Distant Meta-Path Similarities for Text-Based Heterogeneous Information Networks

Chenguang Wang, Yangqiu Song, Haoran Li,

Yizhou Sun, Ming Zhang, and Jiawei Han

Outline

Motivation	Issues in text similarity computation
Distant Similarity	Compute text similarity with semantics
Experiments	Distant similarity is capable

Motivation

D_0

Michael Jordan is an American retired professional basketball player in the NBA.

D_1

A noted basketball fan, former President Barack Obama welcomed Steve Kerr from the greatest team in NBA history.

Are the two documents similar?

"Sports"

Yes

Compute Text Similarity Using Flat Feature

D0
Michael Jordan is an American retired professional basketball player in the NBA.

D1
A noted basketball fan, former President Barack Obama welcomed Steve Kerr from the greatest team in NBA history.

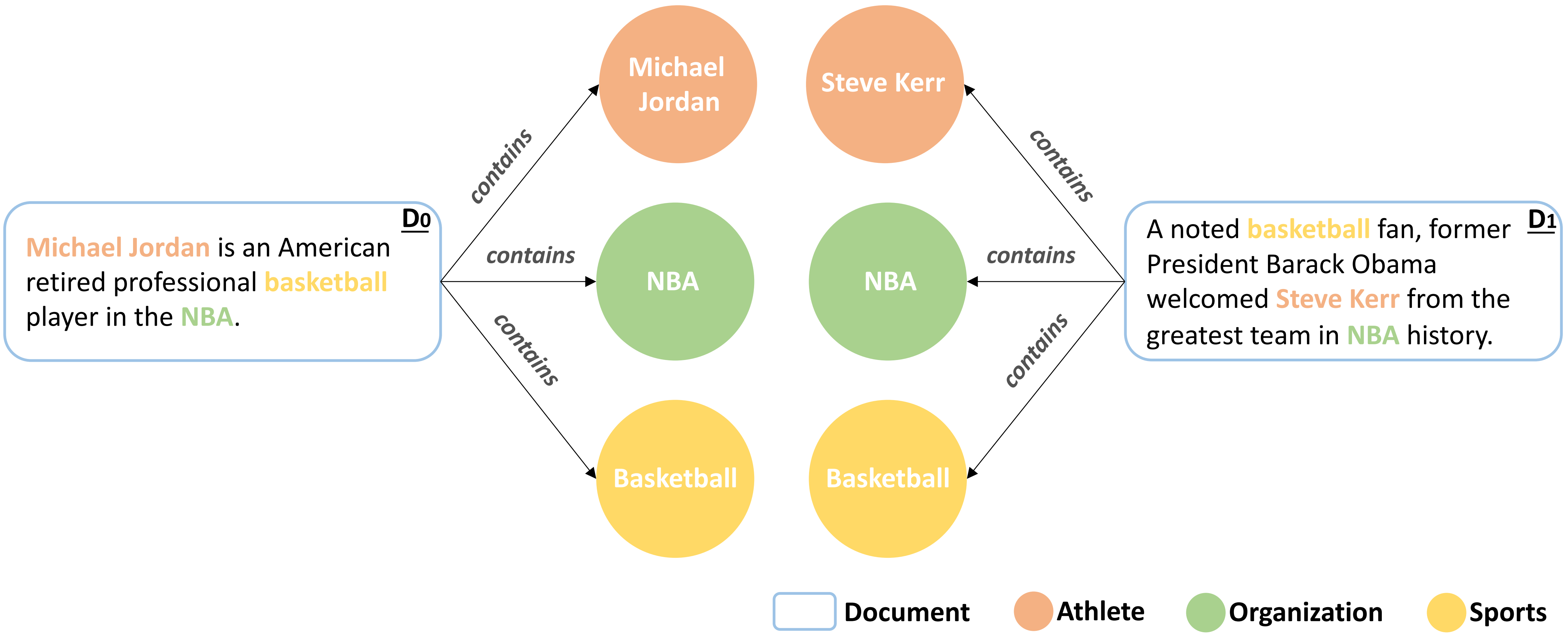
- Represent texts as flat feature vectors
  - e.g., bag of words

D0	w0	w1	w2	...	
D1	w0		w2	w3	...

