KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks

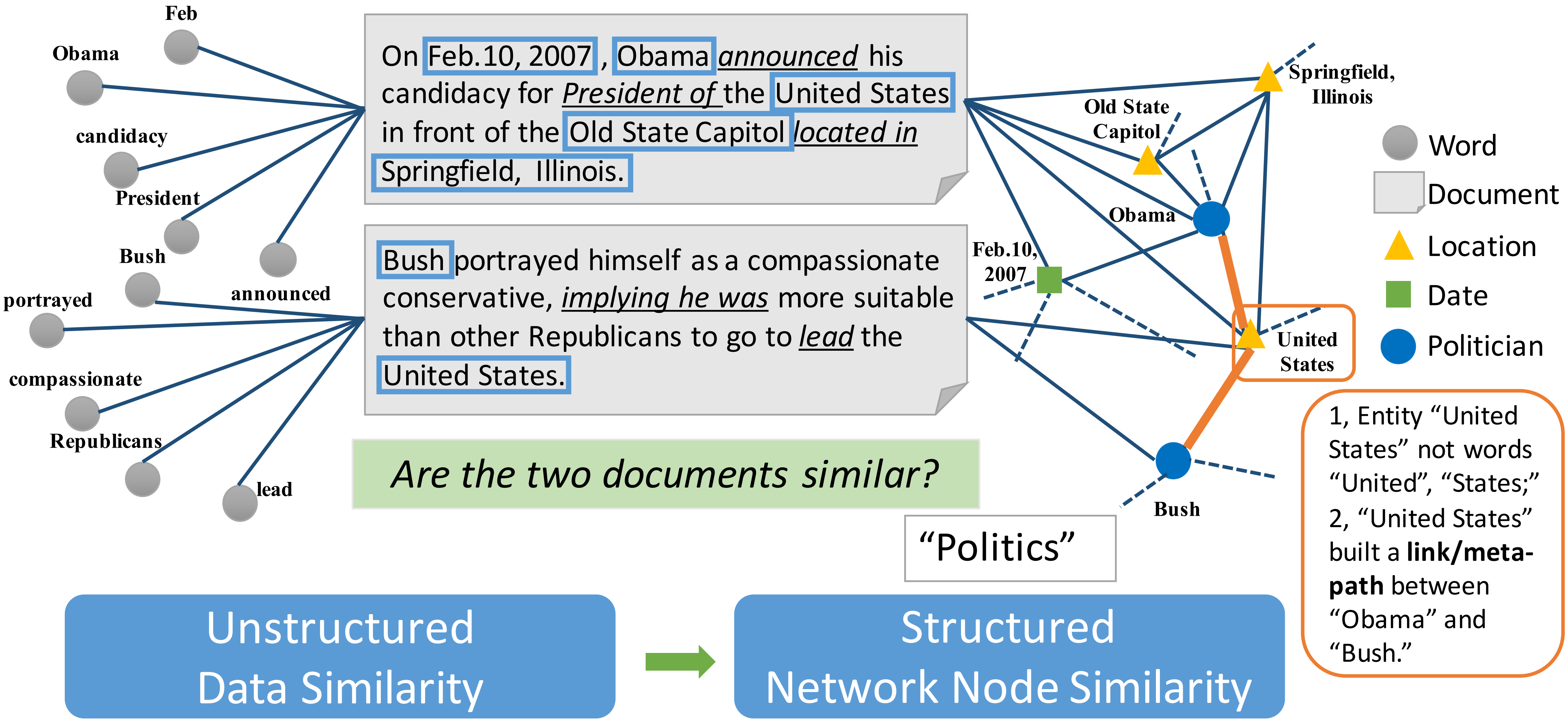
ICDM'15 Atlantic City, USA

Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, Jiawei Han

Outline

Motivation	The problem of current similarity measures.
KnowSim	Our way for computing similarity.
Experiments	The results on benchmark datasets.

