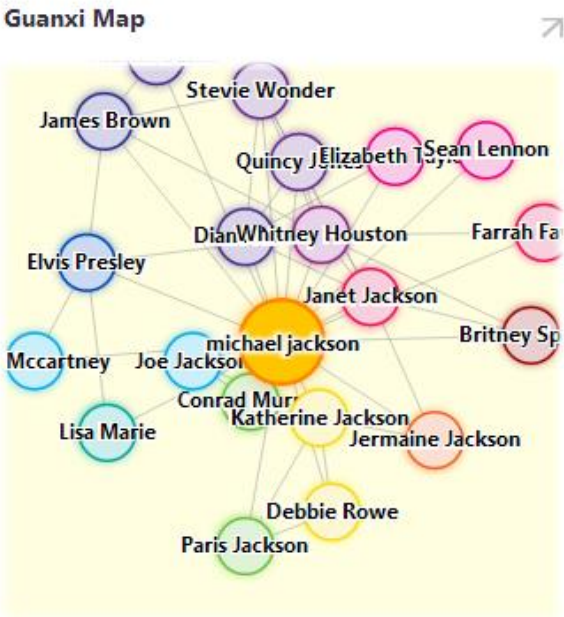
清华大学（英语：Tsinghua University，简称：清华、THU），简称清华学堂、清华学校、国立清华大学，是一所位于北京市海淀区清华园的公立高等院校。最初为1911年清政府利用美国退还的部分庚子赔款所建留美预备学校“游美学务处”及附设“肄业馆”，1925年始设大学部。抗日战争爆发后，清华与北大、南开南迁长沙，组建国立长沙临时大学。 维基百科

地址：中国北京市海淀区双清路30号  
招生人数：43,112 (2013年12月)  
电话号码：+86 10 6279 3001  
成立于：1911年4月29日  
代表色：白色，紫色

用户还搜索了 还有54项

复旦大学 上海市  
北京大学 北京市  
上海交通大学 上海市  
中国人民大学 北京市  
同济大学 上海市

Google 知识图谱



Microsoft 关系图谱



# 1. 领域综述

- 信息抽取 (Information Extraction) 主要分为命名实体识别 (Named Entity Recognition, **NER**) 和关系抽取 (Relation Extraction, **RE**) 两大任务。
- 命名实体识别主要致力于从无结构的文本中识别人名 (*PER*)、地名 (*LOC*)、机构名称 (*ORG*) 等名词实体，目前技术比较成熟，识别率都在90%上下，目前微软亚洲研究院，聂再清研究员领导的小组持有的识别工具已经投入商用。
- **关系抽取** (实体关系挖掘, RE) 是目前研究的主题，同时也是工业界关注的热点话题。该研究在NER的基础上，用于发现实体之间的关系，目前最受关注的是识别实体对 ( $\langle h(\textit{head\_entity}), t(\textit{tail\_entity}) \rangle$ ) 之间的关系 **r**。

维基百科

Barack Hussein Obama II ([/bəˈrɑːk huːˈsem oʊˈbɑːmə/](#); born August 4, 1961) is the 44th and current President of the United States,

$\langle$ Barack Obama, *President of*, U.S. $\rangle$



# 1. 领域综述

- 实体关系挖掘技术的研究在**2008年之前**分为两种不同的研究方向：
  - 固定关系挖掘
  - 开放关系挖掘 (Open RE)
- 上述两种关系挖掘技术的不同点在于：是否有**新关系 (new relationship discovery)**的发现。
- 学生的研究方向主要关注于**固定关系挖掘**。
- 固定关系挖掘基本假设于我们在**圈定种类**的关系类别中，对实体之间的关系进行**预测**，因此属于**监督学习**范畴 (Supervised Learning based Relation Extraction Approaches)。



# 1. 领域综述

- 2008年之前的关系挖掘的研究大多集中在ACE, MUC两类关系标注语料库中探讨如何利用规则方法、统计监督学习方法不断提升对多类别关系分类（预测）的精度。
- ACE和MUC两个人工标注数据库的规模都比较小。以ACE语料为例，共有大约1000篇文本，包含16771个关系实例，23种关系类型。
  - 代表工作有：
    - 基于规则的方法：
      - J. Aitken, “Learning information extraction rules: An inductive logic programming approach”, **ECAI’ 02**.
      - D. McDonald, H.Chen, H. Su, and B. Marshall, “Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser”, **Bioinformatics 2004**.
    - 特征选择的方法：
      - J. Jiang and C. Zhai, “A systematic exploration of the feature space for relation extraction.”, **NAACL’ 07**.
    - 基于句法分析（核函数）的方法：
      - Guodong Zhou, Min Zhang, Donghong Ji and Qiaoming Zhu. Tree kernel based relation extraction with context-sensitive structure parse tree Information. **EMNLP’ 07**.
- **2008年**Sunita Sarawagi在*Foundations and Trends in Databases*发表知名综述长篇论文(117页)“Information Extraction”，对信息抽取，特别是**关系抽取**的研究做了深入总结，特别指出了现有基于语料库标注数据的**局限**。



