知识图谱
实体链接技术

胡伟

南京大学 计算机科学与技术系

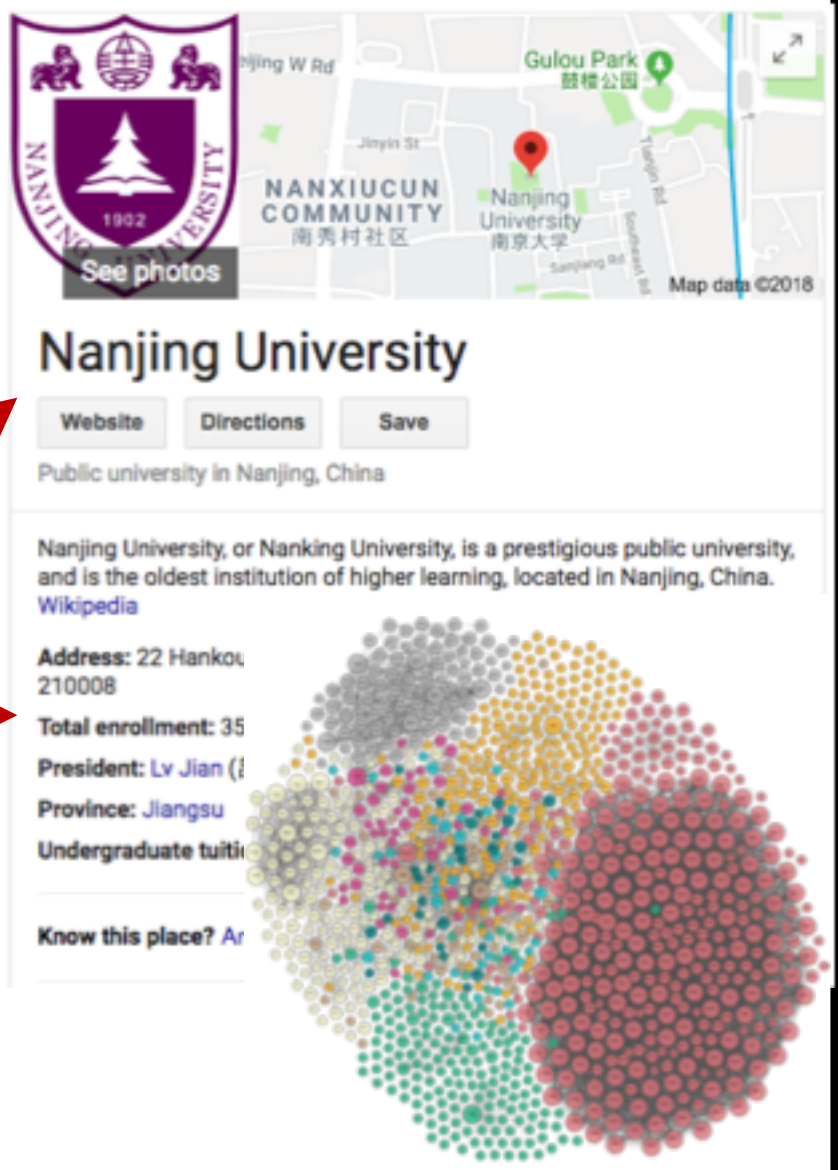
whu@nju.edu.cn

提纲

- 简介
  - 知识图谱
  - 实体链接
- 近期工作介绍
- 总结与展望

知识图谱

- Knowledge graph is a **knowledge base** used by Google to enhance its search results with semantic search information gathered from a wide variety of sources
  - Node: entity / concept
  - Edge: attribute / relationship
- Other famous knowledge bases
  - DBpedia, Freebase, Wikidata, YAGO, WordNet, Probase …
  - Linked Open Data (LOD) cloud



异构性

- Since long long time ago …
  - Syntactic
    - e.g., “Wei Hu” vs. “HU, Wei”
  - Terminological
    - e.g., “notebook” vs. “laptop”
  - Semantic
    - e.g., $\mathrm{hasSon}(x, y)$ vs. $\mathrm{hasChild}(x, y) \sqcap \mathrm{Male}($
  - Pragmatic