Motivation



Feb
Obama
candidacy
President
Bush
announced
portrayed
compassionate
Republicans
lead
On Feb.10, 2007 , Obama announced his candidacy for President of the United States in front of the Old State Capitol located in Springfield, Illinois.
Bush portrayed himself as a compassionate conservative, implying he was more suitable than other Republicans to go to lead the United States.
Are the two documents similar?
Springfield, Illinois
Old State Capitol
Obama
Feb.10, 2007
United States
Bush
"Politics"
Word
Document
Location
Date
Politician
1, Entity "United States" not words "United", "States;"
2, "United States" built a link/meta-path between "Obama" and "Bush."
Unstructured Data Similarity
Structured Network Node Similarity

Document-Based Heterogeneous Information Network Construction

- Machine learning with world knowledge framework [Wang et al. KDD'15]

Documents

World knowledge bases

World Knowledge Specification

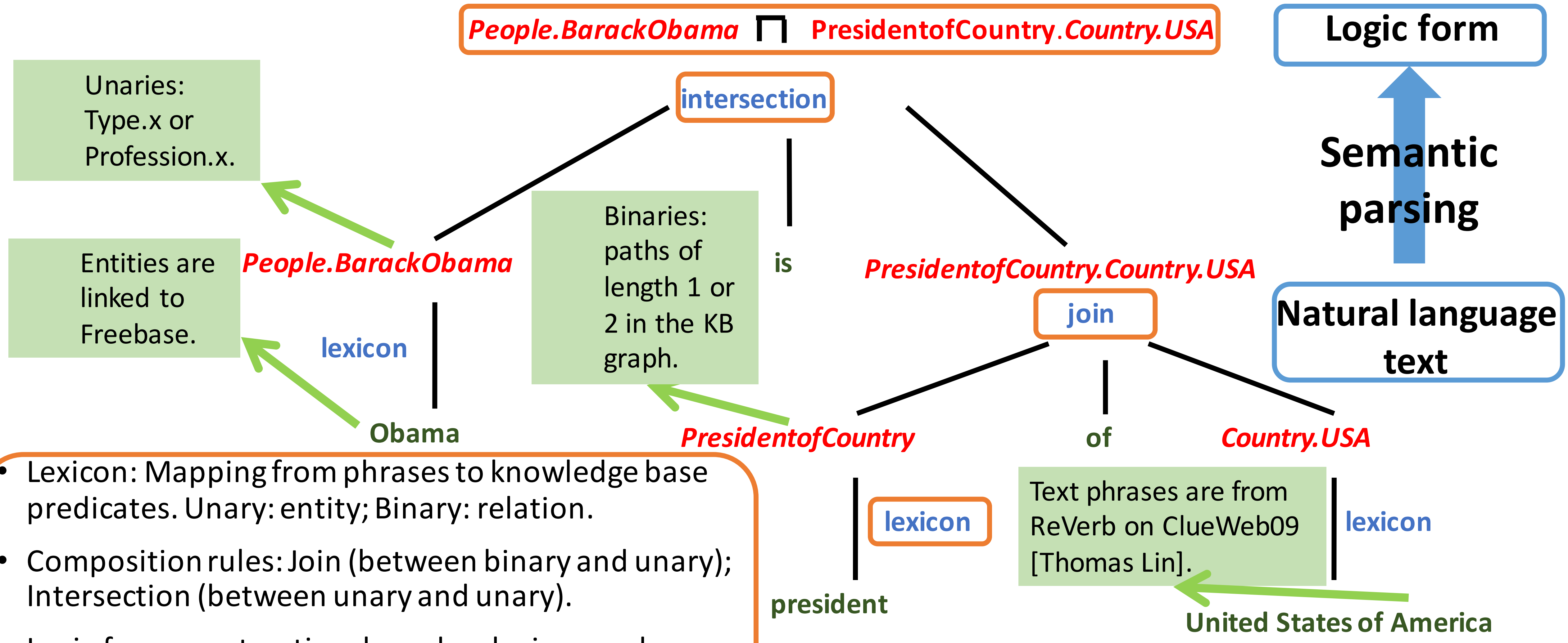
General purpose problem vs. Domain specific problem

Specified World Knowledge Representation

Knowledge representation vs. Data representation

C. Wang et al. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15

World Knowledge Specification: Unsupervised Semantic Parsing for Documents



- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).
- Logic form construction: based on lexicon and composition rules recursively.