Issue: Missing semantics!

Similarity based on flat features → Low similarity score

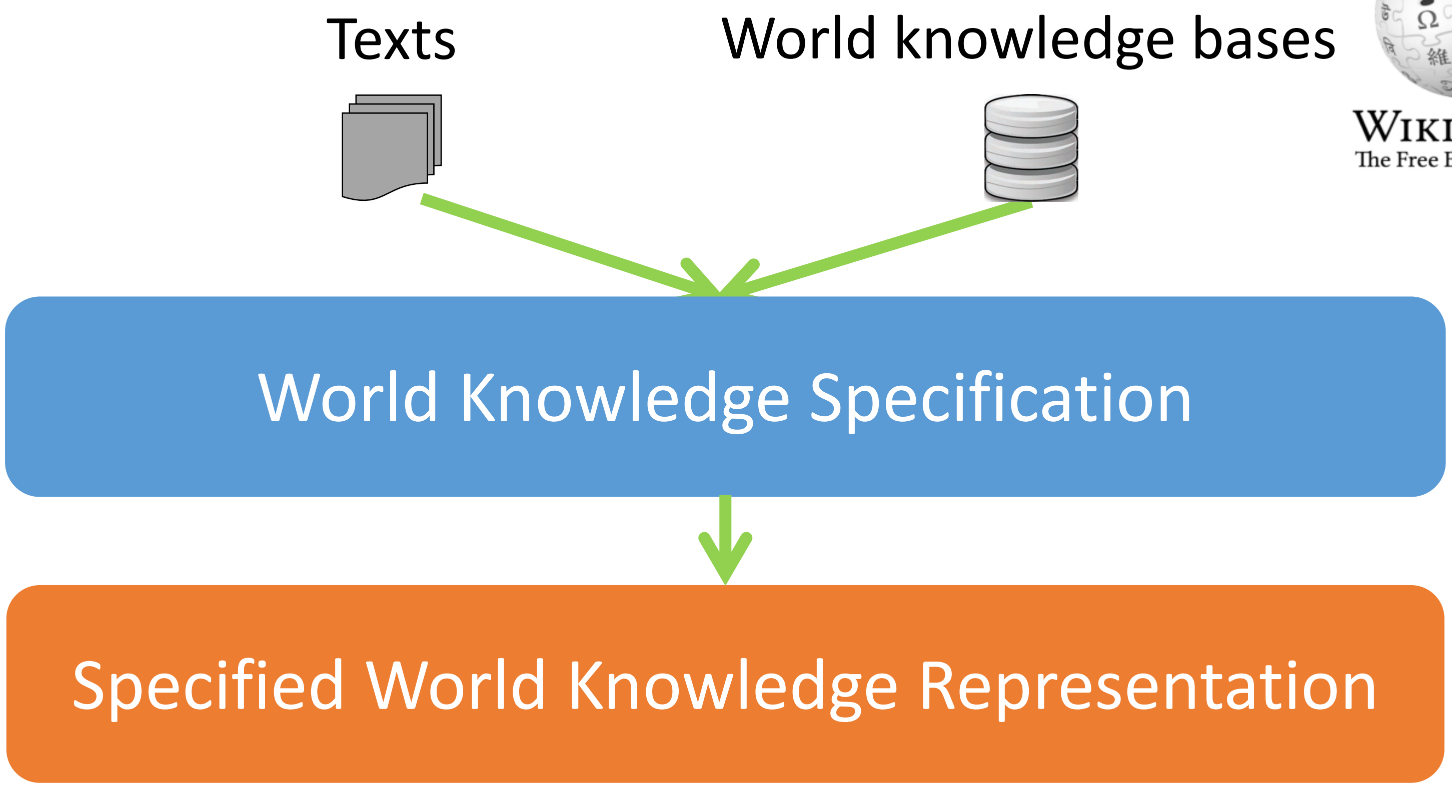
Compute Text Similarity Using HIN

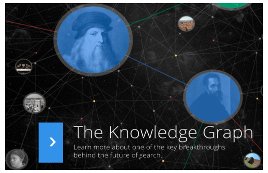


- Heterogeneous information network contains multi-typed nodes and edges.
- Links and types carry rich semantics!
- But traditional approaches are not using them.

Text Based Heterogeneous Information Network Construction

- Grounding texts to world knowledge framework [Wang et al. KDD’15, TKDD’16]





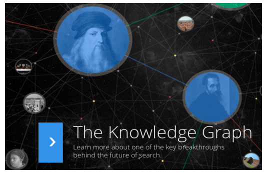
C. Wang et al. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15