Schema-level

ontology matching

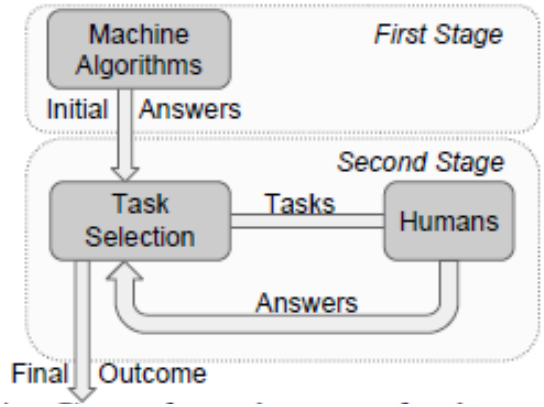
Data-level

entity linkage

- Knowledge graphs have reached a scale in billions of entities

实体链接

- Many different entities refer to the **same** real-world thing
  - typically, denoted by URIs, from distributed data sources
- **Entity linkage (EL):** find different entities referring to the same
  - a.k.a. entity resolution, entity matching … (also widely studied in DB and NLP)
  - important to resolve heterogeneity and achieve interoperability
- **Crowd entity linkage**
  - use humans, in addition to machines, to obtain the truths of EL tasks
  - Key issues
    - *How to present a single EL task?*
    - *How to select "right" human?*
    - *How to pick tasks under budget?*
    - *……*



提纲

- 简介
- 近期工作介绍
  - 任务设计
  - 用户建模
- 总结与展望

Motivation

- Crowd EL
  - request a human to judge which entities in a single EL task refer to the same
  - Little effort has been made on how to present critical information
    - e.g., **important properties and values**
    - to help the human complete the task more efficiently and accurately

e_1 [dbp:Lil_Eazy-E]	e_2 [fb:m.01wf_p_]	e_3 [wd:Q36804]
– rdf:type : Person, MusicalArtist	– alias : Eric Wright, Eazy-E	– rdfs:label : Eazy-E
– rdfs:label : Lil Eazy-E	– date_of_birth : 1963-9-7	– altLabel : Eric Lynn Wright
– owl:sameAs : fb:m.01wf_p_	– gender : male	– date_of_birth : 1963-9-7
– birthDate : 1984-4-23	– genre : gangsta rap, hip hop	– desc : Gangsta rapper, producer
– birthPlace : Compton	– name : Eazy-E	– genre : gangsta rap
– gender : male	– place_of_birth : Compton	– instance_of : human
– genre : Gangsta rap, Hip hop	– profession : rapper, producer	– occupation : musician, rapper
– givenName : Eric Darnell Wright	– type : person, music.artist	– place_of_birth : Compton
(146 property-values in total)	*(391 property-values in total)*	*(141 property-values in total)*

three entities with similar types, names, genders

Related work

1. Display multiple entities in a form of **list**
   - just like what is typically seen from a Web search engine
2. Use **pairwise** presentation
   - compare two entities at a time and aligns similar properties between them

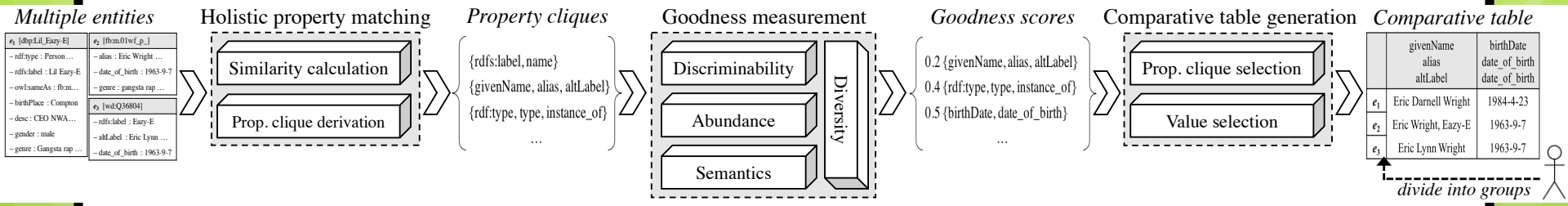
Entity summarization

- Pros & cons for multi-entity linkage (MEL)
  1. List: remember and compare in mind
  2. Pairwise: focus / difficult to scale
  - Both lost transitivity & grouping info

e_1 [dbp:Lil_Eazy-E]	e_2 [fb:m.01wf_p_]	
– rdf:type : Person, MusicalArtist	– type : person, music.artist	
– genre : Gangsta rap, Hip hop	– genre : gangsta rap, hip hop	
– givenName : Eric Darnell Wright – rdfs:label : Lil Eazy-E	– alias : Eric Wright, Eazy-E	
– birthPlace : Compton	– place_of_birth : Compton	⦿ match
– gender : male	– gender : male	◎ nonmatch