World Knowledge Specification: Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple　　　adobe

software company, brand, fruit　　　brand, software company

software company, brand

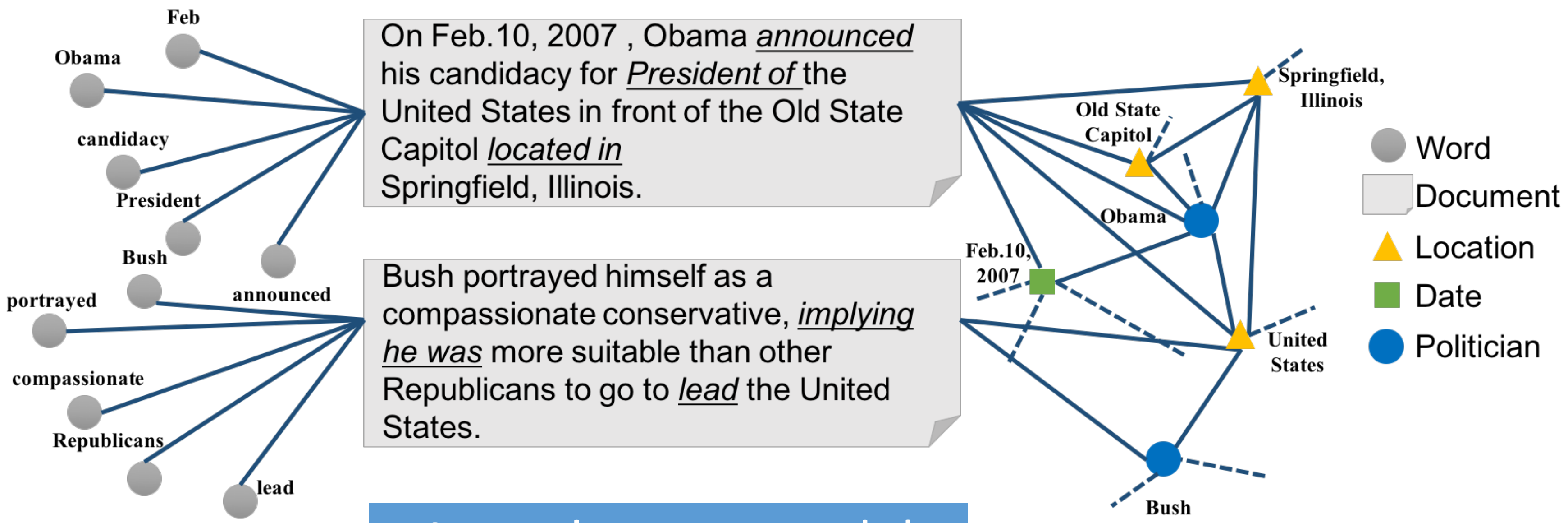
largest probability ones are selected　P(type | related entities)　A cluster of entities of type features

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI'11.

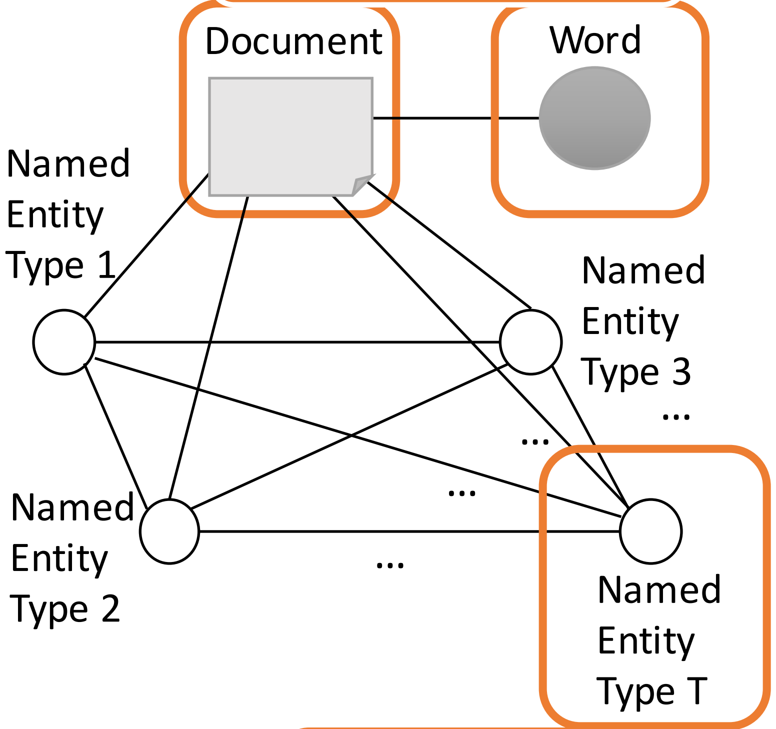
Specified World Knowledge Representation: Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

