清華大學
Tsinghua University

基于表示学习的知识抽取与推断方法研究

申请工学博士学位论文答辩

清华大学计算机科学与技术系：范淼
指导老师 – 郑方

个人简介

范淼，辽宁营口人，生于1991年11月。

http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/Miao_Fan

- **2008年**本科就读于于北京邮电大学软件学院软件工程专业。
- **2012年**保送至清华大学计算机科学与技术系；现为语音与语言技术中心博士研究生，研究方向涉及机器学习与自然语言处理技术。
- **2015年**受国家留学基金委公派至美国纽约大学计算机系联合培养。

- ACM TIST、WWW'17、AAAI'17、COLING'16 审稿人，ACL/ACM 会员。
- CONTENT 国际会议程序委员。

- 以第一作者身份发表 **10** 余篇 SCI 学术期刊、国际重要会议论文（ACL、WI 等）。
- 谷歌学术引用 **120** 次。

- 以第一发明人身份申请美国、中国发明专利共 **2** 项。
- （一作）编著有《Python 机器学习及实践：从零通往 Kaggle 竞赛之路》；2016年11月29日，该书荣登京东计算机与互联网类图书销量榜第 **1**名。

论文摘要

面对互联网文本信息的爆炸性增长，人们已经愈发认识到从无/半结构化的自由文本抽取出结构化的知识进行表示与推断的重要性。知识抽取与推断方法能够帮助人们自动提炼海量自然语言文本中的核心信息，并以<实体，关系，实体>三元组事实的方式进行存储。这使得结构化的知识不论是在信息表述，还是进一步应用方面都具有广阔的前景。一些产品化的知识图谱已经为许多互联网应用如：搜索引擎、问答系统、甚至电商推荐的用户体验和平台性能带来了巨大提升。

然而，多数大型知识图谱的构建一般借助其平台优势，采用众包的方式获取海量人工编辑的知识；自动化知识抽取与推断的研究却受限于昂贵的标注数据、稀疏的文本特征、以及异构数据难以整合等问题。因此，为了解决上述三个问题，本文提出一套基于表示学习的知识自动抽取与推断系列方法。

该研究的主要贡献包括：

1）基于低秩矩阵表示学习的自由文本信息抽取。即在无/半结构化的自由文本方面，利用远程监督假设进行实体消歧与自动构建弱标记训练样本，并提出低秩矩阵补完表示的方法进行文本实体之间的关系抽取。

2）基于低维向量的知识库表示学习与事实推断。即在结构化的知识库方面，对确定与非确定性知识库分别进行有针对性的几何与概率建模，并提出利用低维向量对知识库中的实体与关系进行表示与推断；同时，分析知识库中广泛存在的多关系事实对所提出的表示学习算法的影响，进一步提升表示学习性能。

3）知识库与自由文本的联合低维表示学习。即在自由文本与知识库两种异构数据的整合应用方面，提出利用实体描述或者关系文本为纽带建立词汇与实体、关系之间的联系，学习三者的低维向量表示。

关键词：知识库；自由文本；表示学习；远程监督；关系抽取；事实推断

CONTENTS 目录

清華大學 Tsinghua University

研究背景 选题理由 相关研究 研究目标 研究内容 总结展望 研究成果

Web 2.0时代的特点——自由文本信息过载

1）自由文本信息的异构

唐纳德·特朗普

维基百科，自由的百科全书

"川普"重定向至此。关于与此名称相似的其他条目，详

唐纳德·约翰·特朗普（英语：Donald John Trump[注 2]，
区，为特朗普集团董事长兼总裁，也是特朗普娱乐公司

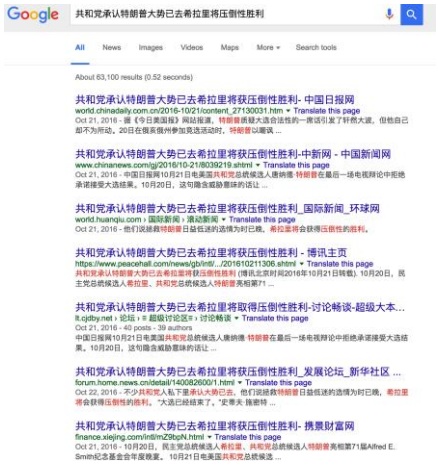
尽管特朗普是个成功的商业大亨与名人，原先他在政治
总统初选但最后退出。他于2015年6月16日，出人意料
的他在共和党全国党代表大会上获得共和党提名，成为
治精英，他常使用简单直白的言论，并利用带着里根
击败主要竞争对手——民主党候选人希拉里·克林顿，
总统。

目录 [隐藏]

