RELIN: Relatedness and Informativeness-based Centrality for Entity Summarization

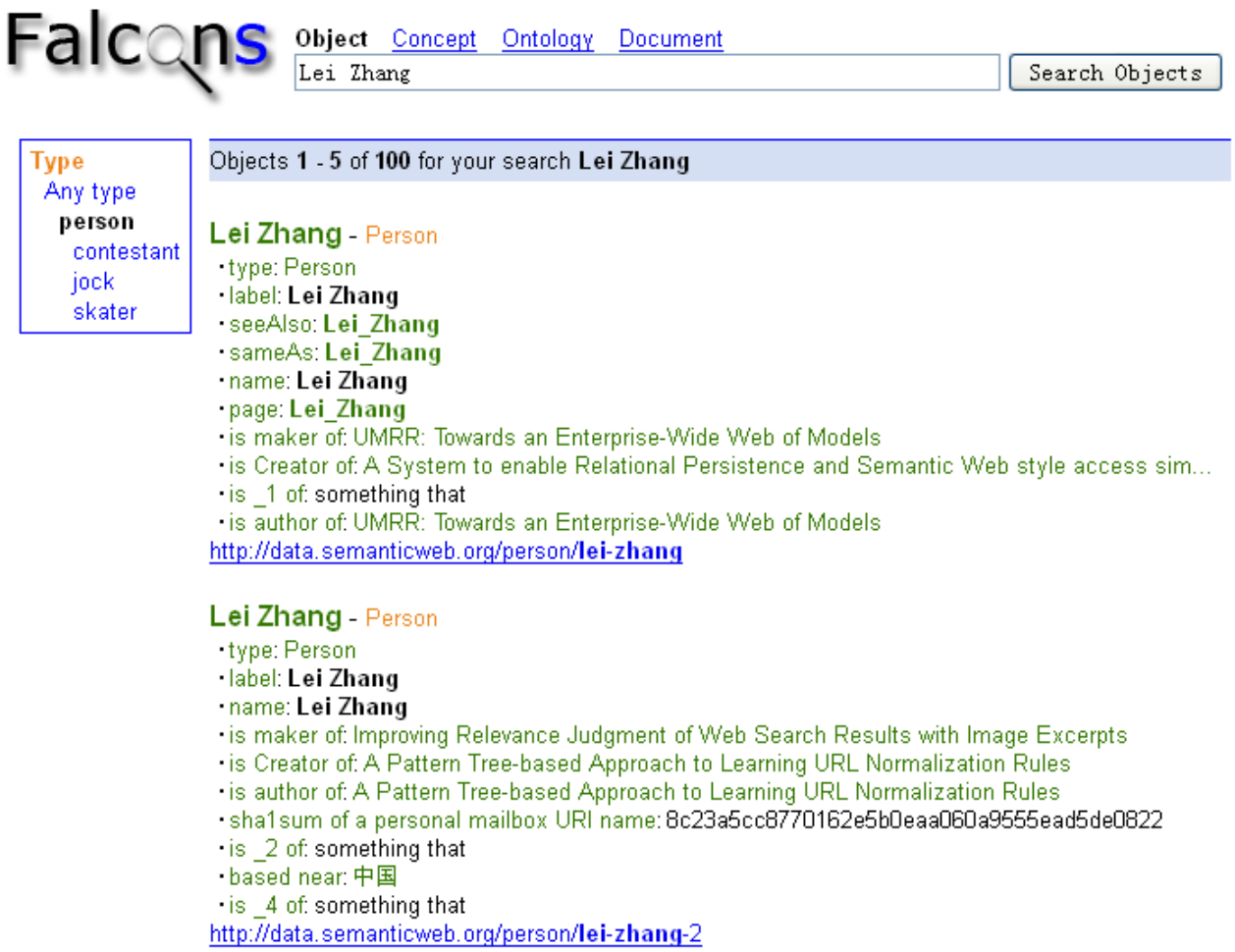
Gong Cheng[1], Thanh Tran[2], Yuzhong Qu[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Institute AIFB, Karlsruhe Institute of Technology, Germany
gcheng@nju.edu.cn

Presented at ISWC2011

Motivation

- DBpedia describes 3.64M entities with 1B RDF triples.
  - 1B/3.64M = 281 RDF triples per entity
- A piece of lengthy entity description is unacceptable in tasks that require quick identification of the underlying entity.