- Our approach: **comparative table**
  - arrange entities and properties in the task as the row and column headers of a table
  - assign their corresponding values in the cells

Approach workflow



1. Holistic property matching
   - similarity calculation → property clique derivation
2. Goodness measurement
   - discriminability, abundance, semantics → diversity
3. Comparative table generation
   - property clique selection → value selection

group?		givenName alias altLabel	rdf:type type instance_of	birthDate date_of_birth date_of_birth
1	e_1	Eric Darnell Wright	Person, MusicalArtist	1984-4-23
2	e_2	Eric Wright, Eazy-E	person, music.artist	1963-9-7
3	e_3	Eric Lynn Wright	human	1963-9-7

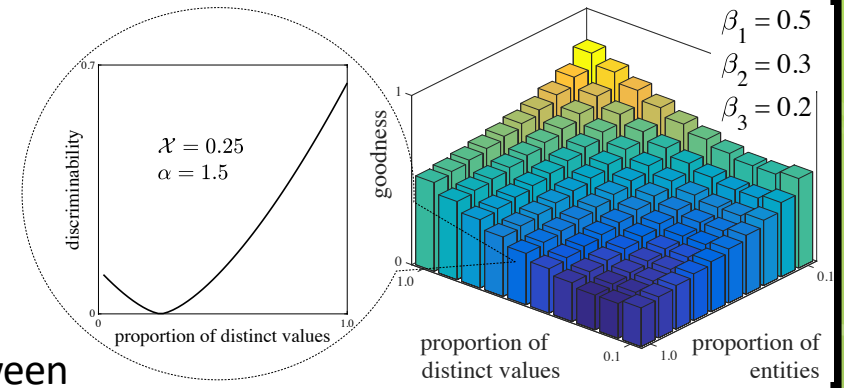
Step 1: holistic property matching

- Property similarity: label, local name, value
  - combined with logistic regression
- Property cliques
  - restrict each property can match at most one other property
  - choose the pairs with highest match probability estimate may lead to conflicts
- **Holistic property matching**
  - maximize the overall match probability estimate among all matched property pairs
    - s.t. 1:1 matching constraint is satisfied
  - NP-hard (3-dimensional assignment)
  - Greedy algorithm

Input: Set of matched property pairs **J**
Output: Set of property cliques **C**
C ← ∅; **J′** ← **J**;
Sort **J′** in descending order according to match probability estimates;
repeat
  Pop $p_a \simeq p_b \in \mathbf{J}'$ holding the highest match probability estimate;
  Find cliques $C_i, C_j \in \mathbf{C}$ containing p_a, p_b, respectively;
  if $C_i = \emptyset$ **and** $C_j = \emptyset$ **then**
    Create a new clique $\{p_a, p_b\}$ and add it in **C**;
  else if C_i *has any property from the same namespace as* p_b **or** C_j *has any property from the same namespace as* p_a **then**
    Drop $p_a \simeq p_b$, due to violate the global 1:1 constraint;
  else
    Merge C_i, C_j into a larger clique C_k;
    Remove C_i, C_j from **C** and add C_k in **C**;
until **J′** = ∅;
return **C**;

Step 2: goodness measurement

- Goodness of property cliques
  1. **Discriminability** measures how well a property clique reveals the commonalities and differences among multiple entities
  2. **Abundance** assesses how adequate of information the property clique provides
  3. **Semantics** gives extra scores to the ones particularly useful, e.g., owl:sameAs
  4. **Diversity** evaluates the redundancy between different property cliques
- 2-phase combination: (1 + 2 + 3) + 4
- Goodness of values
  - Longer length, less redundant



$$div(C_i \mid \mathbf{D}) = \max_{C_j \in \mathbf{D}} div(C_i \mid C_j),$$

$$div(C_i \mid C_j) = \frac{\sum_{e \in \mathbf{S}_i \cap \mathbf{S}_j} \max_{\substack{v_x \in val(e, C_i) \\ v_y \in val(e, C_j)}} S_V(v_x, v_y)}{|\mathbf{S}_i \cup \mathbf{S}_j|},$$

