# Interpretable and Low-Resource Entity Matching via Decoupling Feature Learning from Decision Making

Zijun Yao, Chengjiang Li, Tiansi Dong, Xin Lv, Jifan Yu
Lei Hou, Juanzi Li, Yichi Zhang, Zelin Dai

*Tsinghua University*
*University of Bonn*
*Alibaba Group*

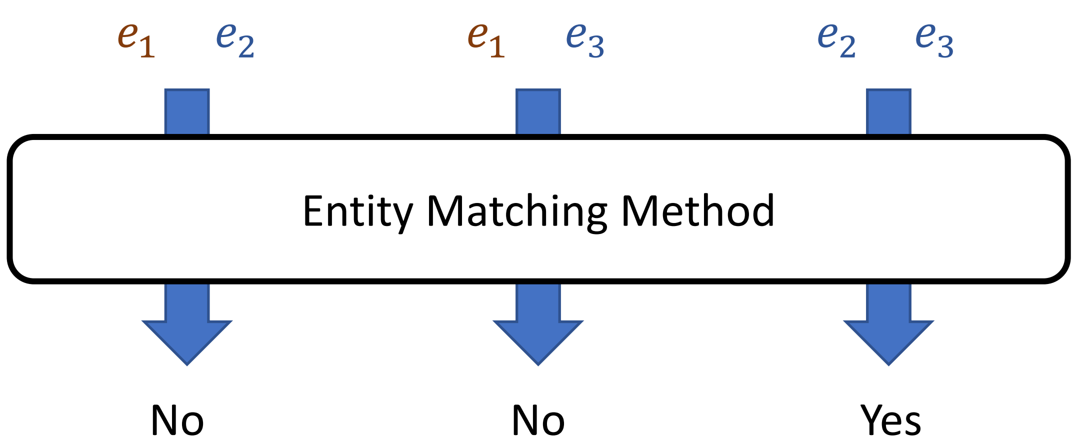
# Index

# Entity Matching

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |

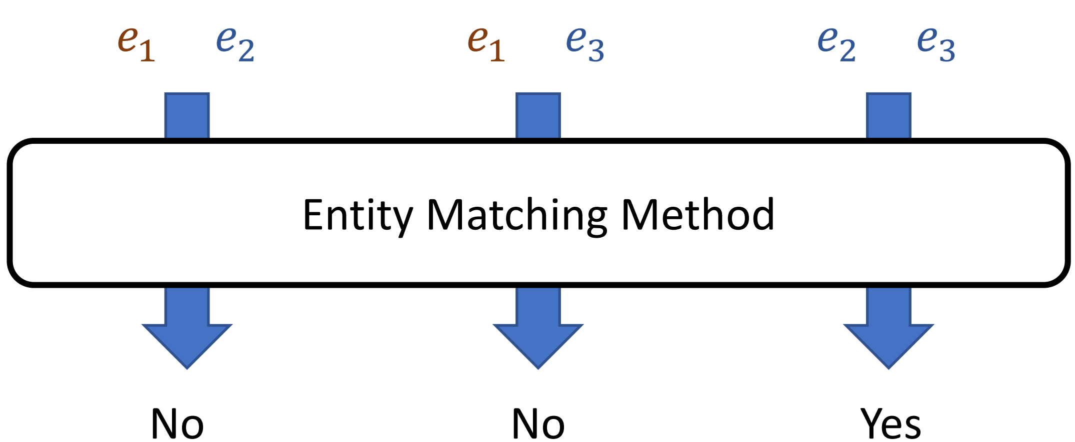
# Entity Matching

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |



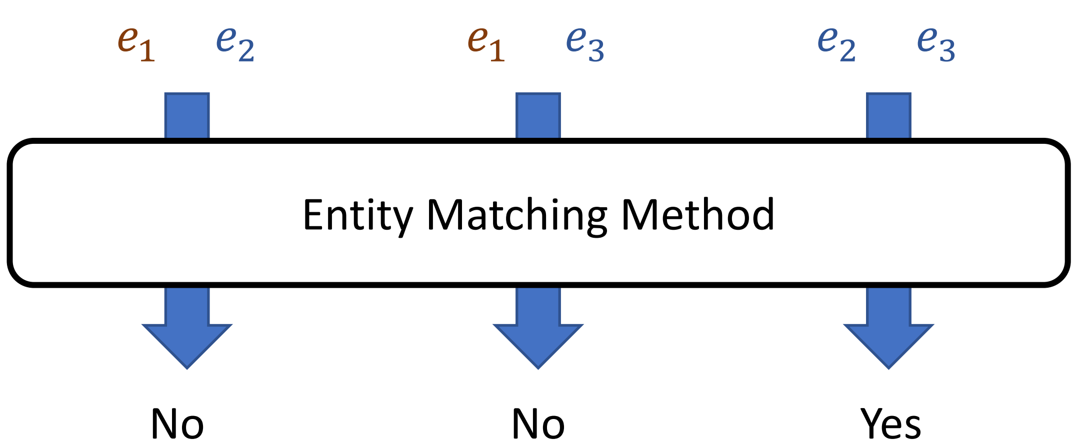
# Entity Matching

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |



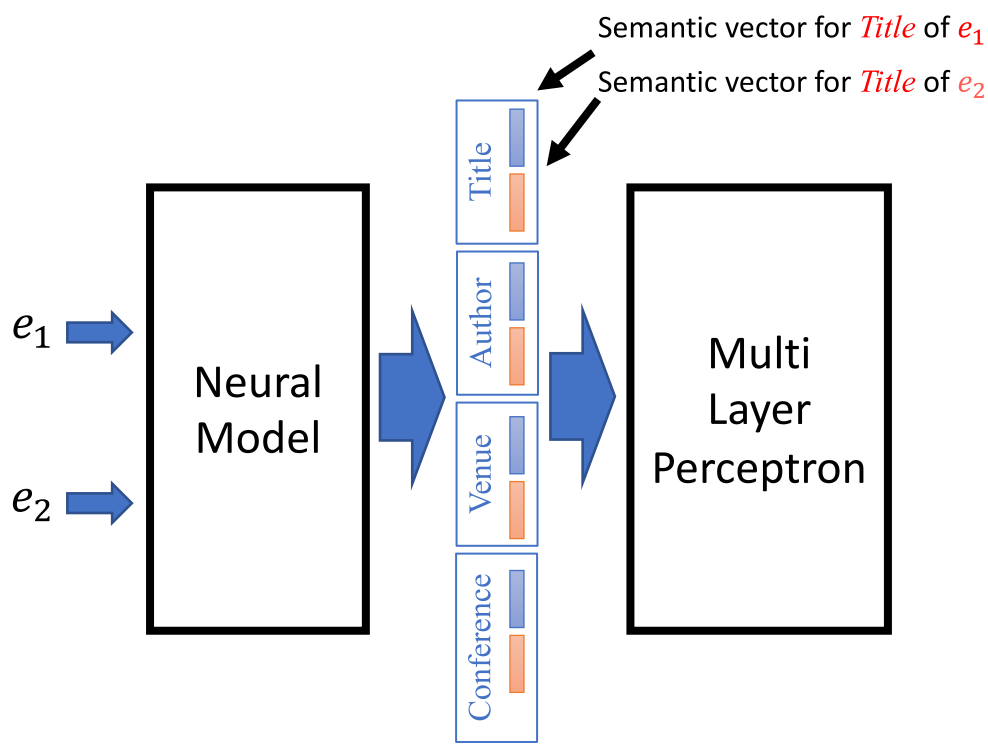
# Entity Matching

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |



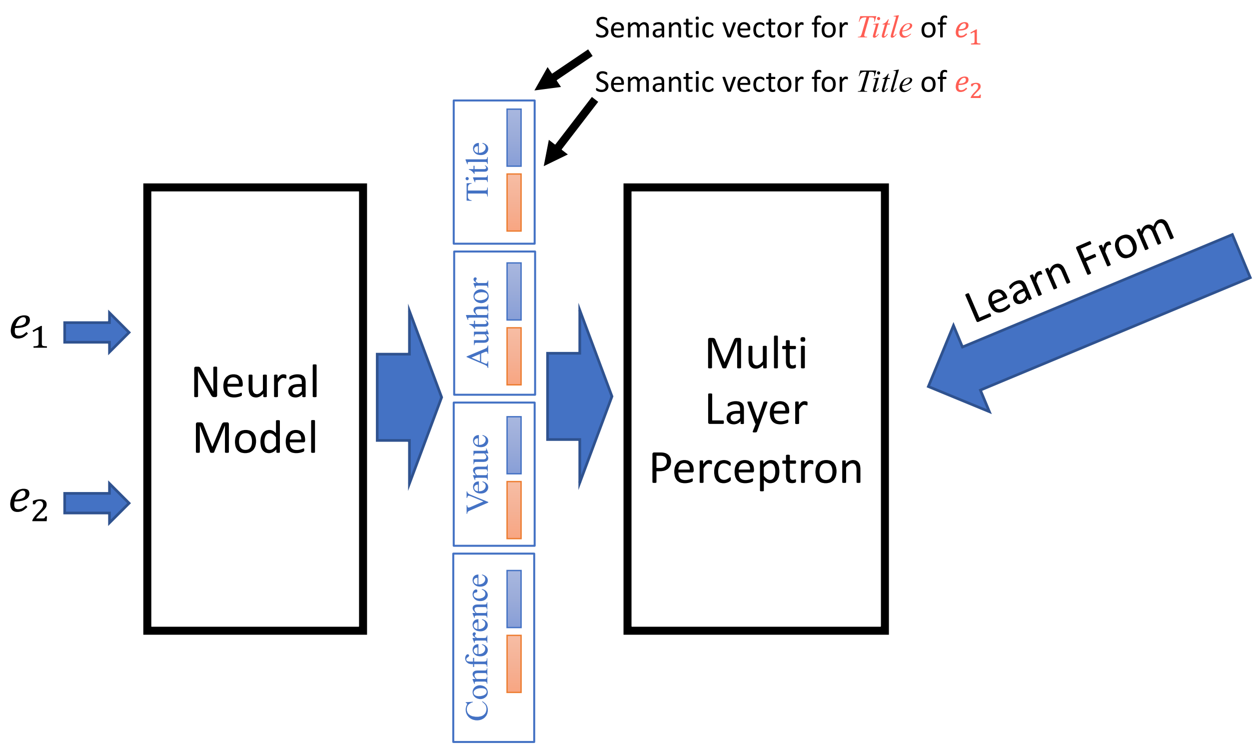
# Entity Matching

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |



# Challenges

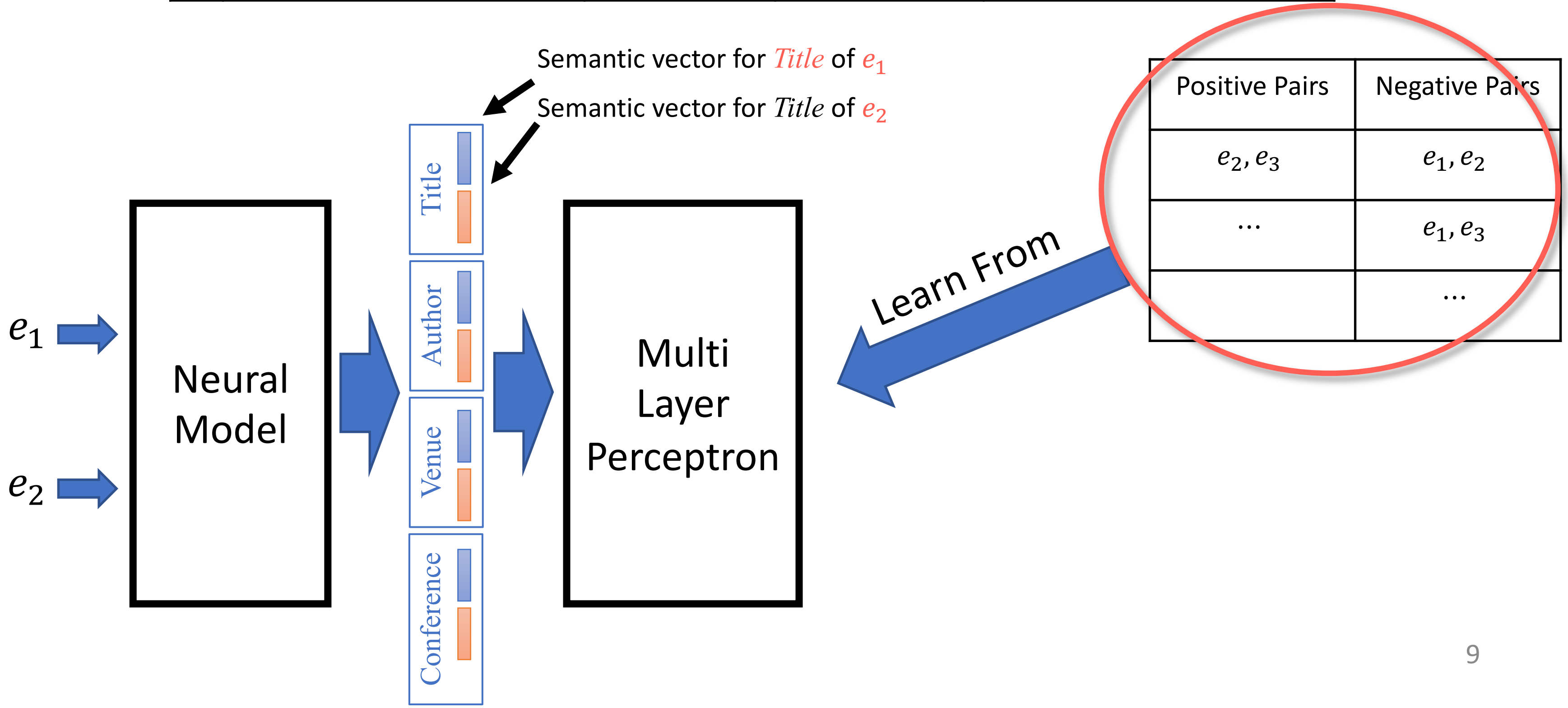
| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |




| Positive Pairs | Negative Pairs |
|---|---|
| $e_2, e_3$ | $e_1, e_2$ |
| … | $e_1, e_3$ |
| | … |