Two entity types in document-based HIN.



A good way to model real world data!

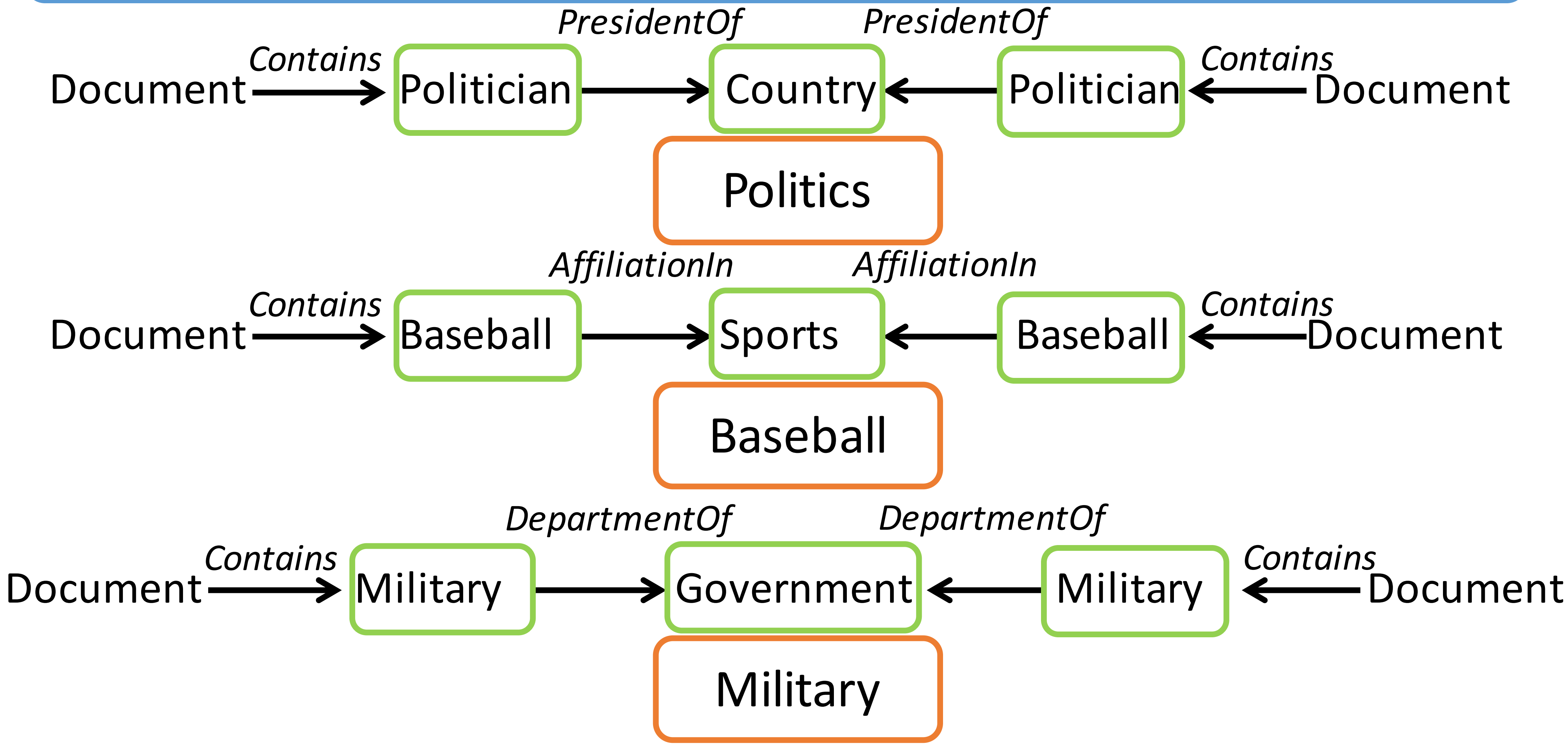


Represent the type of the name in text, e.g, person name.
NOT entity type (node type in HIN).

Network schema: High-level description of a network.

Meta-Path

Meta-path: A **path/link** in the network schema. [Sun et al., 2011]



Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

KnowSim

KnowSim: An unstructured data similarity measure defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

$$KS(d_i, d_j) = \frac{2 \times \boxed{\sum_m^{M'} w_m |\{p_{i \to j} \in P_m\}|}}{\boxed{\sum_m^{M'} w_m |\{p_{i \to i} \in P_m\}| + \sum_m^{M'} w_m |\{p_{j \to j} \in P_m\}|}}$$

Semantic broadness: the number of total meta-paths between themselves.