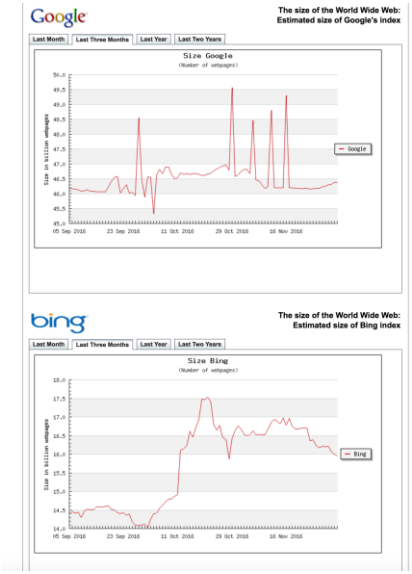
1 简介及其商业经历
2 教育经历
3 经商履历
4 创立品牌
5 媒体经历
5.1 "谁是接班人"
5.2 选美主席
5.3 世界摔角娱乐
6 政治立场
6.1 参选总统

2）自由文本信息的冗余

3）自由文本信息的噪声



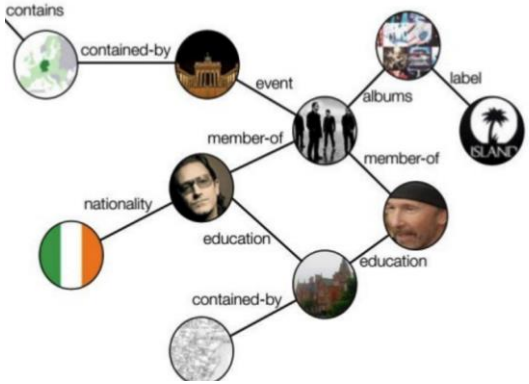
4）自由文本信息的规模



面对自由文本信息过载的解决方案——构建知识库

无/半结构化海量自由文本

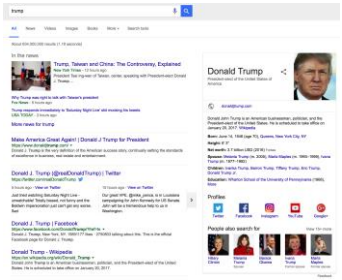
同构化、去冗余、去噪声、压缩规模



结构化知识库/知识图谱

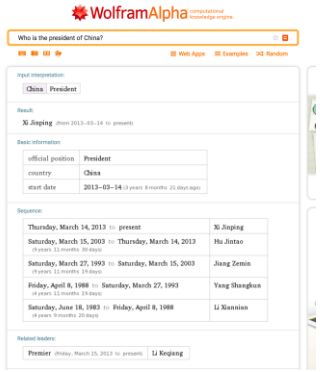
构建知识库/知识图谱的重要意义：

搜索精度、点击率以及用户体验的提升



推荐的常识性搭配

回答事实型问题

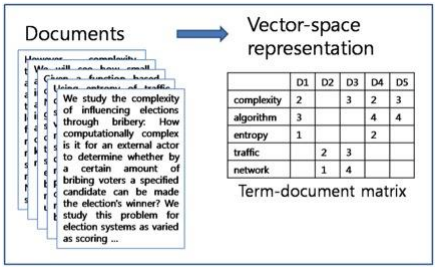


表示学习

1）离散特征表示的**直接编码方法**

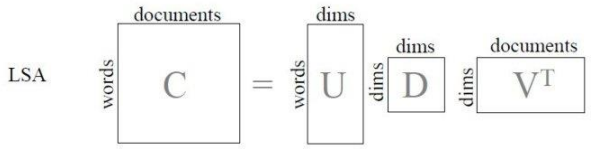
Original data:		One-hot encoding format:					
id	Color	id	White	Red	Black	Purple	Gold
1	White	1	1	0	0	0	0
2	Red	2	0	1	0	0	0
3	Black	3	0	0	1	0	0
4	Purple	4	0	0	0	1	0
5	Gold	5	0	0	0	0	1

One-hot 特征编码

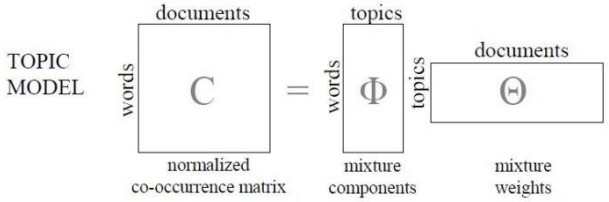


Tf(-idf) 特征编码

2）低维隐含特征表示的**无监督学习方法**

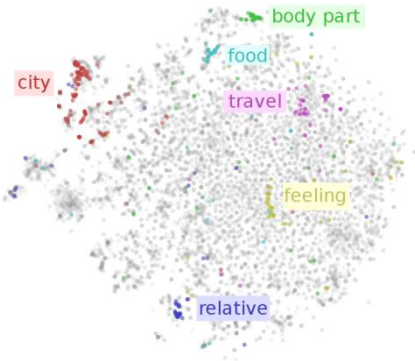


LSA/LSI



PLSA, LDA

3）低维连续特征表示的**监督学习方法**



(Word) Embedding

关系抽取

1）利用人工标注语料的关系抽取方法

Input: documents.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

IE

Output: relation triples.