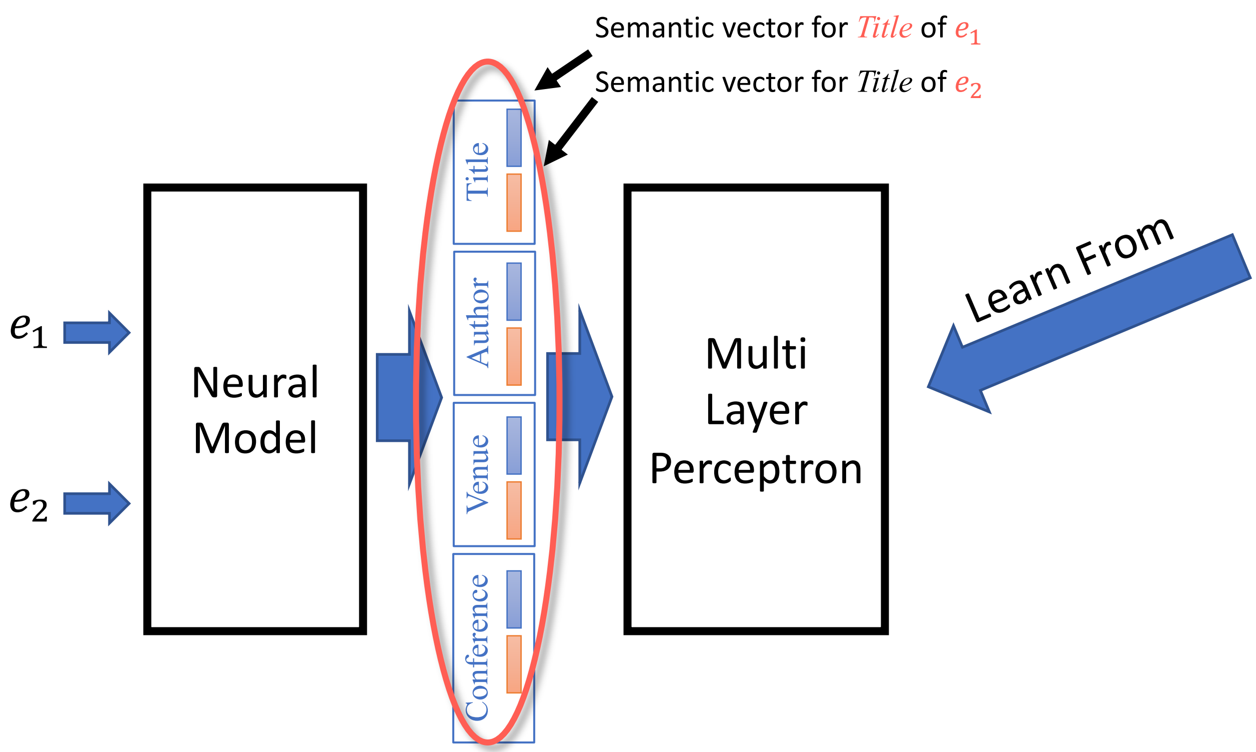
# Challenges

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |



# Challenges

| | Title | Author | Venue | Conference (redundant) |
|---|---|---|---|---|
| $e_1$ | Data Mining Techniques | missing | SIGMOD Conference | International Conference on Management of Data |
| $e_2$ | Data Mining: Concepts and Techniques | J. Han, J. Pei, M. Kamber | SIGMOD Record | missing |
| $e_3$ | Data mining: Concepts & Techniques by Jiawei Han | misplaced | ACM SIGMOD Record | missing |




| Positive Pairs | Negative Pairs |
|---|---|
| $e_2, e_3$ | $e_1, e_2$ |
| … | $e_1, e_3$ |
| | … |