# 1. 领域综述

- **2009年**，斯坦福大学的几位知名教授在ACL上提出一种新的信息抽取方法的范式(*Distant supervision for relation extraction without labeled data*)
  - Google Sites: 315.

| Entity pair                             | <Barack Obama, U.S.>   |
|---|--|
| Relation instances from knowledge bases | <ol style="list-style-type: none"><li>1. President of (Barack Obama, U.S.)</li><li>2. Born in (Barack Obama, U.S.)</li></ol>   |
| Relation mentions from free texts       | <ol style="list-style-type: none"><li>1. <b>Barack Obama</b> is the 44th and current President of the <b>U.S.</b>. (President of &amp; <b>Born in</b>)</li><li>2. <b>Barack Obama</b> ended <b>U.S.</b> military involvement in the Iraq War. (<b>President of</b> &amp; <b>Born in</b>)</li><li>3. <b>Barack Obama</b> was born in Honolulu, Hawaii, <b>U.S.</b>. (<b>President of</b> &amp; Born in)</li><li>4. <b>Barack Obama</b> ran for the <b>U.S.</b> Senate in 2004. (<b>President of</b> &amp; <b>Born in</b>)</li></ol> |





# 1. 领域综述

- 之后，DSRE（Distant Supervision for Relation Extraction）蓬勃兴起，后续的工作也逐渐被各大高校和IT公司的研究部门争相学习和进一步探索。

**Distant Supervision (Mintz2009)**: Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. *Distant supervision for relation extraction without labeled data*. *ACL'09*.

**MIL (Riedel2010)**: Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. *Modeling relations and their mentions without labeled text*. *ECML 2010*.

**MultiR (Hoffman2011)**: Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations*. *ACL'11*.

**MIML (Surdeanu2012)**: Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D. Manning. *Multi-instance Multi-label Learning for Relation Extraction*. *EMNLP-CoNLL'12*.

**Incomplete Knowledge (Bonan2013)**: Bonan, Ralph Grishman, Li Wan, Chang Wang, David Gondek. *Distant Supervision for Relation Extraction with an Incomplete Knowledge Base*. *NAACL'13*.





## 2. 研究动机

- 2008年之前的研究，传统的基于人工标注语料库（ACE、MUC）的规则和监督学习方法的
  - **缺陷：**
    - 人工标注任务量庞大，开销巨大，不适合大规模应用。
    - 模型泛化能力太弱，因为标注数据量较少。
- 2009年至今，Stanford University, Mike Mintz在ACL' 09的论文（The most solid paper in ACL）提出的基于弱标记（**知识库对齐**）的关系挖掘方法的**优势**和**缺陷**。
  - **优势：**
    - 自动通过知识库对齐假设，获取大规模弱标记样本，真正使关系挖掘模型能够应用于实际系统。
  - **缺陷：**
    - 弱标记(weakly labeled)方法的基本假设容易产生一部分**误标记样本**。
    - 大规模的弱标记数据同时产生**高维、稀疏特征**，给训练模型带来极高的**参数复杂度**。

# 3. 科学问题



- 综上，我们的对固定关系挖掘的探究点在于如何寻找能够处理弱标记 (Weakly Labeled) 噪音 (Noisy)、稀疏 (Sparse)，**同时**还能有效应对大规模数据 (Large-scale) 下的计算方法。
- 研究的**着眼点**在于如何通过低维表示寻找**真正**对实体关系预测有价值的信息，同时由于低维表示降低了模型复杂度并且改善了特征的稀疏性，能够在大数据规模的环境下应用。
- 因此学生的研究题目为：
  - 基于低维表示的大规模实体关系挖掘技术
  - **Large-scale entity relation extraction based on low-dimensional representation.**



## 4. 研究意义

- **理论意义:**

- 所要面对的数据弱标记 (Weakly Labeled)、噪音(Noisy)、稀疏(Sparse)、大规模 (Large Scale) 是一个当下亟待解决的机器学习问题 (Learning Algorithm) 。
- 基于低维表示 (Low Dimensional Distributed Representation) 的大规模 (分布式) 计算方案成为最有可能的**突破口!**