NAME	Relation	ORGANIZA
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft.

2）利用自动构建弱标注语料的关系抽取方法

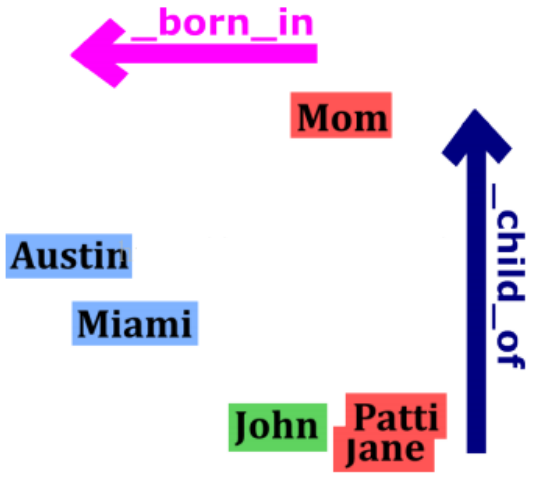
Entity pair	<Barack Obama, U.S.>
Relation instances from knowledge bases	1. **President of (Barack Obama, U.S.)** 2. **Born in (Barack Obama, U.S.)**
Relation mentions from free texts	1. **Barack Obama** is the 44th and current President of the **U.S.**. (President of) 2. **Barack Obama** ended **U.S.** military involvement in the Iraq War. (-) 3. **Barack Obama** was born in Honolulu, Hawaii, **U.S.**. (Born in) 4. **Barack Obama** ran for the **U.S.** Senate in 2004. (Senate of)

事实推断

1）基于知识库的局部图结构进行事实推断的方法

ID	PRA Path (Comment)
	athletePlaysForTeam
1	c $\xrightarrow{\text{athletePlaysInLeague}}$ c $\xrightarrow{\text{leaguePlayers}}$ c $\xrightarrow{\text{athletePlaysForTeam}}$ c (teams wit
2	c $\xrightarrow{\text{athletePlaysInLeague}}$ c $\xrightarrow{\text{leagueTeams}}$ c $\xrightarrow{\text{teamAgainstTeam}}$ c (teams that pla
	athletePlaysInLeague
3	c $\xrightarrow{\text{athletePlaysSport}}$ c $\xrightarrow{\text{players}}$ c $\xrightarrow{\text{athletePlaysInLeague}}$ c (the league that play
4	c $\xrightarrow{\text{isa}}$ c $\xrightarrow{\text{isa}^{-1}}$ c $\xrightarrow{\text{athletePlaysInLeague}}$ c (popular leagues with many players)
	athletePlaysSport
5	c $\xrightarrow{\text{isa}}$ c $\xrightarrow{\text{isa}^{-1}}$ c $\xrightarrow{\text{athletePlaysSport}}$ c (popular sports of all the athletes)
6	c $\xrightarrow{\text{athletePlaysInLeague}}$ c $\xrightarrow{\text{superpartOfOrganization}}$ c $\xrightarrow{\text{teamPlaysSport}}$ c (pop