Entity search --- find entities that match an information need



Pay-as-you-go data integration --- judge whether two entities denote the same

- type: Person
- label: **Lei Zhang**
- seeAlso: **Lei_Zhang**
- sameAs: **Lei_Zhang**
- name: **Lei Zhang**
- page: **Lei_Zhang**
- is maker of: UMRR: Towards an Enterprise-Wide Web of Models
- is Creator of: A System to enable Relational Persistence and Semantic Web style access sim...
- is _1 of: something that
- is author of: UMRR: Towards an Enterprise-Wide Web of Models

- type: Person
- label: **Lei Zhang**
- name: **Lei Zhang**
- is maker of: Improving Relevance Judgment of Web Search Results with Image Excerpts
- is Creator of: A Pattern Tree-based Approach to Learning URL Normalization Rules
- is author of: A Pattern Tree-based Approach to Learning URL Normalization Rules
- sha1sum of a personal mailbox URI name: 8c23a5cc8770162e5b0eaa060a9555ead5de0822
- is _2 of: something that
- based near: 中国
- is _4 of: something that

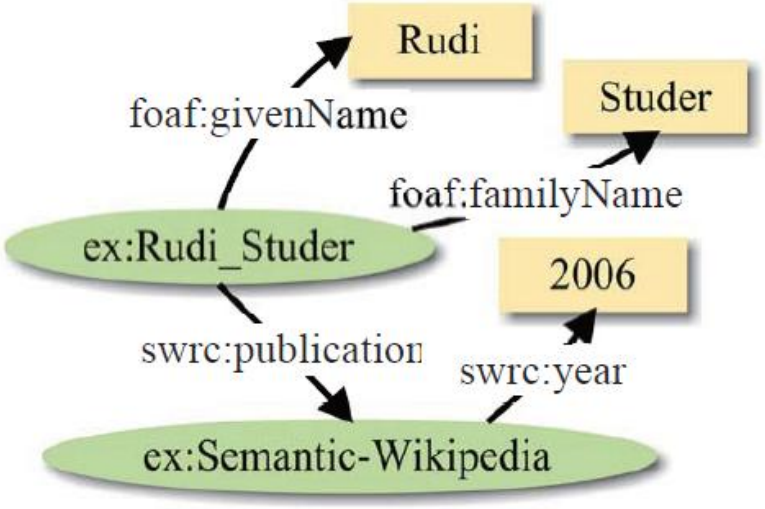
Motivation

- DBpedia describes 3.64M entities with 1B RDF triples.
  - 1B/3.64M = 281 RDF triples per entity
- A piece of lengthy entity description is unacceptable in tasks that require quick identification of the underlying entity.
- Problem: to summarize lengthy entity descriptions