- **应用意义:**

- 真正尝试将实体关系挖掘技术从**小数据集、试验模型**的两大不实用弊端中脱离。
- 充分利用现有**互联网数据** (无结构、半结构互联网文本; 结构化知识库) 探究**实体关系挖掘技术**。



# 5. 研究计划

- 研究目标：给定大规模实体对  $\langle h, t \rangle$ ，以及实体对的上下文环境，探索基于低维表示的方法在已知的关系集合  $R$  中挖掘最佳关系子集  $R'$ ，用来描述  $\langle h, t \rangle$  之间的正确关系。
- 上下文环境（数据源）：

Google 清华大学

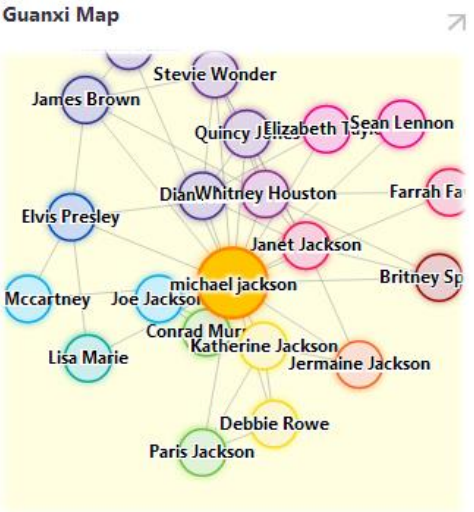
找到约 4,470,000 条结果 (用时 0.47 秒)

清华大学 - Tsinghua University  
www.tsinghua.edu.cn/

清华大学新闻搜索结果

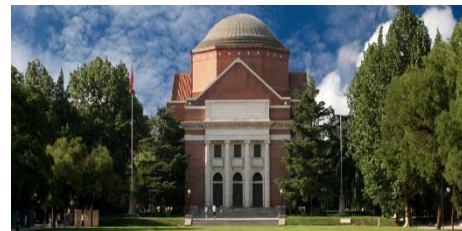
清华大学 - 维基百科，自由的百科全书  
zh.wikipedia.org/zh/清华大学