2）基于知识库全局图结构进行事实计算的推断方法



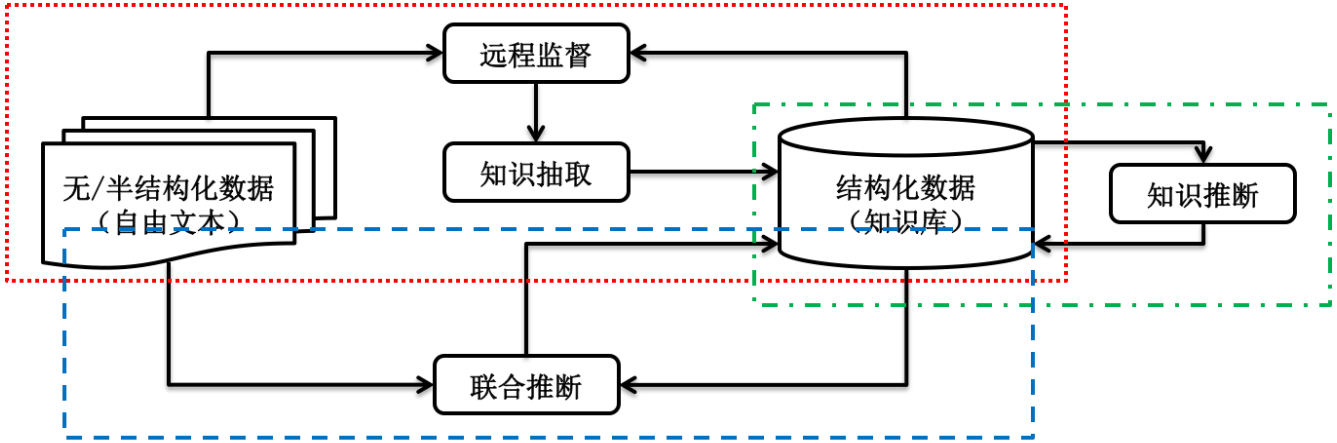
主要挑战

[挑战一] 昂贵的标注数据

[挑战二] 稀疏的文本特征

[挑战三] 难以整合的异构信息

解决方案



第3章：基于低秩矩阵表示学习的自由文本信息抽取

第4章：基于低维向量的知识库表示学习与事实推断

第5章：知识库与自由文本的联合低维表示学习

数据流

1. 基于低秩矩阵表示学习的自由文本信息抽取

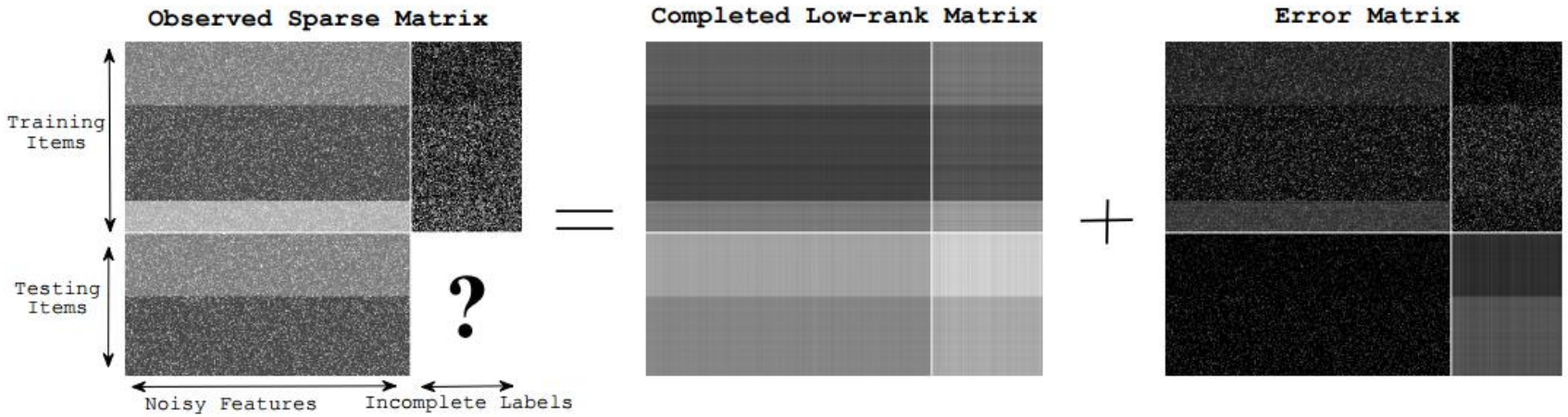
- 在无/半结构化的自由文本方面，利用**远程监督假设**进行实体消歧与自动构建弱标记训练样本，并提出低秩矩阵补完表示的方法进行文本实体之间的关系抽取。

Entity pair	<Barack Obama, U.S.>
Relation instances from knowledge bases	1. President of (Barack Obama, U.S.) 2. Born in (Barack Obama, U.S.)
Relation mentions from free texts	1. **Barack Obama** is the 44th and current President of the **U.S.**. (President of) 2. **Barack Obama** ended **U.S.** military involvement in the Iraq War. (-) 3. **Barack Obama** was born in Honolulu, Hawaii, **U.S.**. (Born in) 4. **Barack Obama** ran for the **U.S.** Senate in 2004. (Senate of)

远程监督假设

1. 基于低秩矩阵表示学习的自由文本信息抽取

$$\mathbf{Z} = \mathbf{Z}^* + \mathbf{E},$$



模型架构

1. 基于低秩矩阵表示学习的自由文本信息抽取

$$\mathbf{Z}^* = \begin{bmatrix} X^* & Y^* \end{bmatrix} = \begin{bmatrix} X^*_{train} & Y^*_{train} \\ X^*_{test} & Y^*_{test} \end{bmatrix} \quad \mathbf{E} = \begin{bmatrix} E_{X_{train}} & E_{Y_{train}} \\ E_{X_{test}} & 0 \end{bmatrix}.$$