# Motivations

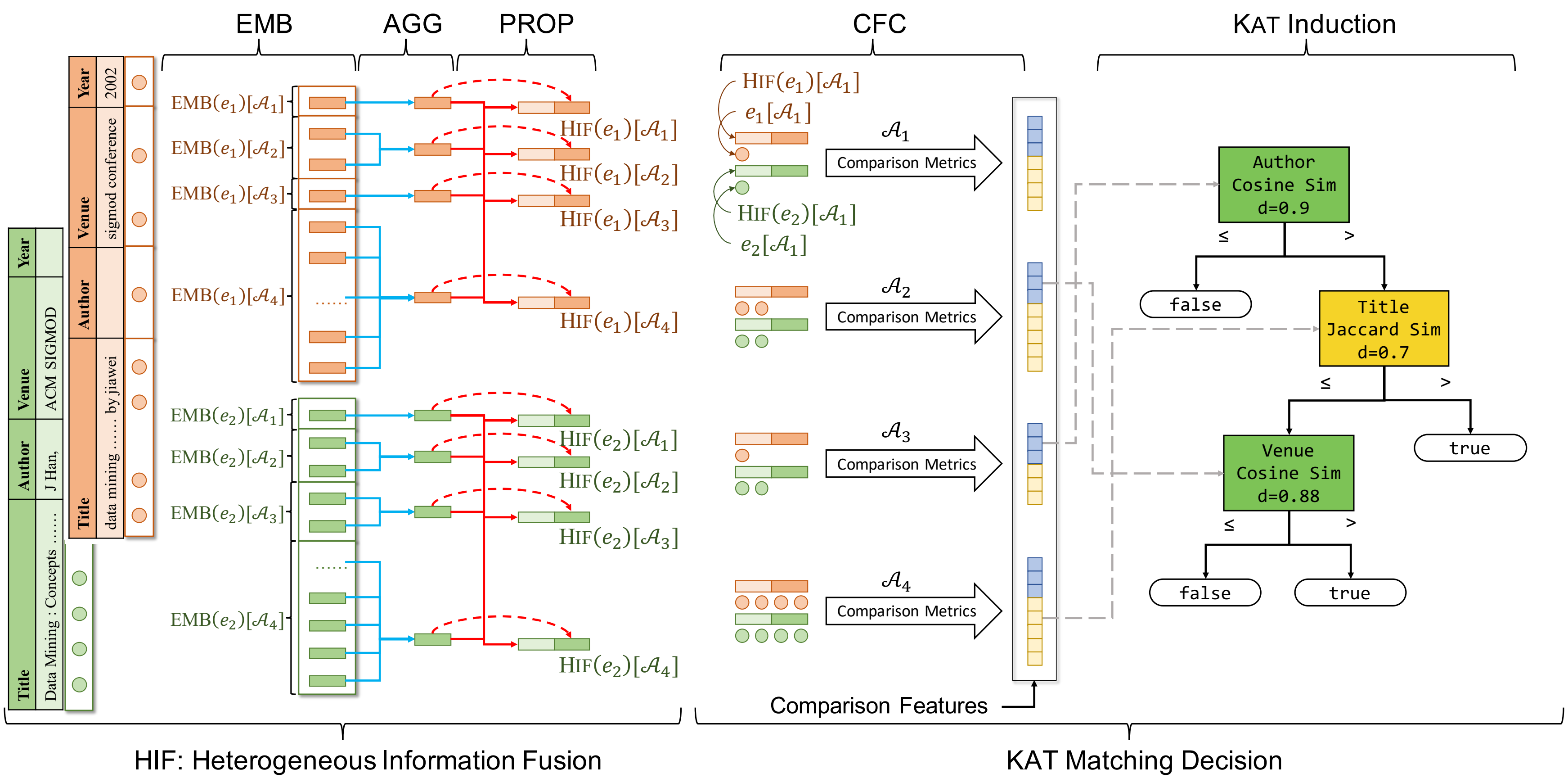
- Representation Learning
  - Make use of unlabeled data
  - Understand heterogeneous attribute values
- Decision Making
  - Using interpretable classifier