Step 3: comparative table generation

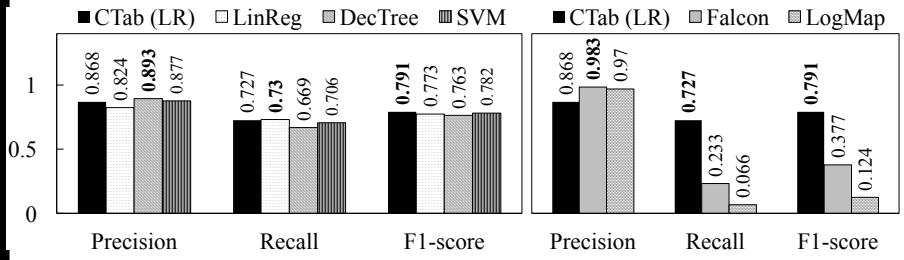
- **Property clique selection**
  - Greedy method
    - Given the maximal number of property cliques in a comparative table, simply select top property cliques with best goodness
      - cannot guarantee each entity to be at least described by several properties
  - **Optimal property clique selection with entity coverage constraint**
    - NP-hard (set cover)
    - $H(N)$-approximation
- **Value selection**
  - model the value selection based on the classic 0/1 knapsack problem with a table table cell size constraint

Input: Property clique set **C**, entity set **E**, least cover times T
Output: Property clique subset **D** $\subseteq$ **C**
D $\leftarrow \emptyset$; **C**′ $\leftarrow$ **C**;
foreach $e_a \in$ **E** **do** $x_a \leftarrow T$;
while **E** $\neq \emptyset$ **do**
  Select $C_j \in$ **C**′ such that $\frac{good(C_j)}{|\mathbf{E} \cap \mathbf{S}_j|}$ is minimized;
  Add C_j in **D** and remove it from **C**′;
  foreach $e_a \in \mathbf{E} \cap \mathbf{S}_j$ **do**
    $x_a \leftarrow x_a - 1$;
    if $x_a = 0$ **then** Remove e_a from **E**;
return **D**;

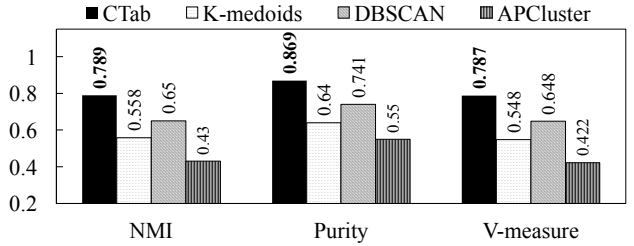
Experiment on holistic property matching

- Multi-entity linkage tasks
  - 10 popular domains, 25 DBpedia entities per domain as seeds
  - Wikipedia disambiguation page, 2~4 Freebase, Wikidata, YAGO entities
  - 250 tasks, 2500 entities, 804 distinct objects

- **Quality of matched property pairs**
  - "Official" property matches
  - Label others by 3 graduate students
    - 484 matches, 1397 non-matches



- **Quality of derived property cliques**
  - Compute connected components
    - 135 reference property cliques



- **Running time**
  - 27s per task for similarity calculation
  - Less than 1s for other steps

Experiment on property clique selection

- 3 experienced humans scored property cliques in each task
  - Highly-useful (3), fairly-useful (2), marginally-useful (1) and useless (0)
- Comparative systems
  - FACES (list) [Gunaratna et al., AAAI 2015]
  - C3D+P (pairwise) [Cheng et al., JWS 2015]
  - Ctab, CTab (entropy), CTab (greedy)
- **Directly rank ref. property cliques**

	Use reference property cliques				K_{Haus}
	P@1	P@5	P@10	nDCG@5	
FACES	0.176	0.310	0.290	0.239	0.753
C3D+P	0.040	0.347	0.511	0.154	0.647
CTab (entropy)	0.180	0.178	0.184	0.092	0.811
CTab (greedy)	0.632	0.660	0.615	0.684	0.647
CTab	**0.756**	**0.754**	**0.643**	**0.798**	**0.615**

K_{Haus}	Discr.	Abund.	Sem.	w/o Div.	Good
CTab (greedy)	0.678	0.686	0.673	0.655	0.647
CTab	0.675	0.633	0.815	0.618	**0.615**

- **Assess property clique selection and goodness measurement together**
  - The Hausdorf version of the Kendall tau distance
    - treat property clique rankings as partial rankings of properties (the properties with the same grade and in the same clique are tied)