优化目标：

$$\arg\min_{\mathbf{Z},\mathbf{b}} \ \mu||\mathbf{Z}||_* + \frac{1}{|\Omega_X|}\sum_{(i,j)\in\Omega_X} C_x(z_{ij}, x_{ij}) + \frac{\lambda}{|\Omega_Y|}\sum_{(i,j)\in\Omega_Y} C_y(z_{i(j+d)} + b_j, y_{ij}),$$

$$C(u,v) = -\log Pr(u|v) = \log(1+e^{-uv})$$

1. 基于低秩矩阵表示学习的自由文本信息抽取

学习算法

Algorithm 1 FPC algorithm for solving DRMC-b

Input:

Initial matrix $\mathbf{Z_0}$, bias $\mathbf{b_0}$; Parameters μ, λ;
Step sizes τ_z, τ_b.

Set $\mathbf{Z} = \mathbf{Z_0}$, $\mathbf{b} = \mathbf{b_0}$.
foreach $\mu = \mu_1 > \mu_2 > \ldots > \mu_F$ **do**
  while relative error $> \varepsilon$ **do**
    Gradient step:
    $\mathbf{A} = \mathbf{Z} - \tau_z g(\mathbf{Z}), \mathbf{b} = \mathbf{b} - \tau_b g(\mathbf{b})$.
    Shrinkage step:
    $\mathbf{U\Sigma V^T} = \mathrm{SVD}(\mathbf{A})$,
    $\mathbf{Z} = \mathbf{U}\ max(\mathbf{\Sigma} - \tau_\mathbf{z}\mu, 0)\ \mathbf{V^T}$.
  end while
end foreach

Output: Completed Matrix $\mathbf{Z}$, bias $\mathbf{b}$.

2. 基于低维向量的知识库表示学习与事实推断

- 在结构化的知识库方面，对确定与非确定性知识库分别进行有针对性的几何与概率建模，并提出利用低维向量对知识库中的实体与关系进行表示与推断。

在词向量空间具备这样的关系

$China - Beijing \approx France - Paris$

在知识向量空间呢?

$China - Beijing \approx capital\_city\_of$

对于一个三元组的事实(h, r, t),

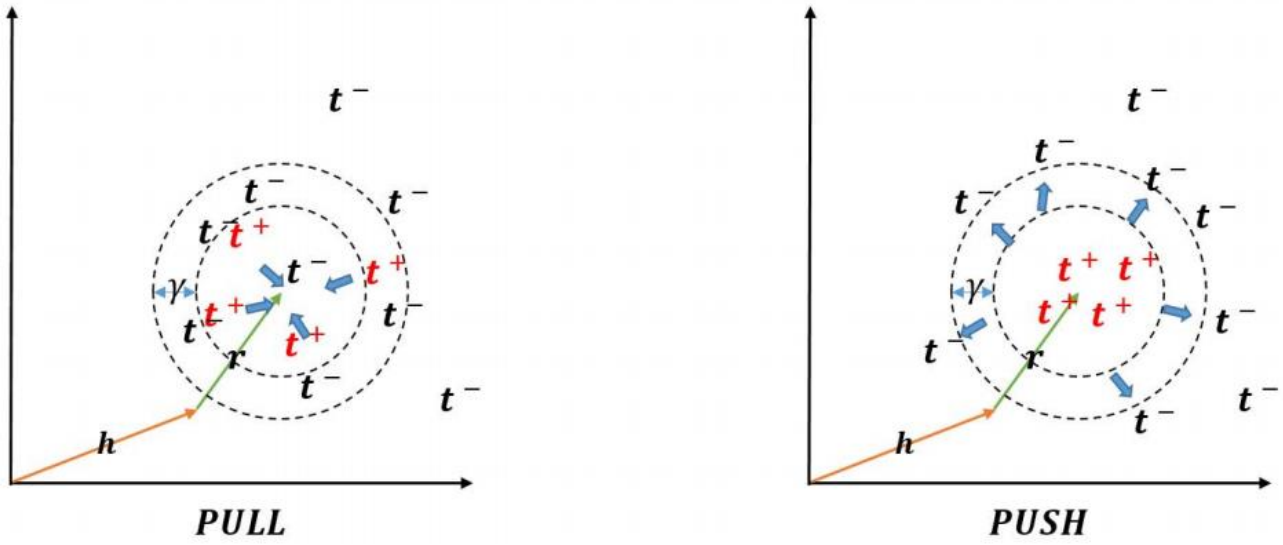
$$\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$$

(h:Beijing, r:capital_city_of, t: China)

(h:Paris, r: capital_city_of, t: France)