Text Based HIN Construction: Unsupervised Semantic Parsing for Documents

C. Wang et al., KDD'15, TKDD'16

Document Trump is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

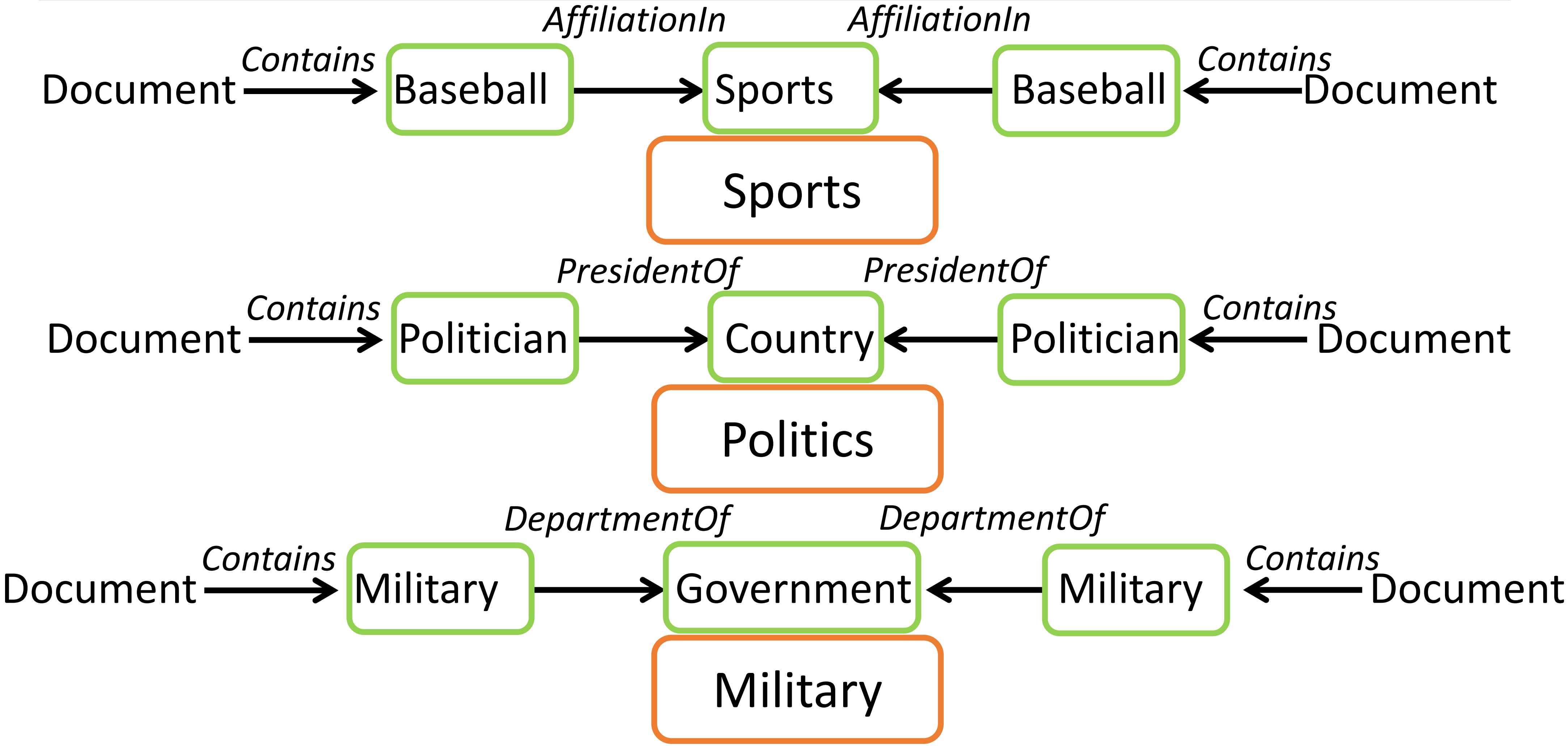


Logic form *People.DonaldTrump* $\sqcap$ PresidentofCountry.*Country.USA*

Advantage: Semantics in the text are preserved in the HIN!

Meta-Path

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



KnowSim: A Meta-Path Based Text Similarity Measure

C. Wang et al., ICDM'15