Experiment on human intervention

- 60 graduate students (top-5/-10), 30 orthogonal tasks per human, 100RMB
  - Task difficulty is not significantly different in statistics among FACES, C3D+P, CTab
- Completion time
  - Participants had to view more entity pairs before decision
- Precision
  - Break the entities in each entity group down to pairs
- Human scoring and comments
  - For CTab, if the least cover times was not satisfied, …

		FACES (L)	C3D+P (P)	CTab (T)	*p*-value	Post-hoc
Top-5	Time (s)	153	208	**96**	0.01%	P < L < T
	Prec.	0.63	0.69	**0.77**	0.07%	L, P < T
Top-10	Time (s)	175	180	**131**	1.13%	L, P < T
	Prec.	0.79	0.77	**0.80**	69.8%	

Questions [from 1: "totally disagree" to 5: "totally agree"]	FACES (L)	C3D+P (P)	CTab (T)	*p*-value	Post-hoc
Q1. The system provided adequate information of entities.	3.11	3.17	**3.70**	0.76%	L, P < T
Q2. The system provided unsuperfluous information of entities.	2.67	**3.30**	3.23	4.46%	L < T, P
Q3. The system helped me easily compare entities of interest.	2.43	3.37	**4.00**	< 0.01%	L < P < T
Q4. I found the system easy to use.	3.00	3.13	**3.70**	2.28%	L, P < T

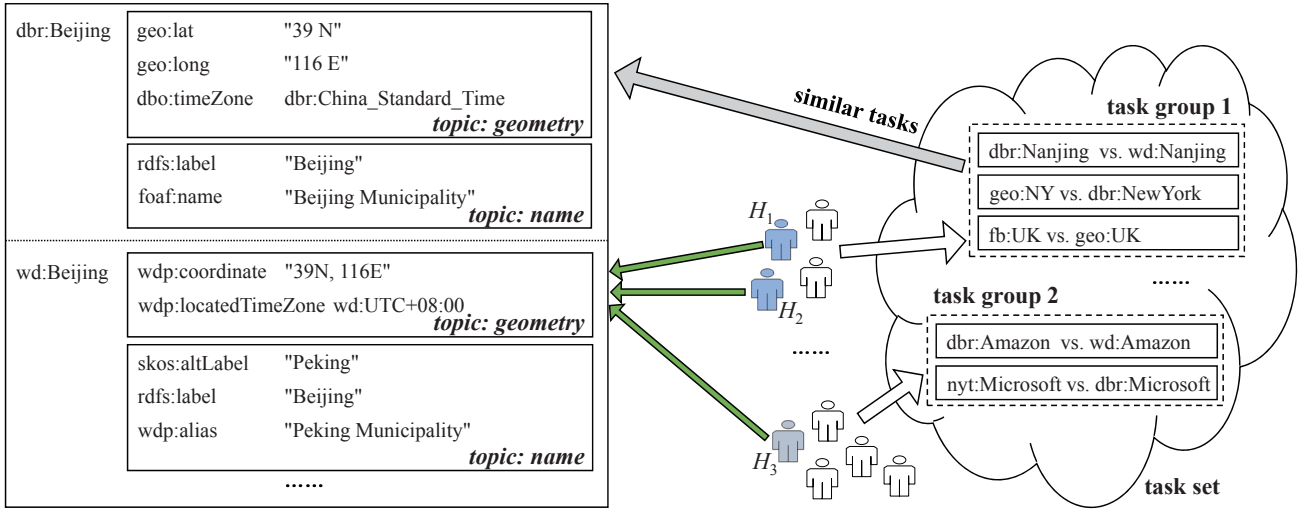
提纲

- 简介
- 近期工作介绍
  - 任务设计
  - 用户建模
- 总结与展望

Motivation

- **Truth inference**
  - identify the task truth from human judgments with inconsistency
- Previous work found that correctly estimating human expertise is crucial
  - A large number assume that humans have **consistent** expertise over all tasks
  - Humans may have **varied expertise on different topics**
    - e.g., Geography vs. Book
      - e.g., American geography vs. Chinese geography
- Our approach: topic modeling
  - model varied human expertise on latent topics
    - leverage similar task clustering for improvement
  - a probabilistic model
    - learn human varied expertise, compute task similarity and infer task truths integrally