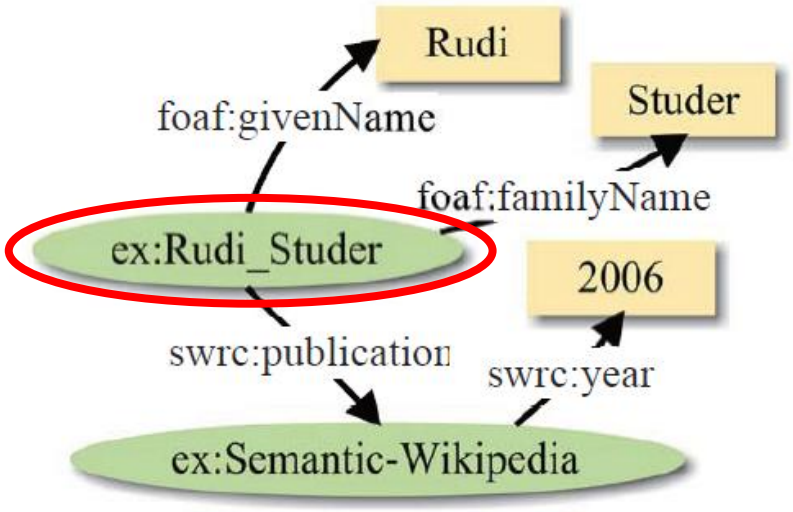
Outline

- Problem statement
- The RELIN model
- Implementation
- Experiments
- Conclusions

Data graph


Rudi
Studer
foaf:givenName
foaf:familyName
ex:Rudi_Studer
2006
swrc:publication
swrc:year
ex:Semantic-Wikipedia

Feature set



FS(ex:Rudi_Studer)	
f_1	<foaf:givenName, "Rudi">
f_2	<foaf:familyName, "Studer">
f_3	<swrc:publication, ex:Semantic-Wikipedia>

Entity summarization

- Entity summarization = feature ranking
- Entity summary = *k* top-ranked features

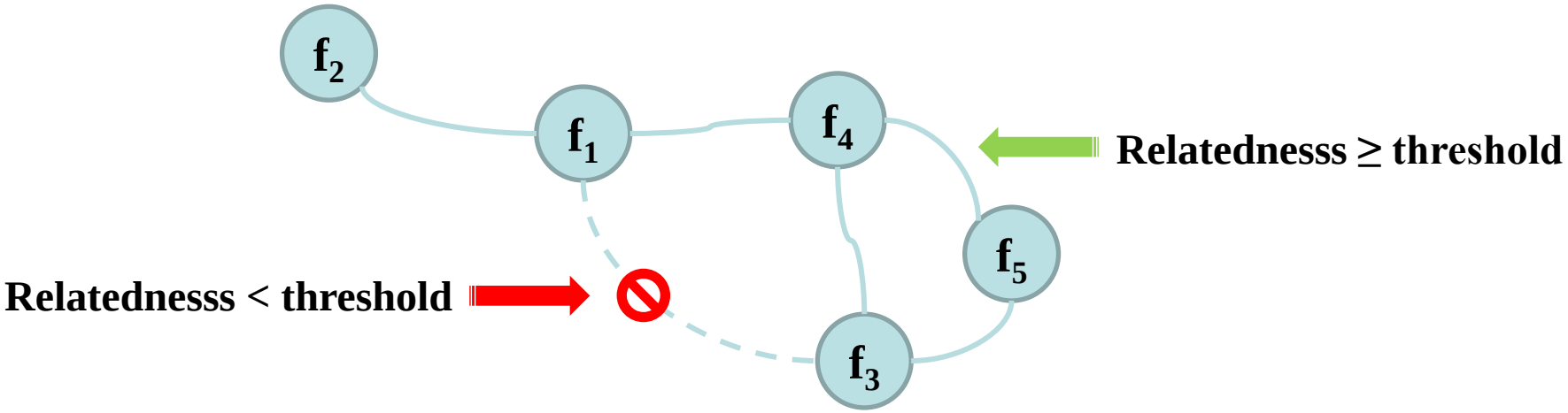
FS(ex:Rudi_Studer)	
f_1	<foaf:givenName, "Rudi">
f_2	<foaf:familyName, "Studer">
f_3	<swrc:publication, ex:Semantic-Wikipedia>