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

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

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

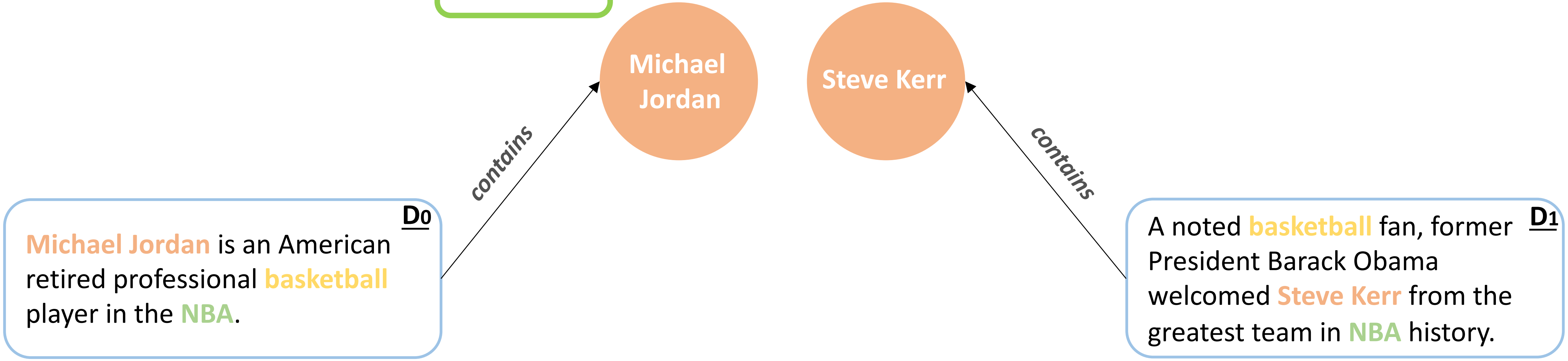
Semantic broadness: the number of total meta-paths between themselves.

- 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.

C. Wang et al. KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM'15

Compute Document Similarity Using HIN

Given: Document $\xrightarrow{\textit{Contains}}$ Athlete $\xleftarrow{\textit{Contains}}$ Document