Running example



- Tasks are formed into groups
  - e.g., TG1 describes “location” topics; TG2 describes “company” topics
- 4 humans have high expertise on TG1; 5 humans have high expertise on TG2
- For the current task about resolving two entities related to Beijing
  - Find its similar tasks from TG1, and identify that Beijing should have “location” topics
- Assume humans H1, H2, H3 provide judgments on the current task
  - Judgments of H1 and H2 would be more preferable

Topic-Expertise-Similar-Tasks Model

- Generative process

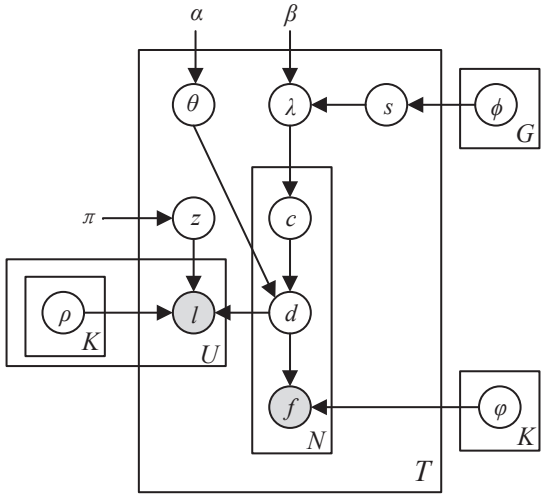
$$P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta}) = \prod_{t=1}^{T} P(z_t|\pi) P(\theta_t|\alpha) P(s_t|\phi_t) P(\lambda_t|\beta, A^t)$$

$$\prod_{n=1}^{N_t} P(c_{t,n}|\lambda_t) P(d_{t,n}|\boldsymbol{\theta}, c_{t,n}) P(f_{t,n}|\varphi_{d_{t,n}})$$

$$\prod_{u=1}^{U_t} P(l_{t,u}|z_t, d_t, \rho_u)$$

  - parameters: $\boldsymbol{\Theta} = \{\alpha, \beta, \pi, \boldsymbol{\phi}, \boldsymbol{\varphi}, \boldsymbol{\rho}\}$
  - observed variables: $\boldsymbol{l}, \boldsymbol{f}$
  - hidden variables: $\boldsymbol{\Delta} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{s}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{z}\}$

- Computing the posterior distribution of the hidden variables is intractable for exact inference
  - **Variational EM algorithm**
    - E-step: fix model parameters, update variational parameters
    - M-step: optimize model parameters



Symbol	Description
R_t	t-th task (incl. two entities and an unknown truth)
z_t	unknown truth for t-th task
H_u	u-th human
$l_{t,u}$	judgment given by u-th human for t-th task
$f_{t,n}$	n-th feature for t-th task
$d_{t,n}$	topic assigned to feature $f_{t,n}$
$\rho_{u,k}$	expertise of u-th human on k-th topic
s_t	task group of t-th task
A^t	indices of the tasks that are in the same group as R_t
$\phi_{g,t}$	probability of t-th task belonging to g-th task group
φ_k	multinomial dist. over features specific to k-th topic
θ_t	multinomial dist. over topics specific to t-th task
λ_t	multinomial dist. over tasks specific to t-th task
$c_{t,n}$	latent variable from which $d_{t,n}$ is sampled, and its value is a task in group s_t
α, β	Dirichlet priors to multinomial dist. θ, λ, resp.
π	prior probability of a task truth