半结构化Web文本（维基百科文本）



结构化的知识库

# 5. 研究计划

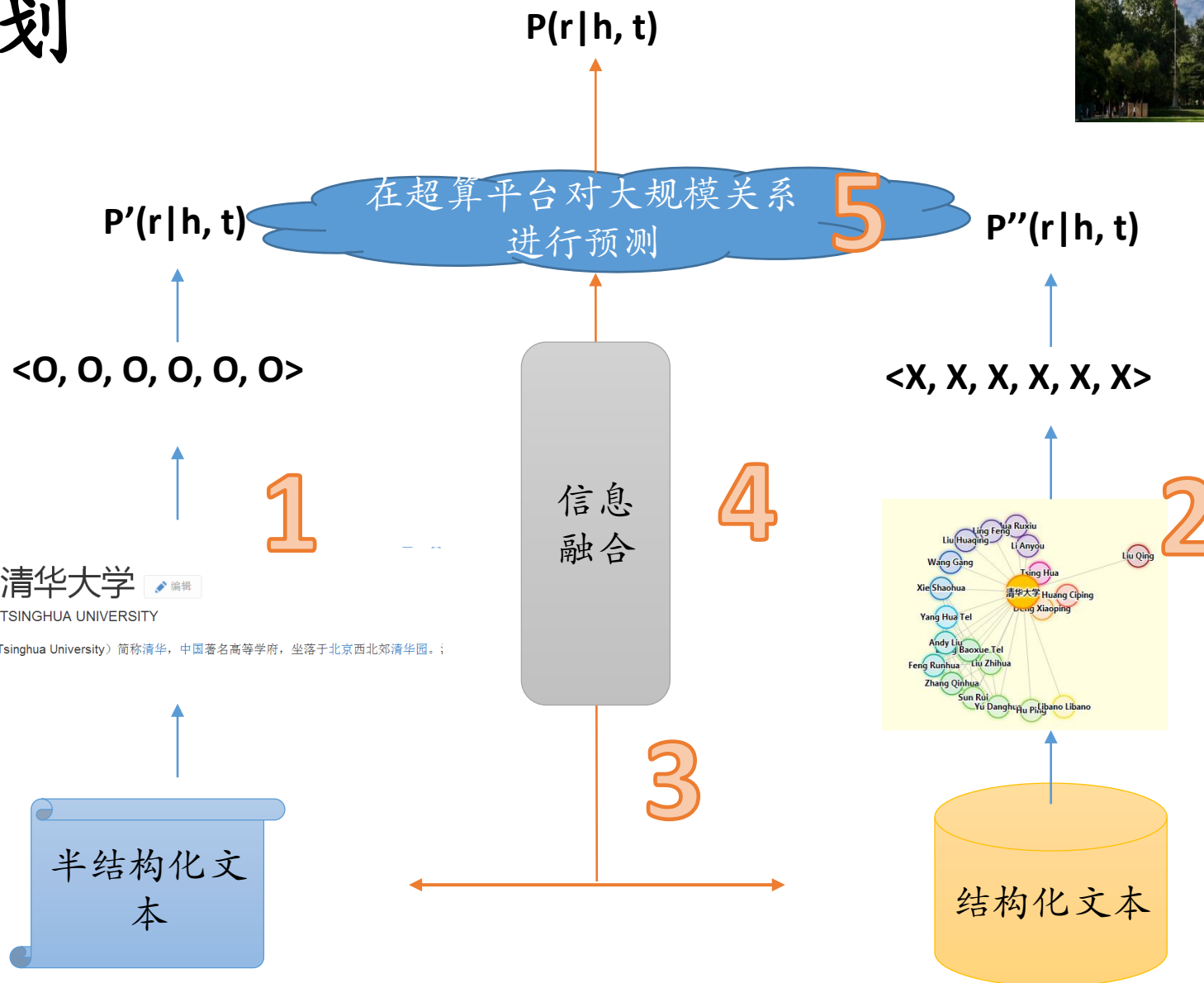


关系挖掘方法

信息的低维表示

上下文信息:

数据源:



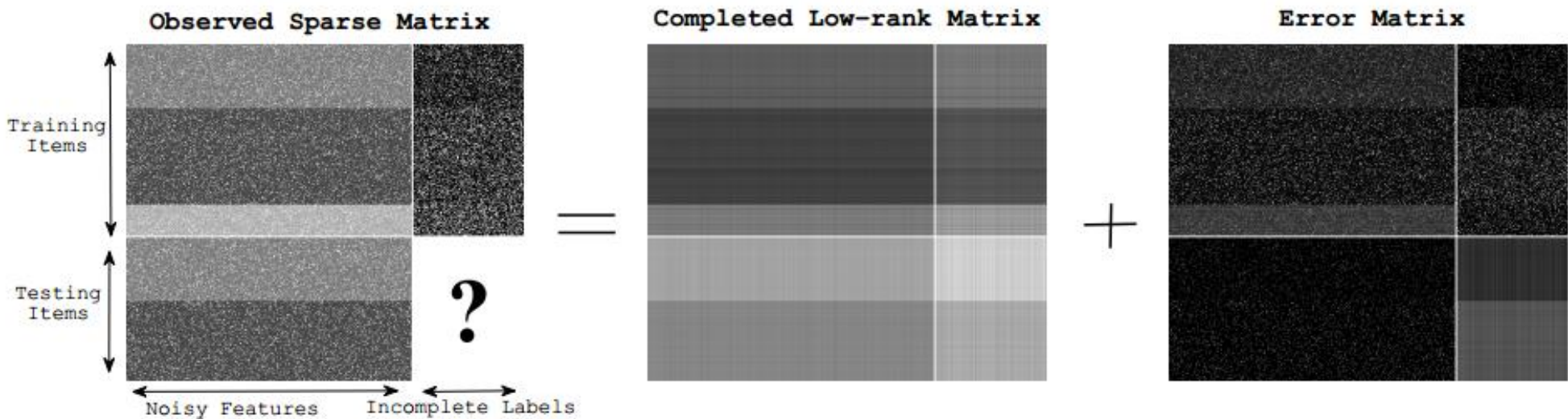




# 5. 研究计划

- **阶段一**：探究低维表示方法半结构化文本关系抽取中的有效性：

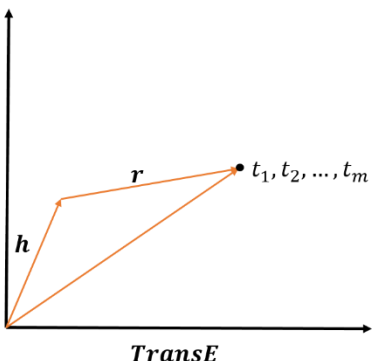
- *Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, Edward Y. Chang. 2014. Distant supervision for relation extraction with *matrix completion*. ACL 2014. *long paper, oral presentation*. (Rank: A)*
- 该论文从低维矩阵补完（低维矩阵表示三元组 $\langle h, r, t \rangle$ ）的角度，采用直推式模型，充分利用测试样本的特征信息，取得突破，主要处理从自由文本中抽取关系实例。