# Motivations

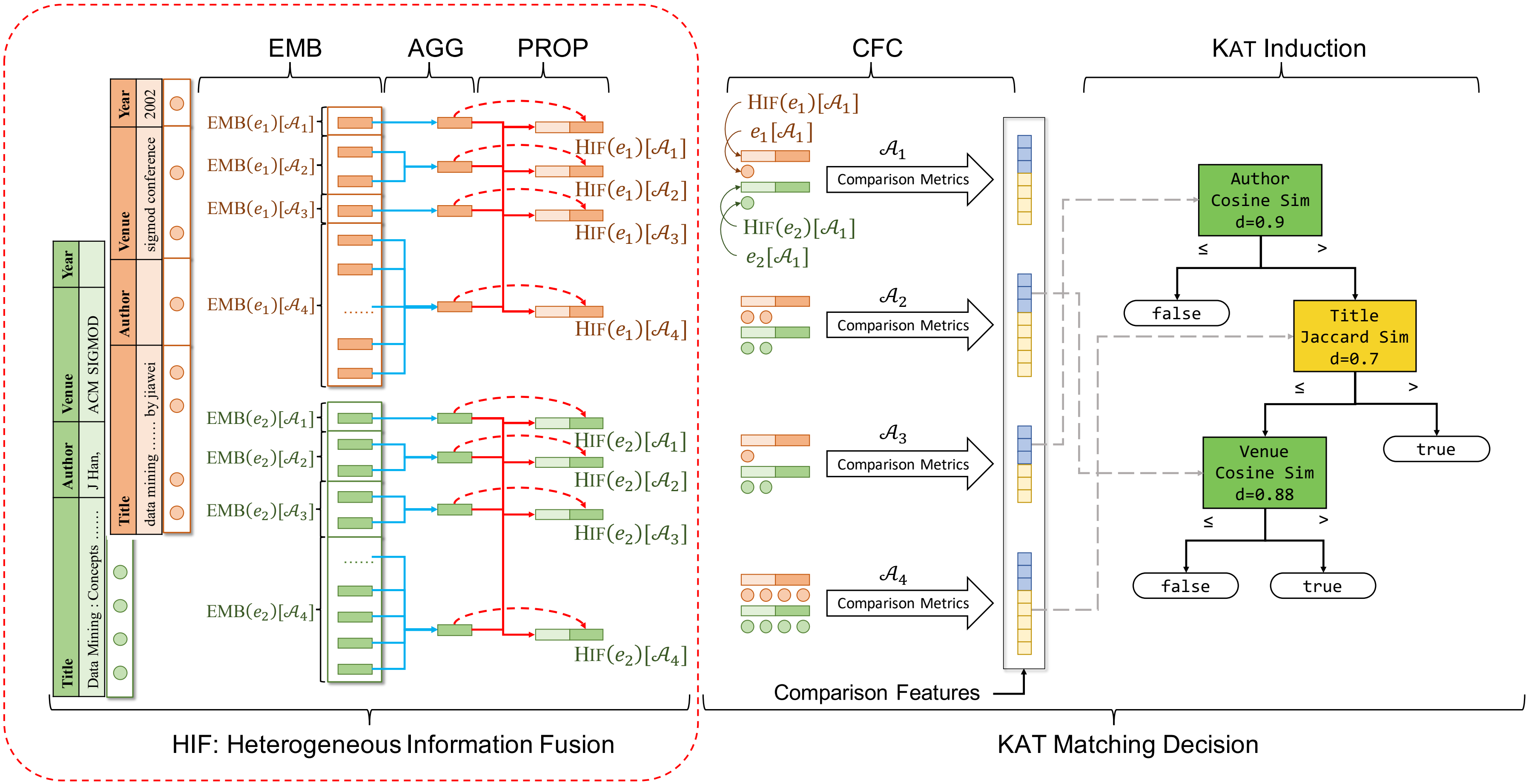
- Representation Learning
  - Make use of unlabeled data
  - Understand heterogeneous attribute values
- Decision Making
  - Using interpretable classifier