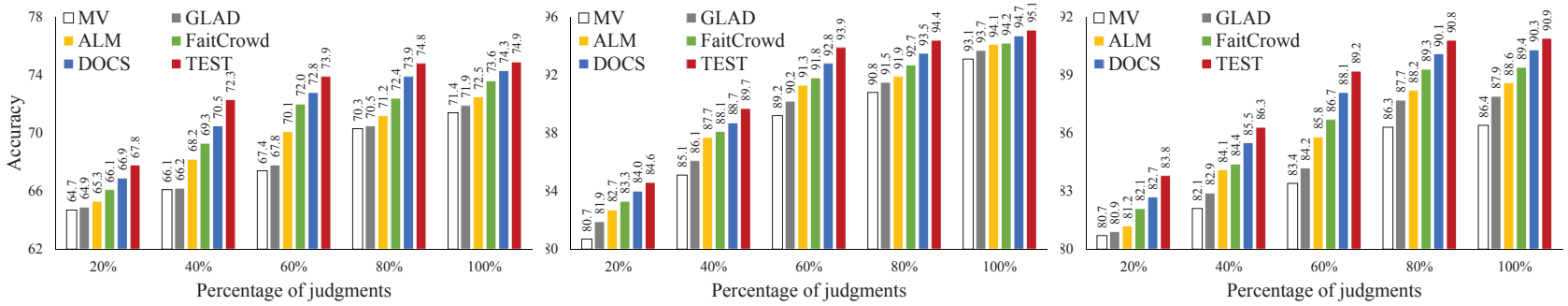
Evaluation

	SView	BTC	Synthetic
ER tasks	4,124	4,324	120,000
Task domains	14	8	3
Human workers	52	33	200
Judgments	18,679	15,996	600,000

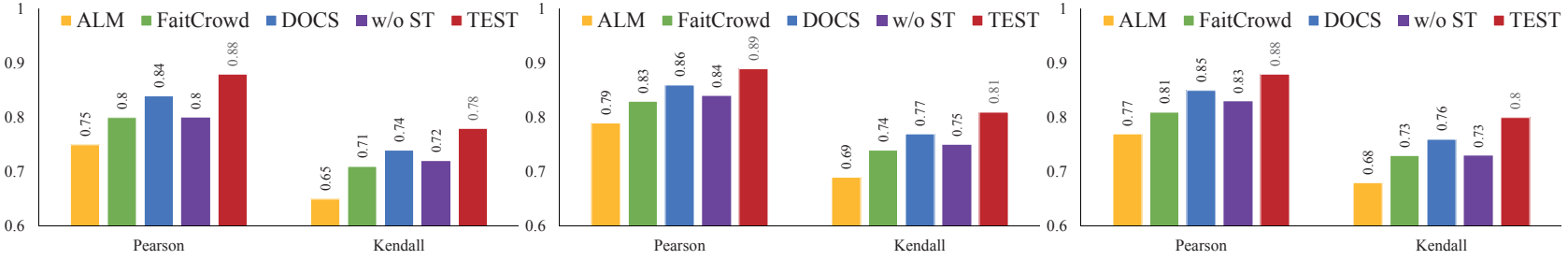
- Datasets
  - Real-world: SView, BTC
  - Synthetic: mix Product, Restaurant, Cora and DBLP-Scholar
- Comparative approaches
  - Baseline: majority voting (MV)
  - State-of-the-art
    - **GLAD:** consistent expertise
    - **ALM:** represent human expertise by a weighted combination of task features
    - **FaitCrowd:** a single topic to each task
    - **DOCS:** use the categories of linked Freebase entities as task domains

Experimental results

- Task truth inference
  - **1.1%** higher than the second best method



- Human expertise estimation
  - **5.3%** higher than the second best method



提纲

- 简介
- 近期工作介绍
- 总结与展望

总结与展望

- 实体链接已经得到了广泛而深入的研究
- 未来研究方向
  1. 表示学习
     - 复杂关系、文字属性
  2. 人机协作
     - 在线链接、用户专长
  3. 评测数据集
     - 大规模、图特征、跨语言

……

参考文献

1. G. Cheng, D. Xu, Y. Qu. C3D+P: a summarization method for interactive entity resolution. Journal of Web Semantics, 2015
2. K. Gunaratna, K. Thirunarayan, A. Sheth. FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In: AAAI, 2015
3. M. Fang, J. Yin, D. Tao. Active learning for crowdsourcing using knowledge transfer. In: AAAI, 2014
4. S. Gong, **W. Hu**, W. Ge, Y. Qu. Modeling topic-based human expertise for crowd entity resolution. Journal of Computer Science and Technology, 2018
5. J. Huang, **W. Hu**, H. Li, Y. Qu. Automated comparative table generation for facilitating human intervention in multi-entity resolution. In: SIGIR, 2018
6. F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, J. Han. FaitCrowd: fine grained truth discovery for crowdsourced data aggregation. In: KDD, 2015
7. V. Verroios, H. Garcia-Molina, Y. Papakonstantinou. Waldo: an adaptive human interface for crowd entity resolution. In: SIGMOD, 2017
8. Y. Zheng, G. Li, R. Cheng. DOCS: domain-aware crowdsourcing system. In: VLDB, 2016

谢谢！

胡伟 (whu@nju.edu.cn)

致谢：江苏省人工智能学会