- Intuition: The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.
- KnowSim is computed in nearly linear time.

Challenges

Number of meta-paths could be very large.

$$KS(d_i,d_j)=\frac{2\times\sum_{m}^{\boxed{M'}} w_m\,|\{p_{i\to j}\in P_m\}|}{\sum_{m}^{M'}\boxed{w_m}\,|\{p_{i\to i}\in P_m\}|+\sum_{m}^{M'} w_m\,|\{p_{j\to j}\in P_m\}|}$$

The weight/importance of each meta-path is different when the domain is different.

#1: How should we generate the large number of meta-paths at the same time?

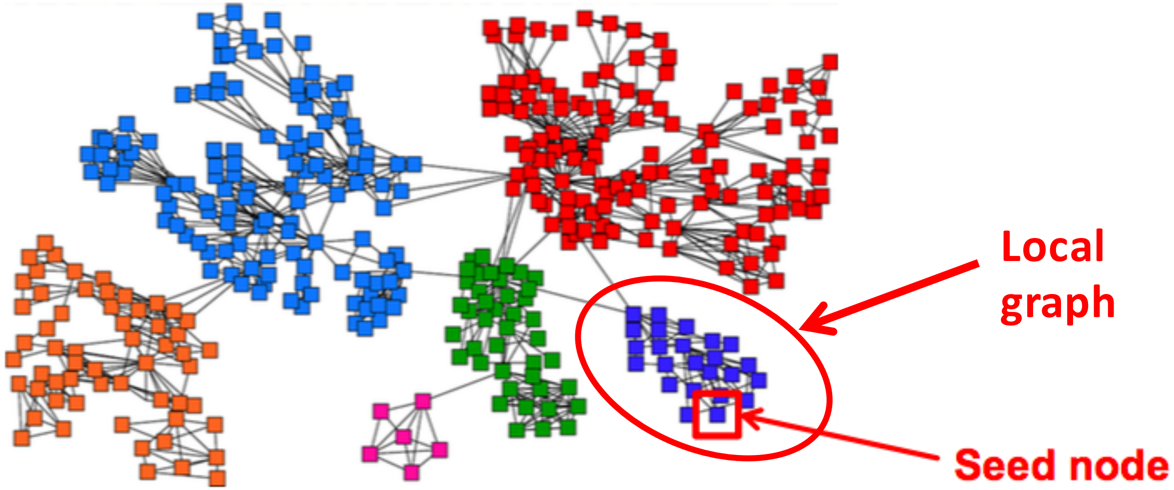
Previous studies only focus on single meta-path, enumeration over the network is OK. In real world, what will happen when thousands of meta-paths are needed?