# Index

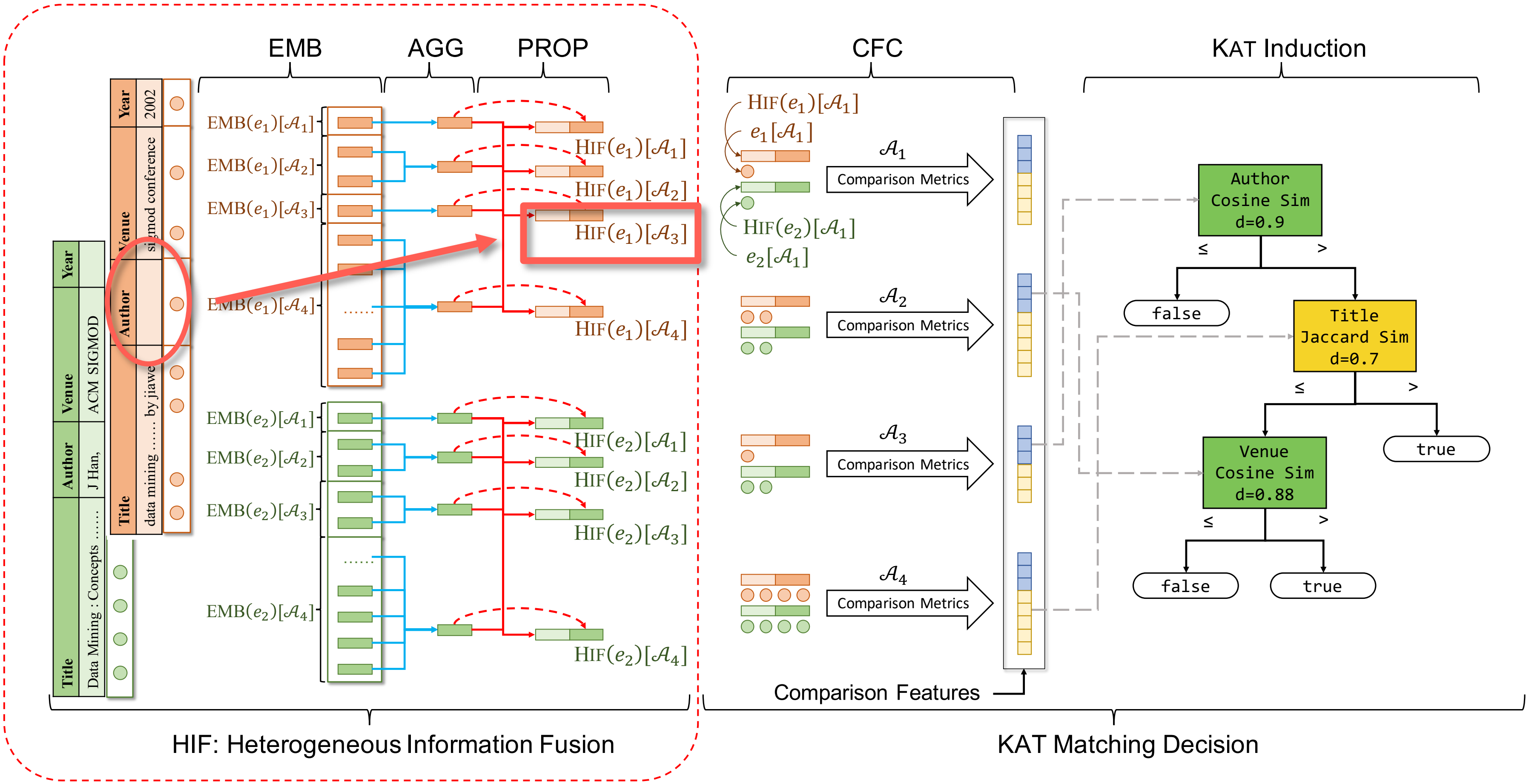
# Overview


EMB
AGG
PROP
CFC
KAT Induction
Year
Venue
Author
Title
2002
sigmod conference
data mining ...... by jiawei
ACM SIGMOD
J Han,
Data Mining : Concepts ......
EMB($e_1$)[$\mathcal{A}_1$]
EMB($e_1$)[$\mathcal{A}_2$]
EMB($e_1$)[$\mathcal{A}_3$]
EMB($e_1$)[$\mathcal{A}_4$]
EMB($e_2$)[$\mathcal{A}_1$]
EMB($e_2$)[$\mathcal{A}_2$]
EMB($e_2$)[$\mathcal{A}_3$]
EMB($e_2$)[$\mathcal{A}_4$]
HIF($e_1$)[$\mathcal{A}_1$]
HIF($e_1$)[$\mathcal{A}_2$]
HIF($e_1$)[$\mathcal{A}_3$]
HIF($e_1$)[$\mathcal{A}_4$]
HIF($e_2$)[$\mathcal{A}_1$]
HIF($e_2$)[$\mathcal{A}_2$]
HIF($e_2$)[$\mathcal{A}_3$]
HIF($e_2$)[$\mathcal{A}_4$]
$e_1$[$\mathcal{A}_1$]
$e_2$[$\mathcal{A}_1$]
$\mathcal{A}_1$
$\mathcal{A}_2$
$\mathcal{A}_3$
$\mathcal{A}_4$
Comparison Metrics
Comparison Features
Author Cosine Sim d=0.9
Title Jaccard Sim d=0.7
Venue Cosine Sim d=0.88
≤
>
false
true
HIF: Heterogeneous Information Fusion
KAT Matching Decision