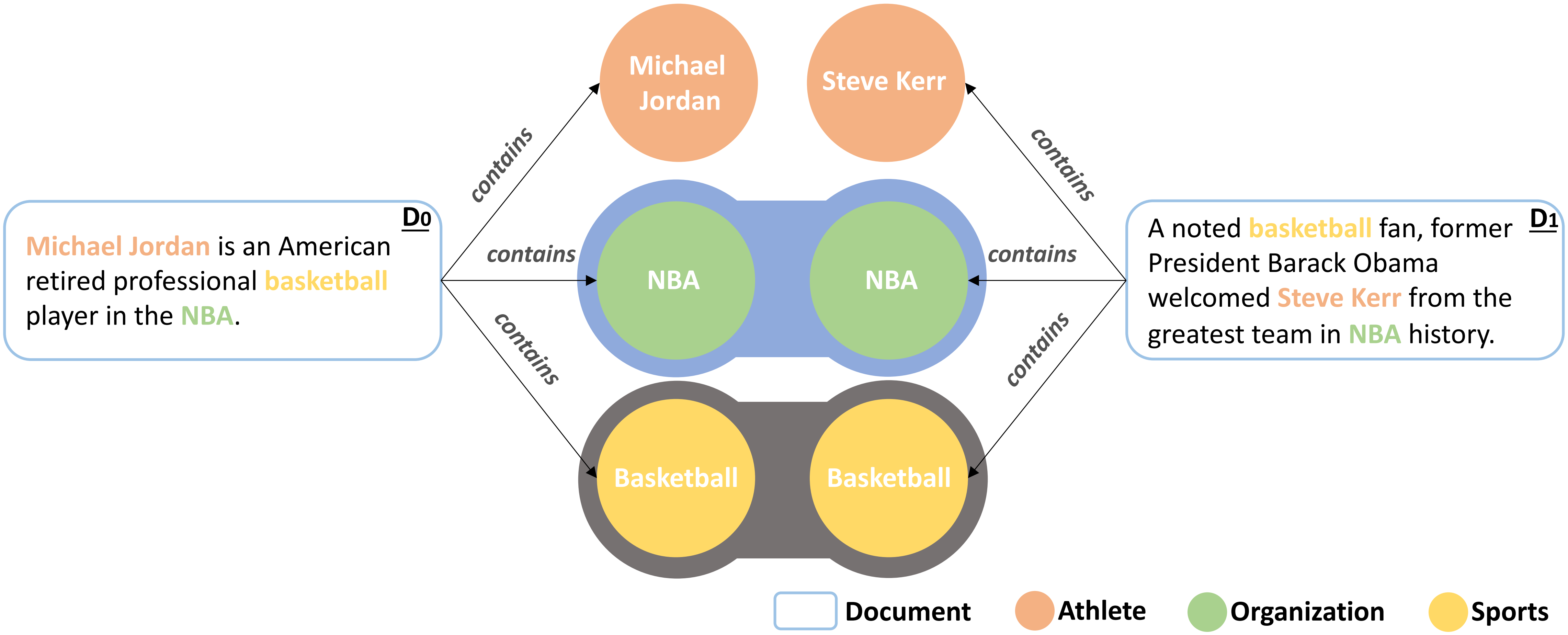


Meta-path similarity assumption: the more instances of meta-path(s) between entities, the more similar the entities are.

No existing meta-path instance → Zero similarity score

Issue: ONLY capturing partial semantics!

Our Approach: Compute Text Similarity Using Distant Similarity Measure on HIN



<u>Distant meta-path similarity assumption:</u> the more similar or the same neighborhood entities of the two entities are, the more similar the two entities should be.

Advantages of Distant Meta-Path Similarity

No existing meta-path instance → Non-Zero similarity score

Capturing full text semantics!

Exploring Similarity Hypothesis Space

Mathematical Assumptions

- set assumption,
- probabilistic assumption

↓

__9__ Families

↓

__53__ Distant meta-path similarity measures

Distant Meta-Path Similarity – **Intersection**

- Hamming Distant Similarity

Number of total meta-paths of entity i

Number of total meta-paths of entity j

$$S_{Ham} = \frac{\boxed{M}\boxed{N}}{\sum_{m=1}^{M}\sum_{k=1}^{N}\left|\boxed{\mathbf{M}_{\mathcal{P}_m}(i,k)} - \boxed{\mathbf{M}_{\mathcal{P}_m}(j,k)}\right|}$$

Meta-path instances between entity i and its neighborhood Entities.