Outline

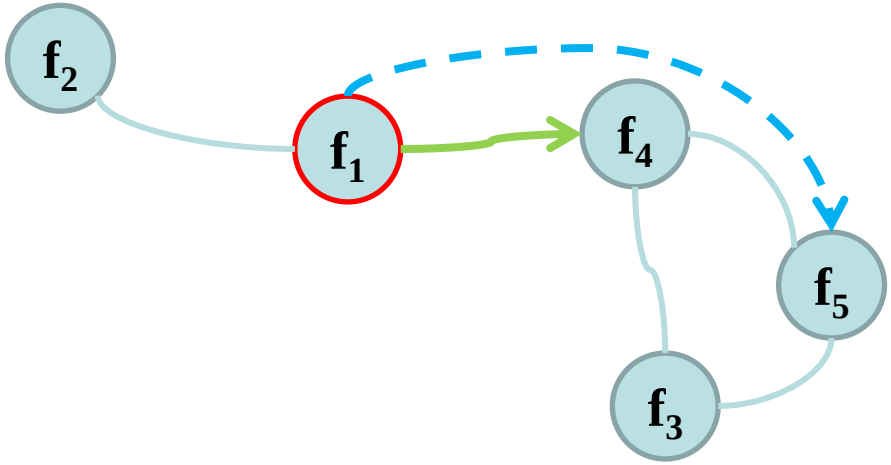
- Problem statement
- The RELIN model
- Implementation
- Experiments
- Conclusions

Centrality-based ranking: concepts

- Widely applied to text summarization and ontology summarization
- By constructing a graph
  - Nodes: data elements to be ranked
  - Edges: connecting related nodes
- and then, measuring node centrality
  - e.g. degree, PageRank, …

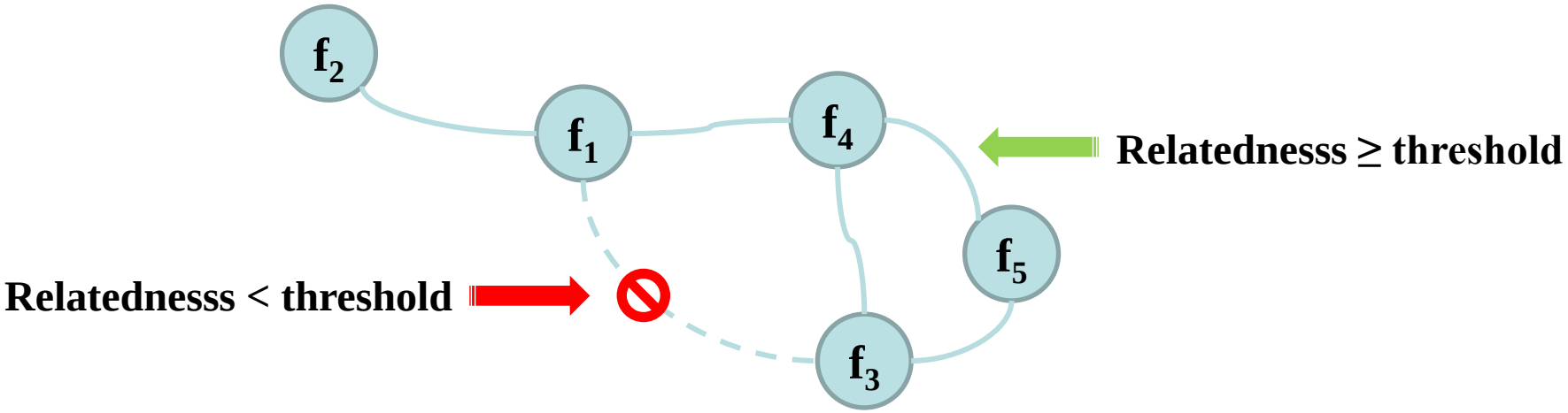


- Simulating a random surfer's behavior who navigates from node to node
- Two types of action
  - Following a random edge (with a uniform probability distribution)
  - Jumping at random (with a uniform probability distribution)
- Ranking based on the stationary distribution of such a Markov chain