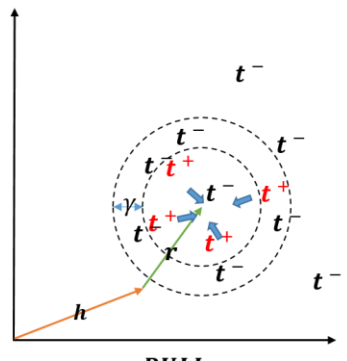
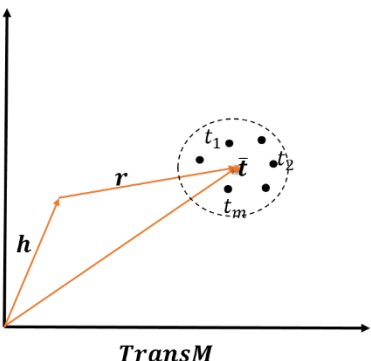


# 5. 研究计划

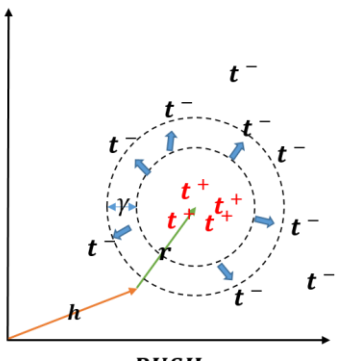
- **阶段二**: 探究低维表示方法在大规模知识图谱上的易用性:
  - *Miao Fan*, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Edward Y. Chang. Large Margin Nearest Neighbor Embedding for Knowledge Representation. *AAAI 2015*. *research paper submission*. (Rank: B, CCF:A)
  - *Miao Fan*, Qiang Zhou, Emily Chang, Thomas Fang Zheng. Transition-based Knowledge Graph Embedding with Relational Mapping Properties. *PACLIC 2014*. *long paper, ORAL*. (Rank: C)
  - 从低维向量表示 (三元组  $\langle h, r, t \rangle$ ) 的角度切入, 设计更加**易于计算**的框架, 便于处理**大规模关系数据 (百万量级)**, 主要应用在知识图自身的**关系推理 (Link prediction)**。



TransE  
PACLIC 2014



PULL  
AAAI 2015



PUSH





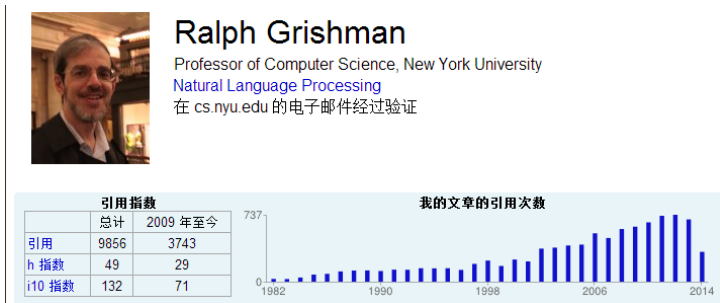
# 5. 研究计划

- **阶段三**: 尝试从**数据源**底层进行**信息整合**

- *Miao Fan*, Qiang Zhou, Deli Zhao, Thomas Fang Zheng, Edward Y. Chang. 2014. Distant Supervision for Entity Linking. **CICLING 2015**. *long paper, submission*.

- **阶段四**: 尝试整合知识库（结构化）和Web文本（半结构化），提出**统一的**基于低维表示的实体关系挖掘理论框架。

- 2015年3月-2016年3月，在**关系挖掘研究领域**的知名教授 Ralph Grishman (曾任ACL, NAACL主席, **Google H-index: 49**) 的联合指导下从事相关科研，加入其创建的“海神计划”，尝试整合两个模型各自的优势，提出统一的框架。



**The Proteus Project**

- Home
- People**
- Research
- Software
- Publications
- For Future Students
- Travel Directions
- Links

**Proteus Project Members**

| Faculty   | Students   |
|---|--|
| <ul style="list-style-type: none"><li>• <a href="#">Ralph Grishman</a></li><li>• <a href="#">Adam Meyers</a></li><li>• <a href="#">Satoshi Sekine</a></li></ul> | <ul style="list-style-type: none"><li>• <a href="#">Lisheng Fu</a></li><li>• <a href="#">Cai Kao</a></li><li>• <a href="#">Xiang Li</a></li><li>• <a href="#">Maria Pershina</a></li><li>• <a href="#">Thien Hau Nguyen</a></li><li>• <a href="#">Wei Xu</a></li></ul> |