Meta-path instances between entity j and its neighborhood Entities.

Definition: The distance is defined as the number of entities with different meta-paths corresponding to the two entities i and j. Larger intersection of entities means more similar.

Distant Meta-Path Similarity – **Inner Product**

- Cosine Distant Similarity

$$S_{Cos} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N} \mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)}{\sqrt{\sum_{m=1}^{M} \sum_{k=1}^{N} \mathbf{M}_{\mathcal{P}_m}(i,k)^2}\sqrt{\sum_{m=1}^{M} \sum_{k=1}^{N} \mathbf{M}_{\mathcal{P}_m}(j,k)^2}}$$

Definition: Inner product is similar to intersection but also considers the weights of each meta-path value.

Distant Meta-Path Similarity – **Lp Minkowski**

- Euclidean L2 Distant Similarity

$$S_{Euc} = \frac{1}{\sqrt{\sum_{m=1}^{M} \sum_{k=1}^{N} |\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)|^2}}$$

Definition: This distance is similar to Hamming distance, but treats the values of each meta-path independently.

Distant Meta-Path Similarity – **L1**

- Sorensen Distant Similarity

$$S_{S\varnothing r} = 1 - \frac{\sum_{m=1}^{M}\sum_{k=1}^{N}|\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)|}{\sum_{m=1}^{M}\sum_{k=1}^{N}(\mathbf{M}_{\mathcal{P}_m}(i,k) + \mathbf{M}_{\mathcal{P}_m}(j,k))}$$

Definition: Using the sum of all the related meta-path values as denominator to normalize the L1 distance in the range of [0, 1] and regards “1 - the distance” as the similarity.

Distant Meta-Path Similarity – **Squared L2**

- Clark Distant Similarity

$$S_{Cla} = \frac{1}{\sqrt{\sum_{m=1}^{M} \sum_{k=1}^{N} \left(\frac{|\mathrm{M}_{\mathcal{P}_m}(i,k) - \mathrm{M}_{\mathcal{P}_m}(j,k)|}{\mathrm{M}_{\mathcal{P}_m}(i,k) + \mathrm{M}_{\mathcal{P}_m}(j,k)}\right)^2}}$$

Definition: The way to normalize the squared L2 norm is similar to the way Sorensen distance normalizes the L1 distance except for the squared value and the way to sum all the values.

Distant Meta-Path Similarity – **Binary**

- Russell-Rao Distant Similarity

$$S_{Rus} = 1 - \frac{MN - \sum_{m=1}^{M} \sum_{k=1}^{N} \mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)}{MN}$$

Definition: Binary similarity is more complicated than intersection since it can introduce a lot of logical operators over the binary values.

Distant Meta-Path Similarity – **Fidelity**

- Hellinger Distant Similarity

$$S_{Hel} = 2 \times (1 - \sqrt{1 - \sum_{m=1}^{M} \sum_{k=1}^{N} \sqrt{\mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)})}$$

Definition: Hellinger distance is originally defined with measure theory based on two probability distributions.

Distant Meta-Path Similarity – **Shannon's Entropy**

- Kullback-Leibler Distant Similarity

$$S_{KL} = \frac{1}{\sum_{m=1}^{M} \sum_{k=1}^{N} \mathrm{M}_{\mathcal{P}_m}(i,k) \ln \frac{\mathrm{M}_{\mathcal{P}_m}(i,k)}{\mathrm{M}_{\mathcal{P}_m}(j,k)}}$$

Definition: Since the entropy is also defined on probabilities, we normalize the frequencies to be probabilities as we did for Hellinger Distance. KL divergence is originally used to evaluate the difference between two distributions. We regard the inverse value as the similarity.