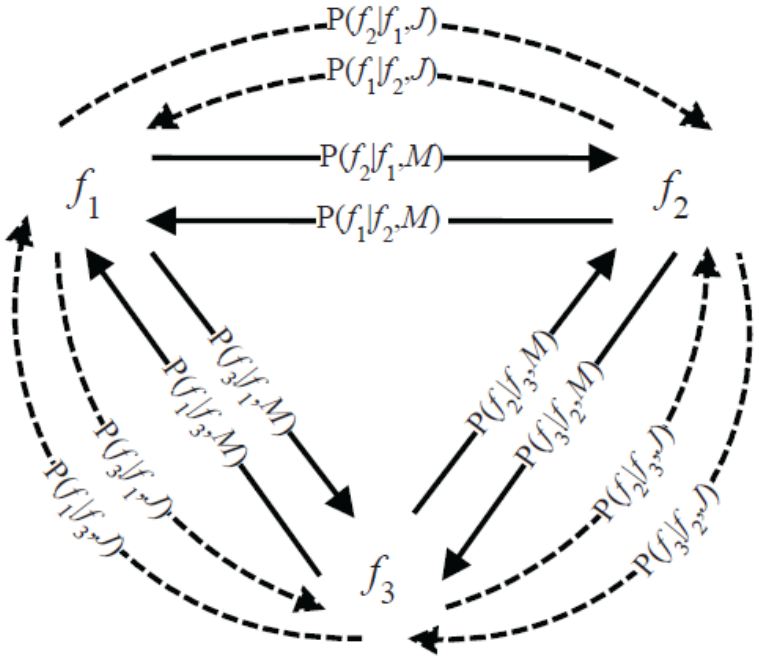


Centrality-based ranking for entity summarization: problems

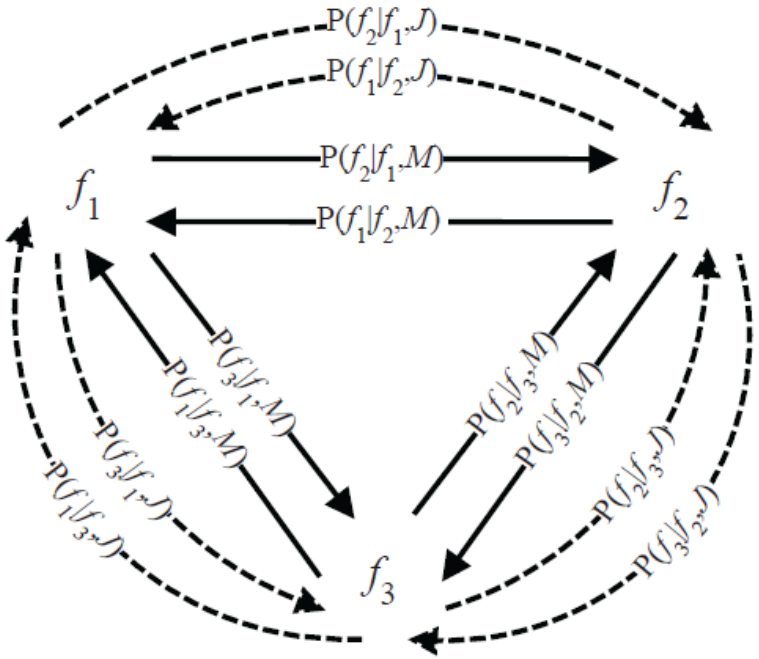
- How to define a good feature
  - Not only capturing the main themes of the entity description
  - But also distinguishing the entity from others
- Loss of information
  - Float-valued function → boolean-valued function



- An extension of PageRank
  - Following a random edge (⟶)
    within a complete graph, with a probability proportional to the relatedness between the two associated nodes, i.e. no threshold needed
  - Jumping at random (⇢)
    with a probability proportional to the amount of information carried by the target that helps to identify the entity



- Two kinds of action
  - **Relational move** --- more likely to a feature that carries related information about the theme currently under investigation
  - **Informational jump** --- more likely to a feature that provides a large amount of new information for clarifying the identity of the underlying entity
- Two non-uniform probability distributions