启发图解

2. 基于低维向量的知识库表示学习与事实推断



模型架构

2. 基于低维向量的知识库表示学习与事实推断

$$f_r(h,t) = ||h + r - t||, \qquad (1)$$

$$\mathcal{L}_{pull} = \text{Min} \sum_{(h,r,t)\in\Delta} \sum_{(h^+,r,t^+)\in\Delta^+_{(h,r,t)}} (||h-h^+||+||t-t^+||). \qquad (2)$$

优化目标：

$$\mathcal{L}_{push} = \text{Min} \sum_{(h,r,t)\in\Delta} \sum_{(h^-,r,t^-)\in\Delta^-_{(h,r,t)}} [\gamma+f_r(h,t)-f_r(h^-,t^-)]_+. \qquad (3)$$

$$\mathcal{L} = \text{Min} \quad \mu\mathcal{L}_{pull} + (1-\mu)\mathcal{L}_{push}. \qquad (4)$$

2. 基于低维向量的知识库表示学习与事实推断

Algorithm 1 The Learning Algorithm of **LMNNE**

Input:

Training set $\Delta = \{(h, r, t)\}$, entity set E, relation set R; dimension of embeddings d, margin γ, learning rate α and β for $\mathcal{L}_{pull}$ and $\mathcal{L}_{push}$ respectively, convergence threshold ϵ, maximum epoches n and the trade-off μ.

foreach $\mathbf{r} \in R$ **do**
  $\mathbf{r} := \text{Uniform}(\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}})$
  $\mathbf{r} := \frac{\mathbf{r}}{|\mathbf{r}|}$
end foreach
$i := 0$
while $Rel.loss > \epsilon$ and $i < n$ **do**
  foreach $\mathbf{e} \in E$ **do**
    $\mathbf{e} := \text{Uniform}(\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}})$
    $\mathbf{e} := \frac{\mathbf{e}}{|\mathbf{e}|}$
  end foreach
  foreach $(h, r, t) \in \Delta$ **do**
    $(h', r, t') := \text{Sampling}(\Delta'_{(h,r,t)})$
    if $(h', r, t') \in \Delta^{+}_{(h,r,t)}$ **then**
      Updating: $\nabla_{(h,r,t,h',t')}\mathcal{L}_{pull}$ with: $\alpha\mu$
    end if
    if $(h', r, t') \in \Delta^{-}_{(h,r,t)}$ **then**
      Updating : $\nabla_{(h,r,t,h',t')}\mathcal{L}_{push}$ with: $\beta(1-\mu)$
    end if
  end foreach
end while

Output:

All the embeddings of e and r, where $e \in E$ and $r \in R$.

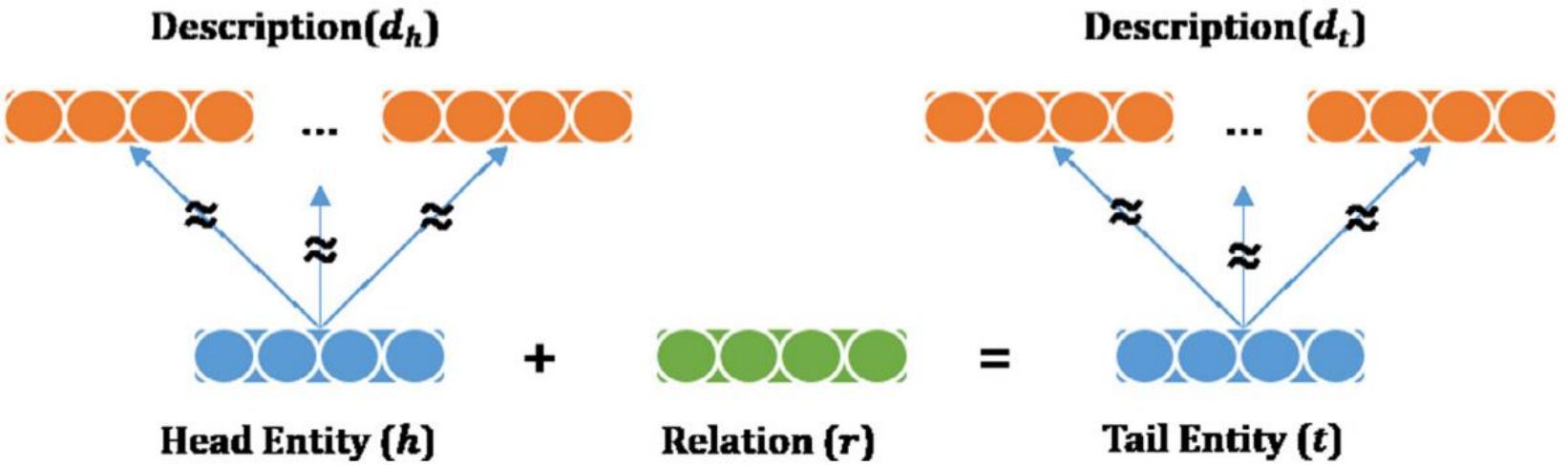
学习算法

$$\mathcal{L}_{pull} = \text{Min} \sum_{(h,r,t)\in\Delta} \sum_{(h^+,r,t^+)\in\Delta^{+}_{(h,r,t)}} (||h-h^+||+||t-t^+||). \quad (2)$$