Distant Meta-Path Similarity – **Hybrids**

- Avg(L1, L∞) Distant Similarity

$$S_{Avg} = \frac{2}{\sum_{m=1}^{M}\sum_{k=1}^{N}|\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)| + \max_{j,k}|\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)|}$$

Definition: We include some combinations of the above similarities.
average of city block and Chebyshev distances.

Experiments

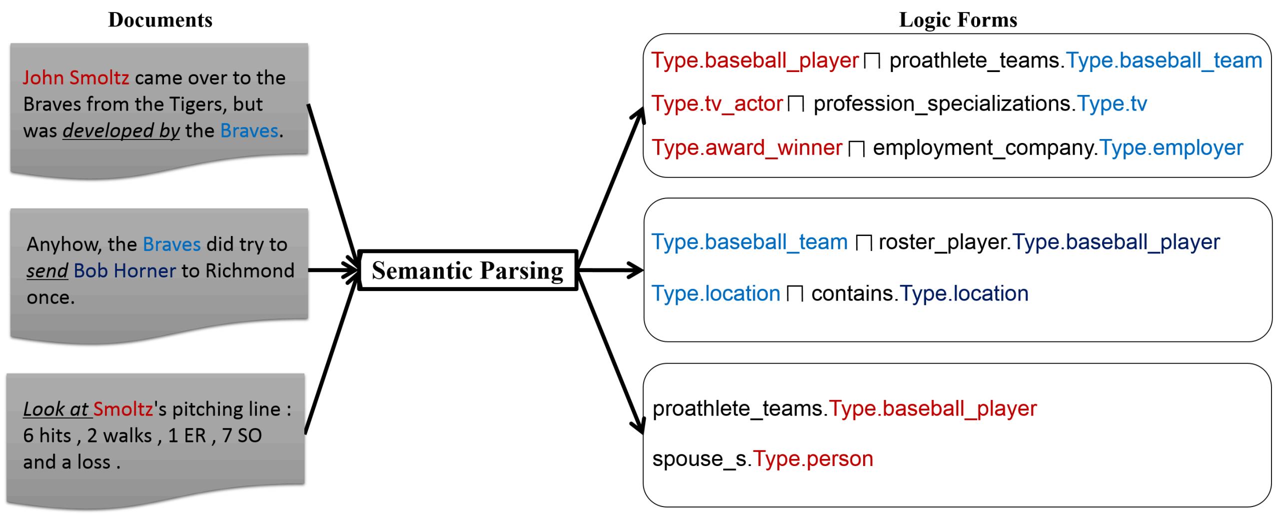
Mathematical Assumptions
- set assumption,
- probabilistic assumption

9 Families

53 Distant meta-path similarity measures

Which distant meta-path similarity is the best?

Experiments



- Four sub-datasets are constructed

20NewsGroup

RCV1-GCAT

Document datasets

Sub-datasets	#(Document)	#(word)	#(Entity)	#(Total)	#(Types)
20NG-SIM	3000	22686	5549	31235	1514
20NG-DIF	3000	25910	6344	35254	1601
GCAG-SIM	3596	22577	8118	34227	1678
GCAT-DIF	2700	33345	12707	48752	1523

Each sub-datasets consists of three similar or distinct topics.