- Actions (given the current feature f_q)
  - $P(M|f_q)$: the probability of performing a relational move from f_q
  - $P(J|f_q)$: the probability of performing an informational jump from f_q
  - subject to $P(M|f_q) + P(J|f_q) = 1$
- Targets for actions (given FS the feature set)
  - $P(f_p|f_q,M)$: the probability of performing a relational move from f_q to f_p
  - $P(f_p|f_q,J)$: the probability of performing an informational jump from f_q to f_p
  - subject to $\sum_{f_p \in FS} P(f_p \mid f_q, M) = 1$ and $\sum_{f_p \in FS} P(f_p \mid f_q, J) = 1$
- Result
  - $\mathbf{x}(t)$: |FS|-dimensional vector
  - $\mathbf{x}_p(t)$: the probability that the surfer visits f_p at step t
  - Finally,

$$\mathbf{x}_p(t+1) = \sum_{f_q \in FS} \mathbf{x}_q(t) \cdot \left(P(M \mid f_q) P(f_p \mid f_q, M) + P(J \mid f_q) P(f_p \mid f_q, J) \right)$$

  - and

$$\lim_{t \to \infty} \mathbf{x}(t) = \mathbf{x}^*$$

Outline

- Problem statement
- The RELIN model
- Implementation
- Experiments
- Conclusions

Actions

- $P(M|f_q) = 1 - \lambda$
- $P(J|f_q) = \lambda$
- λ: to be tuned in experiments

Relatedness --- P(f_p|f_q,M)

- Relatedness between features (i.e. property-value pairs) combines
  - Relatedness between properties (i.e. resources)
  - Relatedness between values (i.e. resources)
- Relatedness between resources = relatedness between resource names
  - URI: label or local name
  - Literal: lexical form
- Distributional relatedness between resource names
  - More related = more often co-occur in certain contexts (e.g. documents)
- Estimated via "pointwise mutual information + Google"

$$\mathrm{PMI}(s_i, s_j) = \log \frac{\mathrm{P}(s_i, s_j)}{\mathrm{P}(s_i) \cdot \mathrm{P}(s_j)}$$

$$\mathrm{P}(s_i, s_j) = \frac{\mathrm{Hits}(s_i, s_j)}{N}$$

$$\mathrm{P}(s_j) = \frac{\mathrm{Hits}(s_j)}{N}$$

Informativeness --- P(f_p|f_q,J)

- Self-information

$$\mathrm{SelfInfo}(o) = -\log(\mathrm{P}(o))$$

- o: informational jump from f_q to f_p
- P(f_p|f_q): the probability that f_p belongs to a feature set given f_q also does so
- Estimated via a statistical analysis of the data set

$$\mathrm{P}(f_p|f_q) = \frac{|\{e \in E \mid f_p, f_q \in \mathrm{FS}(e)\}|}{|\{e \in E \mid f_q \in \mathrm{FS}(e)\}|}$$

- Approximation: P(f_p|f_q) = P(f_p)

Outline

- Problem statement
- The RELIN model
- Implementation
- Experiments
- Conclusions

Experiments

- Intrinsic evaluation
- Extrinsic evaluation

Intrinsic evaluation --- design

- Task
  - To manually construct ideal entity summaries as the gold standard
- Participants
  - 24 students majoring in computer science
- Test cases
  - 149 entity descriptions randomly selected from DBpedia 3.4
- Assignment
  - 4.43 participants per entity description
- Output
  - Top-5 features
  - Top-10 features

Intrinsic evaluation --- results

- Metric: overlap between summaries

- Agreement between participants about ideal summaries
  - 2.91 when *k*=5
  - 7.86 when *k*=10
- Quality of summaries computed under different approach settings

		$k = 10$	$k = 5$
Baselines	*OntoSum*	3.69	1.01
	RandomRank	3.36	0.76
Ours	RELIN, with $\lambda = 0.00$	3.58	1.61
	RELIN, with $\lambda = 0.15$	3.84	1.73
	RELIN, with $\lambda = 0.50$	4.40	1.99
	RELIN, with $\lambda = 0.85$	**4.88**	2.29
	RELIN, with $\lambda = 1.00$	4.86	**2.40**