#2: How should we decide the weight of each meta-path?

Previous studies treat them equally. In real world, different meta-path should contribute differently in various domains.

Meta-Path Dependent Random Walk

Intuition: Discovering compact sub-graph based on seed document nodes.



- Compute Personalized PageRank around seed nodes.
- The random walk will get trapped inside the blue sub-graph.

- Algorithm outline
  - Run **PPR** (approximate connectivity to seed nodes) with teleport set = {**S**}
  - Sort the nodes by the decreasing **PPR** score
  - **Sweep** over the nodes and find compact **sub-graph**.
  - Use the sub-graph instead of the whole graph to compute # of meta-paths between nodes.

Meta-Path Selection

- Maximal Spanning Tree based Selection [Sahami, 1998]
  - Intuition: meta-paths that only weakly influence the remaining domain variables are candidates for elimination (Select meta-paths with the largest dependencies with others).

$$\frac{\sum_{j \neq i}^{M} \cos(\boldsymbol{D}_{.,j_1}, \boldsymbol{D}_{.,j_2})}{M-1}$$

- Laplacian Score based Selection [He, 2006]
  - Intuition: Laplacian score represents the power of a meta-path in discriminating documents from different clusters.

$$L_j = \frac{\widetilde{\boldsymbol{D}_{.,j}}^T \boldsymbol{L} \boldsymbol{D}_{.,j}}{\widetilde{\boldsymbol{D}_{.,j}}^T \Lambda \boldsymbol{D}_{.,j}}$$

Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
GCAT (Government/Social)	1	16	60,608

GCAT is top category in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				

The number is reported in [X. Dong et al. KDD'14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.

Text Similarity Results

Datasets	Similarity Measures	BOW	BOW+TOPIC	BOW+ENTITY	BOW+TOPIC+ENTITY
20NG	Cosine	0.2400	0.2713	0.2473	0.2768
	Jaccard	0.2352	0.2632	0.2369	0.2650
	Dice	0.2400	0.2712	0.2474	0.2767
KnowSim+UNI	0.2860	KnowSim+MST	0.2891	KnowSim+LAP	**0.2913 (+5.2%)**
GCAT	Cosine	0.3490	0.3639	0.2473	0.3128
	Jaccard	0.3313	0.3460	0.2319	0.2991
	Dice	0.3490	0.3638	0.2474	0.3156
KnowSim+UNI	0.3815	KnowSim+MST	0.3833	KnowSim+LAP	**0.4086 (+12.3%)**

Finding #1: Our method KnowSim is better than traditional measures.

KnowSim can better leverage world knowledge (entity, meta-path) rather than just treating them as flat features (e.g., BOW+ENTITIY).

Finding #2: More world knowledge will lead to better performance.

Laplacian score based meta-path selection method (KnowSim+LAP) performs the best.

Spectral Clustering Using KnowSim Matrix

Datasets	Similarity Measures	BOW	BOW+TOPIC	BOW+ENTITY	BOW+TOPIC+ENTITY
20NG	Cosine	0.3440	0.3461	0.3896	0.4247
	Jaccard	0.3547	0.3517	0.3850	0.4292
	Dice	0.3440	0.3457	0.3894	0.4248
KnowSim+UNI	0.4304	KnowSim+MST	0.4412	KnowSim+LAP	**0.4461 (+3.9%)**
GCAT	Cosine	0.3932	0.4352	0.2394	0.4106
	Jaccard	0.3887	0.4292	0.2497	0.4159
	Dice	0.3932	0.4355	0.2392	0.4112
KnowSim+UNI	0.4463	KnowSim+MST	0.4653	KnowSim+LAP	**0.4736(+8.8%)**

Finding: we can get the same results according to the clustering NMI.
KnowSim is a better similarity measure.
We can infer that KnowSim could have positive impact on other similarity-based applications, e.g., document classification and ranking.

Conclusion

Problem: Document similarity as network node similarity.

Approach: World knowledge specification;
KnowSim: unstructured data similarity defined on network.

Results: Document similarity results and its application (clustering) show the power.

Thank You! ☺