$$\mathcal{L}_{push} = \text{Min} \sum_{(h,r,t)\in\Delta} \sum_{(h^-,r,t^-)\in\Delta^{-}_{(h,r,t)}} [\gamma+f_r(h,t)-f_r(h^-,t^-)]_+. \quad (3)$$

3. 知识库与自由文本的联合低维表示学习

- 在自由文本与知识库两种异构数据的整合应用方面，提出利用实体描述或者关系文本为纽带建立词汇与实体、关系之间的联系，学习三者的低维向量表示。



模型架构

3. 知识库与自由文本的联合低维表示学习

优化目标：

$$\arg\max_{h,r,t,d_h,d_t} \sum_{(h,r,t,d_h,d_t)\in\Delta} \log Pr(h,r,t,d_h,d_t), \quad (1)$$

$$\log Pr(h,r,t,d_h,d_t) = \log Pr(h,r,t) + \log Pr(d_h,d_t|h,r,t). \quad (2)$$

$$\log Pr(h,r,t) = \frac{\log Pr(h|r,t) + \log Pr(r|h,t) + \log Pr(t|h,r)}{3}. \quad (3)$$

$$\log Pr(d_h,d_t|h,r,t) = \log Pr(d_h|h) + \log Pr(d_t|t).$$

3. 知识库与自由文本的联合低维表示学习

Algorithm 1 : The Learning Algorithm of *RLKB*

Input:

The training knowledge base $\Delta = \{(h, r, t, d_h, d_t)\}$, entity set E, relation set R, vocabulary set V of entity descriptions; dimension of embeddings k, number of negative samples n, learning rate r, the bias α and β.

foreach $\mathbf{e} \in E$ **do**
  $\mathbf{e} := \text{Uniform}(-\frac{6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
foreach $\mathbf{r} \in R$ **do**
  $\mathbf{r} := \text{Uniform}(-\frac{6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
foreach $\mathbf{w} \in V$ **do**
  $\mathbf{w} := \text{Uniform}(-\frac{6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
while not adequate rounds **do**
  foreach $(h, r, t, d_h, d_t) \in \Delta$ **do**
    foreach $i \in \text{range}(n)$ **do**
      $\Delta'_{(h,r,t,d_h,d_t)}$ appends a negative sample: $\langle h'_i, r'_i, t'_i, d'_{h,i}, d'_{t,i} \rangle$
      /*$\Delta'_{(h,r,t,d_h,d_t)}$ is the set of n negative samples, given the positive knowledge (h, r, t, d_h, d_t).*/
    end foreach
    Conduct gradient ascent with learning rate r on $\log Pr(h, r, t, d_h, d_t)$, and update the embeddings based on Eq. (20).
  end foreach
  Check the probability over the validation set.
  Set $\Delta'_{(h,r,t,d_h,d_t)}$ empty.
end while

Output:

All the embeddings of h, t, r, w, where $h, t \in E$, $r \in R$ and $w \in \{d_h, d_t\}$.

学习算法

3. 知识库与自由文本的联合低维表示学习

Entity	/m/01n4w_ (Washington and Lee University)
Nearest@10	/m/0kw4j (American University) /m/017v3q (College of William & Mary) /m/01nnsv (George Washington University) /m/0pspl (Georgetown University) /m/0438f (James Madison University) /m/037s9x (Washington & Jefferson College) /m/02zr0z (Virginia Union University) /m/0g8rj (University of Virginia) /m/07t90 (University of Washington) /m/04wlz2 (Hampton University)

Keyword	*Colleges*
Nearest@10	*Laboratory* *Universities* *University* *Instituion* *Nonsectarian* *Granting* *Eiga* *Students* *Drumlins*

学习效果

研究工作总结

[贡献一] 基于低秩矩阵表示学习的自由文本信息抽取

[贡献二] 基于低维向量的知识库表示学习与事实推断

[贡献三] 知识库与自由文本的联合低维表示学习

未来工作展望

[展望一] 多关系与复合型知识的表示学习研究

[展望二] 词汇、实体和关系的统一空间表示模型

[展望三] 分布式架构下知识向量化表示学习的无损算法

已发表的重要学术期刊、会议论文:

- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng, Ralph Grishman. Distributed Representation Learning for Knowledge Bases with Entity Descriptions. *Pattern Recognition Letters* (Special Issue on Pattern Recognition Techniques on Data Mining), 2016. SCI 5-year Impact Factor (2015): 2.002.
- **Miao Fan**, Qiang Zhou, Andrew Abel, Thomas Fang Zheng, Ralph Grishman. Probabilistic Belief Embedding for Large-scale Knowledge Population. *Cognitive Computation*, 2016. SCI 5-year Impact Factor (2015): 1.714.
- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng. Learning Embedding Representations for Knowledge Inference on Imperfect and Incomplete Repositories. IEEE/WIC/ACM International Conference on Web Intelligence (*WI'16*), 2016. Regular paper, oral presentation. EI Index: xxx.
- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng. Distant Supervision for Entity Linking. The 29th Pacific Asia Conference on Language, Information and Computation (*PACLIC'15*), 2015. Regular paper, oral presentation. EI Index: 20162002400438.
- **Miao Fan**, Kai Cao, Yifan He, Ralph Grishman. Jointly Embedding Relations and Mentions for Knowledge Population. The 10th Recent Advances in Natural Language Processing (*RANLP'15*), 2015. Poster paper. EI Index: 20155101692258.
- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng, Ralph Grishman. Large Margin Nearest Neighbor Embedding for Knowledge Representation. IEEE/WIC/ACM Web Intelligence Conference (*WI'15*), 2015. Regular paper, oral presentation. EI Index: xxxx.

已发表的重要学术期刊、会议论文:

- **Miao Fan**, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, Edward Y. Chang. Distant Supervision for Relation Extraction with Matrix Completion. The 52th Annual Meeting of the Association for Computational Linguistics (*ACL'14*), 2014. Regular paper, oral presentation. EI Index: 20143718156957.
- **Miao Fan**, Qiang Zhou, Emily Chang, Thomas Fang Zheng. Transition-based Knowledge Graph Embedding with Relational Mapping Properties. The 28th Pacific Asia Conference on Language, Information and Computing (*PACLIC'14*), 2014. Regular paper, oral presentation.
- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng. Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge. The 15th International Conference on Intelligent Text Processing and Computational Linguistics (*CICLing'14*), 2014. Regular paper, poster presentation. EI Index: 20142017719416.
- **Miao Fan**, Qiang Zhou, Thomas Fang Zheng. Content-based Semantic Tag Ranking for Recommendation. IEEE/WIC/ACM International Conference on Web Intelligence (*WI'12*), 2012. Short paper, oral presentation. EI Index: 20132316402034.
- **Miao Fan**, Yingnan Xiao, Qiang Zhou. Bringing the Associative Ability to Social Tag Recommendation. *ACL'12* Workshop on Graph-based Methods for Natural Language Processing. Workshop paper, oral presentation. EI Index: 20133616706762.

参与的重大科研项目：

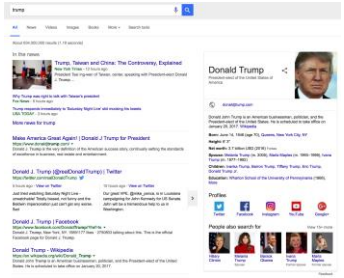
- 国家973计划项目：互联网环境中文言语信息处理与深度计算的基础理论和方法。项目编号：2013CB329304。立项部门：中华人民共和国科技部。时间：2013-2017。
- 国家自然科学基金项目：互联网话语理解的认知与计算建模。项目编号：61433018。立项部门：国家自然科学基金委。时间：2015-2019。
- 国家自然科学基金项目：汉语语篇中连贯关系和隐含角色的分析标注研究。项目编号：61373075。立项部门：国家自然科学基金委。时间：2014-2017。

出版的专著：

- **范淼**、李超，编著。《Python机器学习及实践：从零开始通往Kaggle竞赛之路》。清华大学出版社。2016年10月第1版。ISBN：978-7-302-44287-5。中国版本图书馆CIP数据核字（2016）第164306号。

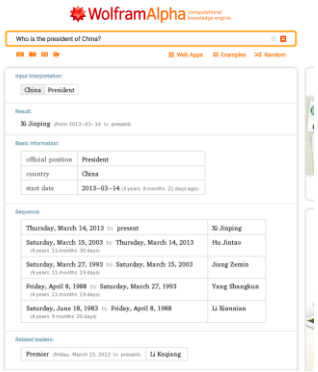
讨论：这项研究到底如何用在如下的领域?

搜索精度、点击率以及用户体验的提升



推荐的常识性搭配

回答事实型问题



清華大學
Tsinghua University

谢谢聆听

基于表示学习的知识抽取与推断方法研究

清华大学计算机科学与技术系：范淼
指导老师 – 郑方