Extrinsic evaluation --- design

- Task
  - To manually confirm entity mappings by using summaries
- Participants
  - 19 students majoring in computer science
- Test cases
  - 47 pairs of entity descriptions (DBpedia 3.4 ↔ Freebase Dec. 2009)
  - Gold-standard judgments based on owl:sameAs links
    - 24 correct and 23 incorrect
- Assignment
  - 3.62 participants per pair, per approach setting
- Output
  - Judgment: correct or incorrect

Extrinsic evaluation --- results

- Metrics
  - Accuracy of the judgments
    - 1.0 = consistent with the gold standard
    - 0.0 = inconsistent
  - Time spent
    - Normalized by the average time per judgment spent by the participant
    - 1.0 = medium efficiency
    - Smaller value = higher efficiency

- Results

	k	Accuracy	Time
OntoSum	5	0.56	**0.84**
RandomRank	5	0.60	**0.87**
RELIN, with $\lambda = 0.85$	5	**0.70**	**0.92**
RELIN, with $\lambda = 0.85$	10	**0.68**	1.12
ReturnsAll	n/a	0.60	1.41

Discussion

- Automatically computed summaries are still not as good as handcrafted ones.

	k=5	*k*=10
Agreement between ideal summaries	2.91	7.86
Agreement between computed summaries and ideal summaries	2.40	4.88

- User-specific notion of informativeness
  - Longitude and latitude are highly informative, but …
- Information redundancy
  - Longitude + latitude = point
  - What if multiple sources …
- Summarization = what + how (to present)

Outline

- Problem statement
- The RELIN model
- Implementation
- Experiments
- Conclusions