Research Staff

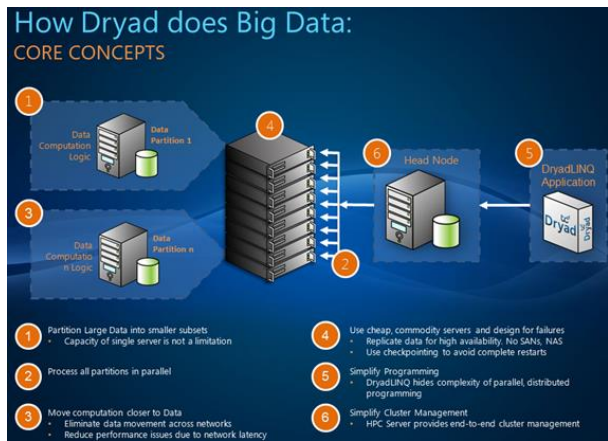
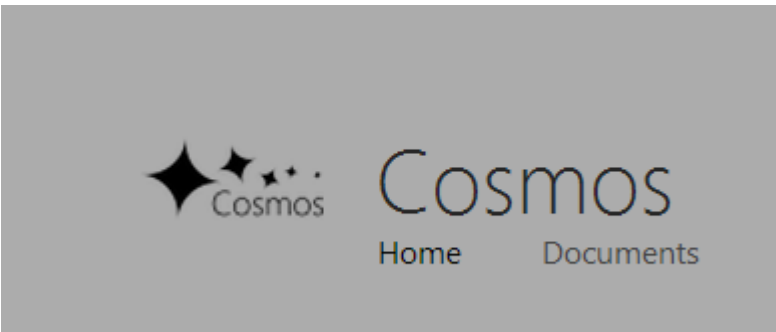
- [Yifan He](#)

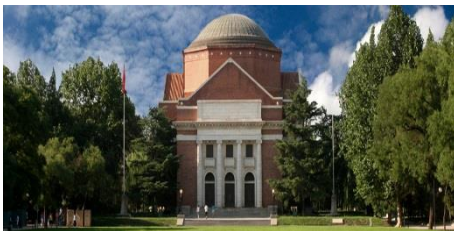
Recent Graduates



# 5. 研究计划

- **阶段五**：参考阶段二，设计更加**易于计算**的框架，处理**世界级关系数据（19亿）**，使用微软“宇宙”**超级计算平台**(Cosmos)，对Web文本（Wikipedia），大规模知识库（Freebase）进行统一处理，推断实体关系，为**实体搜索，知识图谱**等应用提供实际理论参考。



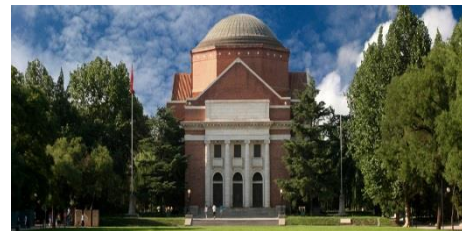


# 5. 研究计划

- 整体研究时间表:

| 起始时间    | 研究内容                              | 研究成果 (计划)                                   |
|---------|-----------------------------------|---|
| 2014年3月 | 基于低维表示的 <b>无结构化文本</b> 实体关系挖掘技术    | ACL' 14                                     |
| 2014年9月 | 基于低维表示的 <b>结构化知识</b> 实体关系推断技术     | AAAI' 15, PACLIC' 14                        |
| 2015年3月 | 基于低维表示的实体关系挖掘 (推断) <b>整合技术</b>    | 正在进行  |
| 2016年3月 | 基于低维表示的 <b>大规模 (分布式)</b> 实体关系挖掘技术 | 在微软COSMOS平台已经成功完成1<br>百亿级别数据模型的 <b>原型开发</b> |
| 2017年3月 | 整理博士论文, 准备博士论文答辩                  | -   |

