



# 基于低维表示的大规模实体关系挖掘

## Large-scale relation extraction based on low-dimensional representation

范淼

直博三年级

语音和语言技术中心

指导教师：郑方、周强

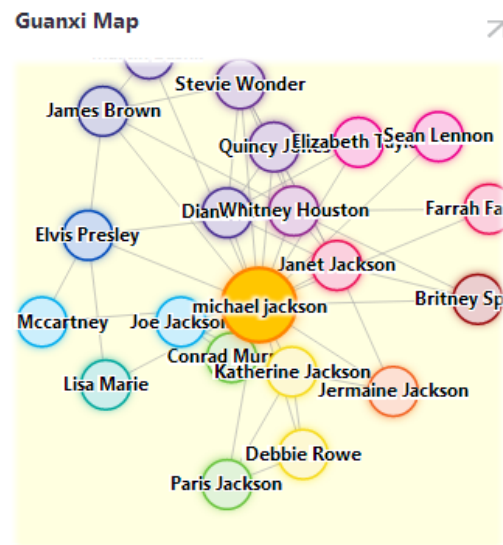
[fanmiao.cslt.thu@gmail.com](mailto:fanmiao.cslt.thu@gmail.com)

# 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 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。
- NER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 ( $\langle e_i, e_j \rangle$ ) 之间的关系 **r**。

维基百科

Barack Hussein Obama II ([🔊](#)<sup>i</sup>/[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* 发表知名综述长篇文章(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. <b>President of (Barack Obama, U.S.)</b></li><li>2. <b>Born in (Barack Obama, U.S.)</b></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>. (<b>President of</b> &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; <b>Born in</b>)</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。
- 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
  - **Large-scale relation extraction based on low-dimensional representation.**
- 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。

## 4. 研究计划



- 首先探究低维表示的有效性:

- 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*.
- 该论文从低维矩阵补完的角度, 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。

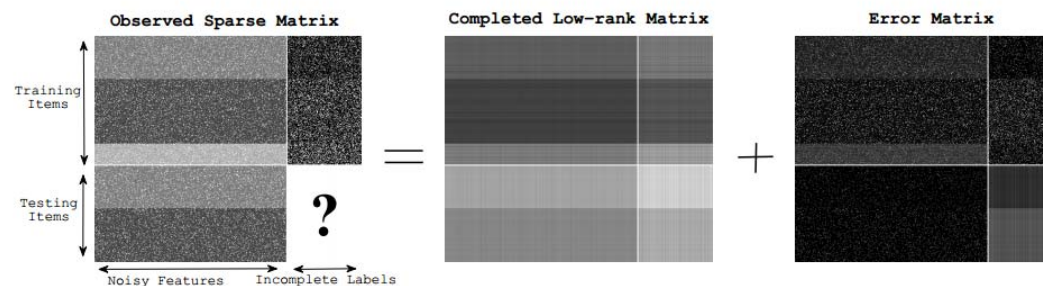


Figure 2: The procedure of noise-tolerant low-rank matrix completion. In this scenario, distantly supervised relation extraction task is transformed into completing the labels for testing items (entity pairs) in a sparse matrix that concatenates training and testing textual features with training labels. We seek to recover the underlying low-rank matrix and to complete the unknown testing labels simultaneously.

## 4. 研究计划



- 然后探究低维表示的大规模易计算框架：
  - Miao Fan, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Edward Y. Chang. 2014. Transition-based Knowledge Graph Embedding with Relational Mapping . **CIKM 2014**. *long paper submitted*.
  - 该论文依然从低维表示的角度切入，设计计算机易于计算的框架，便于处理大规模关系数据，主要应用在知识图自身的关系推理（*Link prediction*）。