Conclusions

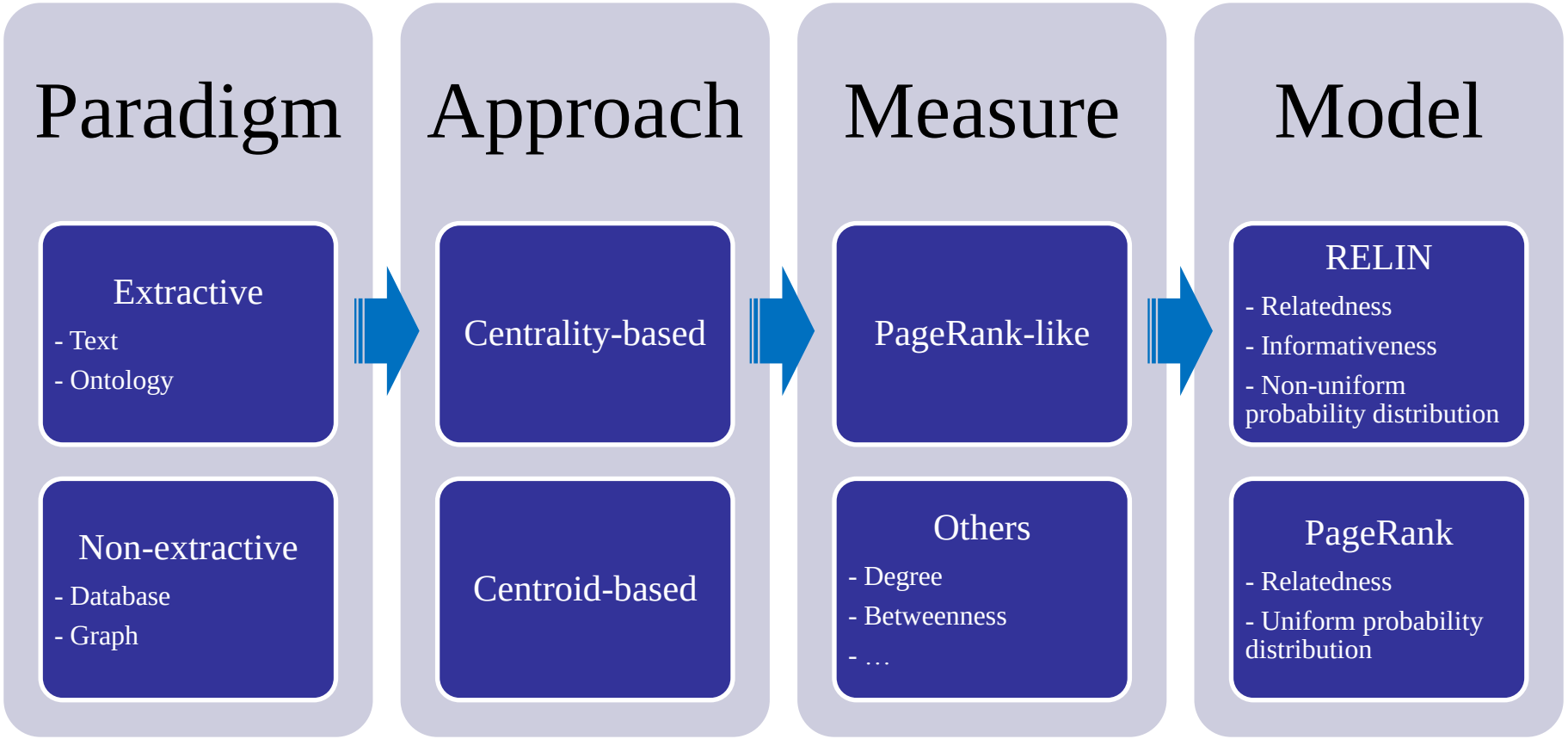
- Problem of entity summarization
  - Extractive
  - About identifying the entity that underlies a lengthy description
- The RELIN model
  - Variant of the random surfer model
  - Non-uniform probability distributions
  - Informativeness + relatedness
- Implementation
  - Based on linguistic and information theory concepts
  - Using information captured by the labels of nodes and edges in the data graph
- Experiments
  - Closer to handcrafted ideal summaries
  - Assisting users in confirming entity mappings more accurately

Future work --- application-specific entity summarization

- type: Person
- label: **Lei Zhang**
- seeAlso: **Lei_Zhang**
- sameAs: **Lei_Zhang**
- name: **Lei Zhang**
- page: **Lei_Zhang**
- is maker of: UMRR: Towards an Enterprise-Wide Web of Models
- is Creator of: A System to enable Relational Persistence and Semantic Web style access sim...
- is _1 of: something that
- is author of: UMRR: Towards an Enterprise-Wide Web of Models

- type: Person
- label: **Lei Zhang**
- name: **Lei Zhang**
- is maker of: Improving Relevance Judgment of Web Search Results with Image Excerpts
- is Creator of: A Pattern Tree-based Approach to Learning URL Normalization Rules
- is author of: A Pattern Tree-based Approach to Learning URL Normalization Rules
- sha1sum of a personal mailbox URI name: 8c23a5cc8770162e5b0eaa060a9555ead5de0822
- is _2 of: something that
- based near: 中国
- is _4 of: something that

Related work --- summarization



Related work --- ranking

- Different goals --- to best identify the underlying entity
  - B. Aleman-Meza et al., Ranking Complex Relationships on the Semantic Web. IEEE Internet Comput. 2005.
  - R. Delbru et al., Hierarchical Link Analysis for Ranking Web Data. ESWC 2010.
  - T. Franz. TripleRank: Ranking Semantic Web Data By Tensor Decomposition. ISWC 2009.
  - …
- Exploitation of data semantics at different levels --- use labels of nodes and edges
  - T. Penin et al., Snippet Generation for Semantic Web Search Engines. ASWC 2009.
  - X. Zhang et al., Ontology Summarization Based on RDF Sentence Graph. WWW 2007.
  - …