# 6. 研究进度

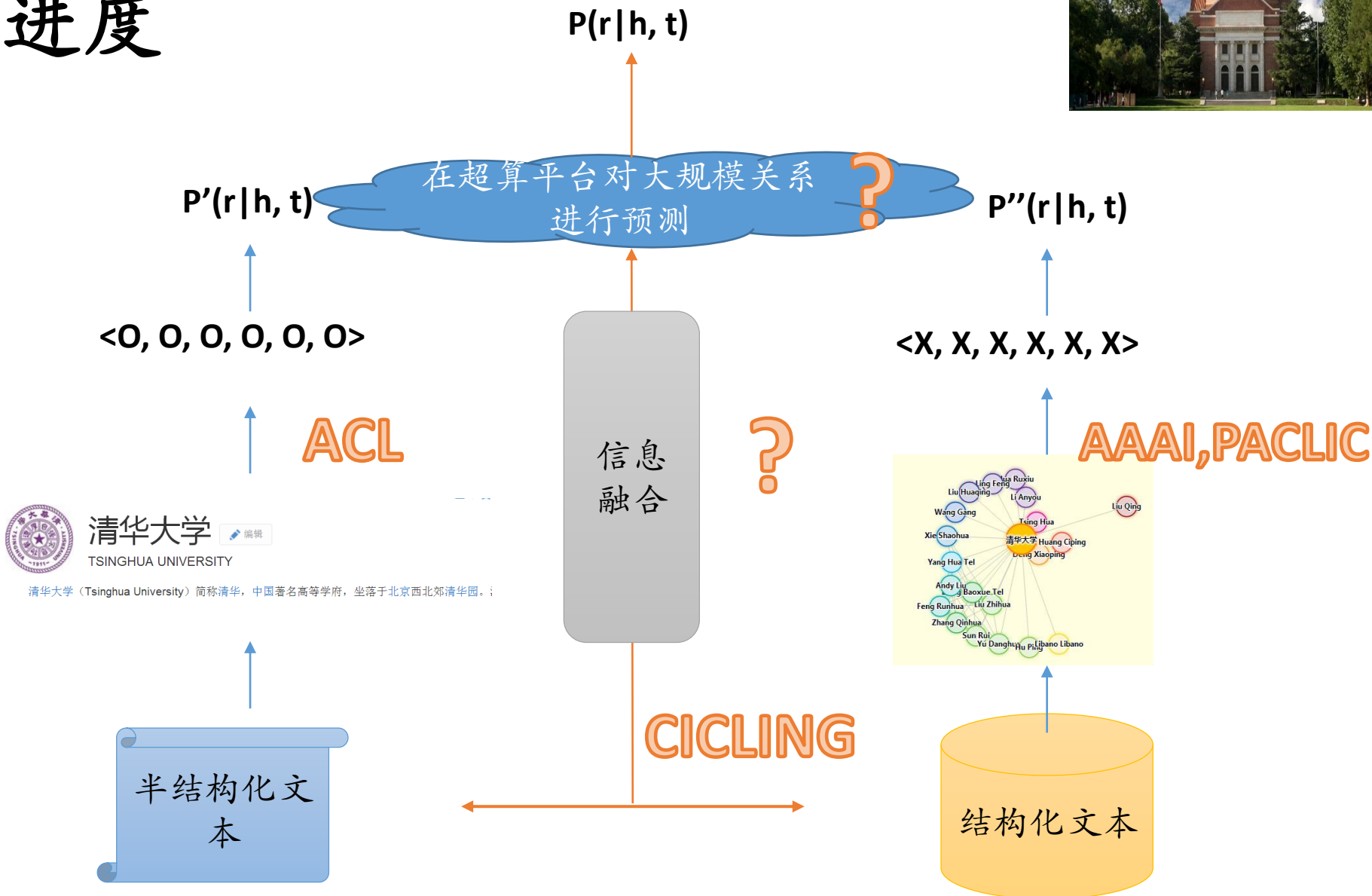


关系挖掘方法

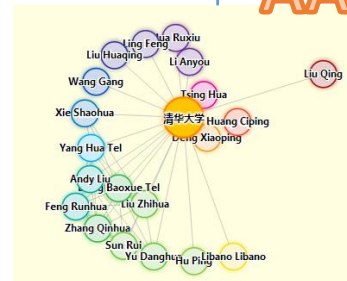
信息的低维表示

上下文信息:

数据源:



清华大学 (Tsinghua University) 简称清华, 中国著名高等学府, 坐落于北京西北郊清华园。



## 6. 研究进度 (按照计算机系的重要会议排名)



- **DBLP**: <http://dblp.uni-trier.de/pers/hd/f/Fan:Miao.html>

- [1] Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, Edward Y. Chang: *Distant supervision for relation extraction with matrix completion*. [ACL 2014](#). *long paper, ORAL*. (Rank: A)
- [2] Miao Fan, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Edward Y. Chang: *Large Margin Nearest Neighbor Embedding for Knowledge Representation*. [AAAI 2015](#). *research paper submitted*. (Rank: B, CCF: A)
- [3] Miao Fan, Qiang Zhou, Emily Chang, Thomas Fang Zheng: *Transition-based Knowledge Graph Embedding with Relational Mapping Properties*. [PACLIC 2014](#). *long paper, ORAL*. (Rank: C)
- [4] Miao Fan, Qiang Zhou, Deli Zhao, Thomas Fang Zheng, Edward Y. Chang: *Distant Supervision for Entity Linking*. [CICLING 2015](#). *long paper, submitted*. (EI Compendex)
- [5] Miao Fan, Qiang Zhou, Thomas Fang Zheng: *Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge*. [CICLING 2014](#): 188-200. *long paper, POSTER*. (EI Compendex)
- [6] Miao Fan, Qiang Zhou, Thomas Fang Zheng: *Content-Based Semantic Tag Ranking for Recommendation*. [Web Intelligence \(WI\) 2012](#): 292-296. *short paper, ORAL*. (Rank: B)
- [7] Miao Fan, Yingnan Xiao, Qiang Zhou: *Bringing the associative ability to social tag recommendation*. [ACL 2012 Workshop on Textgraph-7](#). *ORAL*.



## 7. 参考文献 (部分)

- [1] J. Aitken, “Learning information extraction rules: An inductive logic programming approach” , **ECAI’ 02**.
- [2] D. McDonald, H.Chen, H. Su, and B. Marshall, “Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser” , **Bioinformatics 2004**.
- [3] J. Jiang and C. Zhai, “A systematic exploration of the feature space for relation extraction.” , **NAACL’ 07**.
- [4] Guodong Zhou, Min Zhang, Donghong Ji and Qiaoming Zhu. Tree kernel based relation extraction with context-sensitive structure parse tree Information. **EMNLP’ 07**.
- [5] Sunita Sarawagi. Information Extraction. 2008. Foundations and Trends in Databases.

注：学生开题前共阅读相关领域重要文献近百篇，这里仅列出报告中涉及的**重点**20篇。



## 7. 参考文献 (部分)



- [6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. *Distant supervision for relation extraction without labeled data*. *ACL' 09*.
- [7] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. *Modeling relations and their mentions without labeled text*. *ECML 2010*.
- [8] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations*. *ACL' 11*.
- [9] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D. Manning. *Multi-instance Multi-label Learning for Relation Extraction*. *EMNLP-CoNLL' 12*.
- [10] Bonan, Ralph Grishman, Li Wan, Chang Wang, David Gondek. *Distant Supervision for Relation Extraction with an Incomplete Knowledge Base*. *NAACL' 13*.





## 7. 参考文献 (部分)

- [11] J. Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. **2003**. A neural probabilistic language model. *Journal of Machine Learning Research (JRML)* 3:1137 – 1155.
- [12] Bordes, A.; Weston, J.; Collobert, R.; Bengio, Y.; et al. **2011**. Learning structured embeddings of knowledge bases. In *AAAI*.
- [13] J. Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. **2014**. A semantic matching energy function for learning with multirelational data. *Machine Learning* 94(2):233 – 259
- [14] Chaiken, R.; Jenkins, B.; Larson, P.- ° A.; Ramsey, B.; Shakib, D.; Weaver, S.; and Zhou, J. 2008. Scope: easy and efficient parallel processing of massive data sets. *Proceedings of the VLDB Endowment* 1(2):1265 – 1276
- [15] Sunita Sarawagi. Information Extraction. 2008. *Foundations and Trends in Databases*.

注：学生开题前共阅读相关领域重要文献近百篇，这里仅列出报告中涉及的重点20篇。

## 7. 参考文献 (部分)



- [16] Weston, J.; Bordes, A.; Yakhnenko, O.; and Usunier, N. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1366 – 1371. Seattle, Washington, USA: Association for Computational Linguistics.
- [17] Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In Advances in Neural Information Processing Systems, 926 – 93
- [18] Schwenk, H., and Gauvain, J.-L. 2002. Connectionist language modeling for large vocabulary continuous speech recognition. In ICASSP, 765 – 768. IEEE
- [19] Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A threeway model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11), 809 – 816
- [20] Mikolov, T.; tau Yih, W.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In HLT-NAACL, 746 – 751. The Association for Computational Linguistics.

# 欢迎各位老师提问!



求知若饥、虚心若愚  
[fanmiao.cs@tsinghua.edu.cn](mailto:fanmiao.cs@tsinghua.edu.cn)