More entities in GCAT

Evaluation Tasks

- Clustering:
  - Spectral Clustering Using Similarities

- Classification:
  - SVM Classification Using Similarities

Results

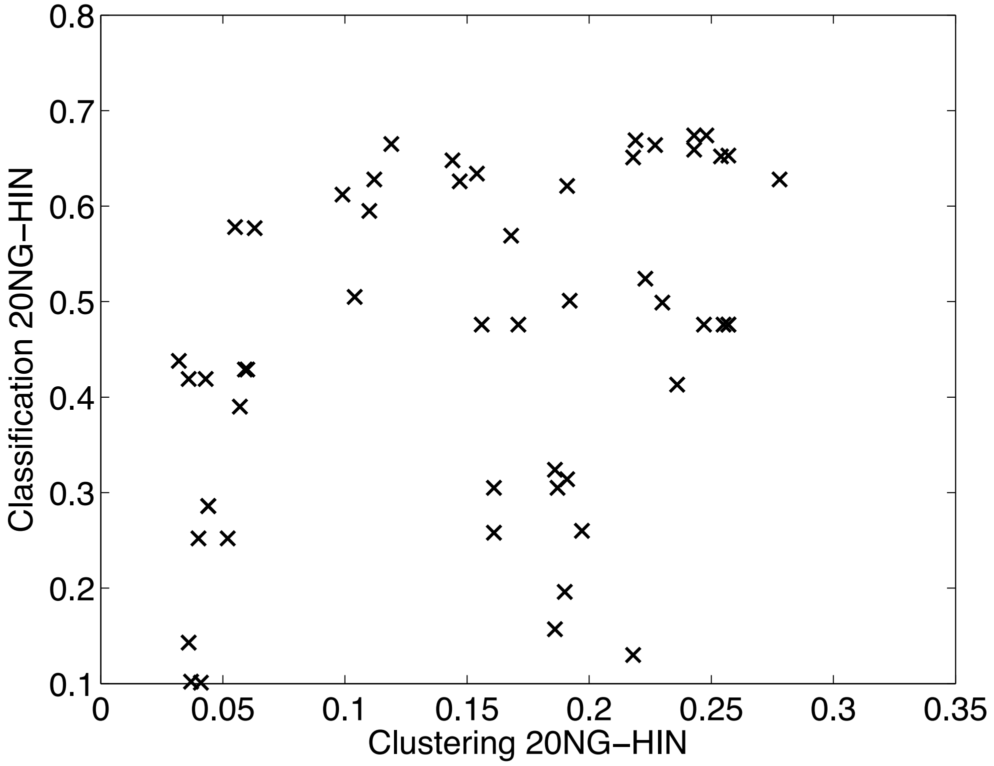
	20NG-HIN		GCAT-HIN	
	Clust.	Class.	Clust.	Class.
KnowSim	0.223	52.4%	0.299	81.6%
Avg(PathSim)	0.218	13.0%	0.329	69.4%
Mean(neighborhood)	**0.221**	32.7%	0.314	75.5%
1. Intersection	0.218	65.1%	0.328	92.1%
2. Wave Hedges	0.057	39.0%	0.159	57.2%
3. Czekanowski	0.119	66.5%	0.229	90.7%
4. Motyka	0.219	66.9%	0.286	85.1%
5. Ruzicka	0.059	42.9%	0.043	46.9%
6. Tanimoto	0.044	28.6%	0.038	41.2%
7. Hamming	0.168	56.9%	0.188	83.6%
Mean(Intersection)	0.126	52.3%	0.182	71.0%
8. Inner Product	0.154	63.4%	0.237	92.7%
9. Harmonic Mean	0.191	62.1%	0.205	89.2%
10. Cosine	0.248	**67.4%**	0.242	**93.1%**
11. Kumar-Hassebrook	0.104	50.5%	0.233	82.4%
12. Jaccard	0.104	50.5%	0.225	82.4%
13. Dice	0.04	25.2%	0.037	56.5%
14. Correlation	0.243	**67.4%**	0.251	**93.1%**
Mean(Inner product)	0.155	55.2%	0.204	84.2%

Neighborhood meta-path similarity

Inner product family distant meta-path similarity

Findings:
#1: Distant similarities are generally better than existing similarities
#2: Cosine distant similarity is consistently good for general use

Correlation Between Clustering and Classification Results



Finding:

Pearson correlation coefficient between clustering and classification results are high, and <u>significant at 0.01 level.</u>

The best distant similarity can be trusted.

Takeaways

Text Representation	Represent text as HIN with rich semantics
Distant Similarity	Compute text similarity in HIN with full text semantics
Data	Please download: https://github.com/cgraywang/TextHINData

Thank You! ☺