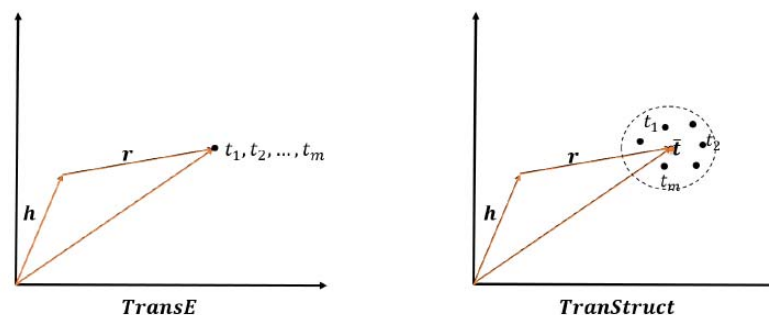
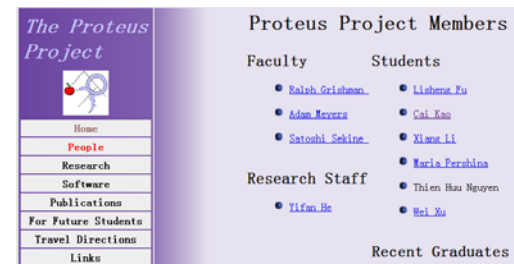


Figure 1: The differences between TransE and TransM when modeling ONE-TO-MANY relation instances, i.e.  $(h, r, t_1)$ ,  $(h, r, t_2)$ , ...,  $(h, r, t_m)$ .



## 4. 研究计划

- 学生另一篇相关工作：
  - Miao Fan, Qiang Zhou, Deli Zhao, Thomas Fang Zheng, Edward Y. Chang. 2014. Distant supervision for Entity Linking. **WISE 2014**. *long paper submitted*.
- 后续计划：
  - 2014年9月-2015年3月，结合ACL, CIKM的工作，尝试整合两个模型各自的优势，提出统一的框架。
  - 2015年3月-2016年3月，赴美NYU学习交流，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”。



- 而后回国完成博士论文答辩。

## 5. 发表（在审）论文



- [1] 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: 839-849](#). *long paper, oral presentation*.
- [2] Miao Fan, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Edward Y. Chang. 2014. Transition-based Knowledge Graph Embedding with Relational Mapping . [CIKM 2014](#). *long paper submitted*.
- [3] Miao Fan, Qiang Zhou, Deli Zhao, Thomas Fang Zheng, Edward Y. Chang. 2014. Distant supervision for Entity Linking. [WISE 2014](#). *long paper submitted*.
- [4] Miao Fan, [Qiang Zhou](#), [Thomas Fang Zheng](#): Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge. [CICLing \(2\) 2014](#): 188-200.
- [5] Miao Fan, [Qiang Zhou](#), [Thomas Fang Zheng](#): Content-Based Semantic Tag Ranking for Recommendation. [Web Intelligence 2012](#): 292-296.
- [6] Miao Fan, Yingnan Xiao, Qiang Zhou: Bringing the associative ability to social tag recommendation. [ACL'12 Workshop on Textgraph-7](#).

## 6. 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.

## 6. 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*.

## 6. 参考文献 (部分)



- [11] 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: 839-849](#). *long paper, oral presentation*.
- [12] Miao Fan, Deli Zhao, Qiang Zhou, Thomas Fang Zheng, Edward Y. Chang. 2014. Transition-based Knowledge Graph Embedding with Relational Mapping . [CIKM 2014](#). *long paper submitted*.
- [13] Miao Fan, Qiang Zhou, Deli Zhao, Thomas Fang Zheng, Edward Y. Chang. 2014. Distant supervision for Entity Linking. [WISE 2014](#). *long paper submitted*.
- [14] Miao Fan, [Qiang Zhou](#), [Thomas Fang Zheng](#): Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge. [CICLing \(2\) 2014](#): 188-200.
- [15] Miao Fan, [Qiang Zhou](#), [Thomas Fang Zheng](#): Content-Based Semantic Tag Ranking for Recommendation. [Web Intelligence 2012](#): 292-296.
- [16] Miao Fan, Yingnan Xiao, Qiang Zhou: Bringing the associative ability to social tag recommendation. [ACL'12 Workshop on Textgraph-7](#).



# 谢谢各位老师！



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