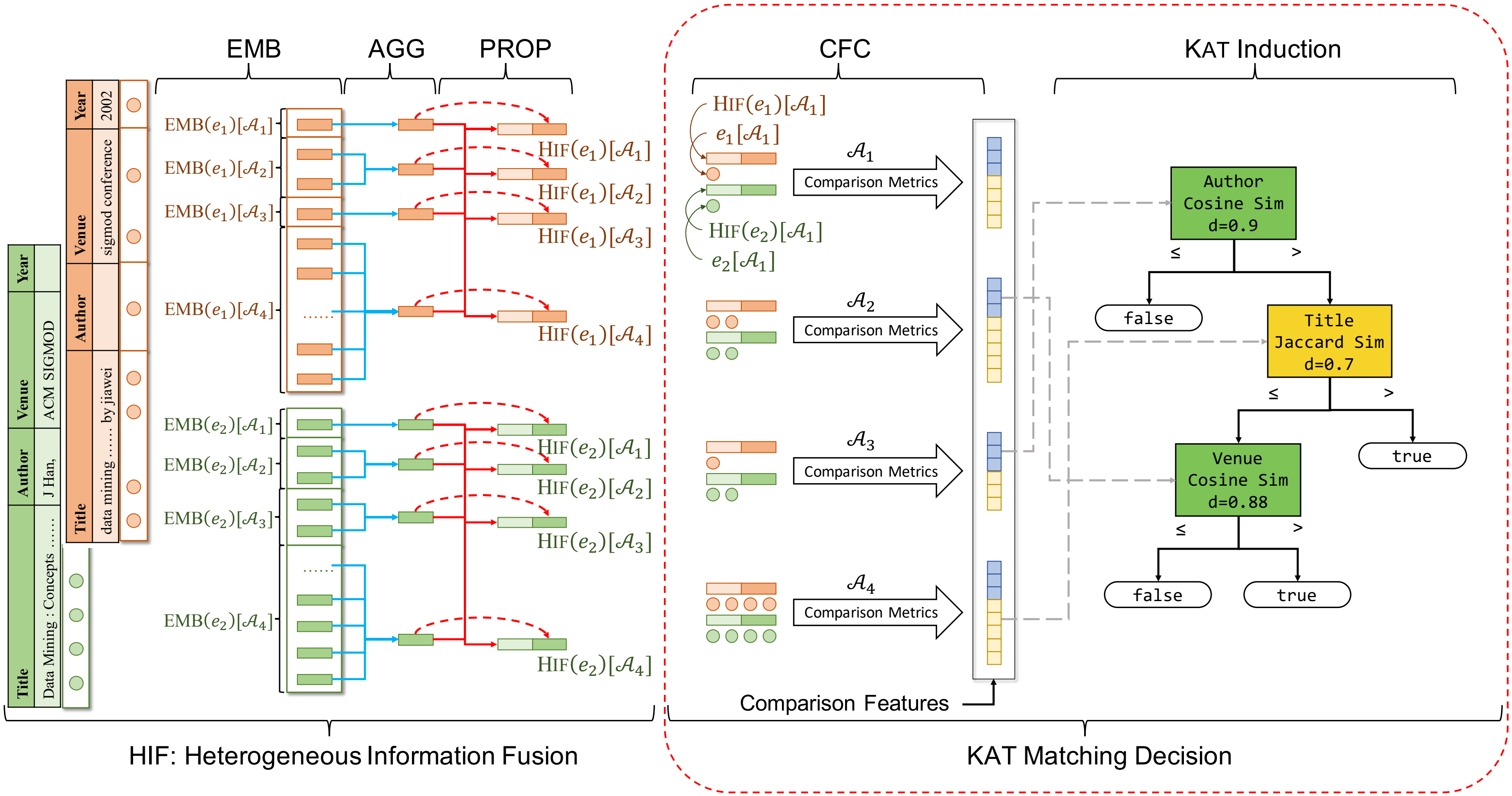
# Overview


EMB
AGG
PROP
CFC
KAT Induction
Year
Venue
Author
Title
2002
sigmod conference
by jiawei
data mining ......
ACM SIGMOD
J Han,
Data Mining : Concepts ......
EMB($e_1$)[$\mathcal{A}_1$]
EMB($e_1$)[$\mathcal{A}_2$]
EMB($e_1$)[$\mathcal{A}_3$]
EMB($e_1$)[$\mathcal{A}_4$]
EMB($e_2$)[$\mathcal{A}_1$]
EMB($e_2$)[$\mathcal{A}_2$]
EMB($e_2$)[$\mathcal{A}_3$]
EMB($e_2$)[$\mathcal{A}_4$]
HIF($e_1$)[$\mathcal{A}_1$]
HIF($e_1$)[$\mathcal{A}_2$]
HIF($e_1$)[$\mathcal{A}_3$]
HIF($e_1$)[$\mathcal{A}_4$]
HIF($e_2$)[$\mathcal{A}_1$]
HIF($e_2$)[$\mathcal{A}_2$]
HIF($e_2$)[$\mathcal{A}_3$]
HIF($e_2$)[$\mathcal{A}_4$]
$e_1$[$\mathcal{A}_1$]
$e_2$[$\mathcal{A}_1$]
$\mathcal{A}_1$
$\mathcal{A}_2$
$\mathcal{A}_3$
$\mathcal{A}_4$
Comparison Metrics
Comparison Features
Author Cosine Sim d=0.9
Title Jaccard Sim d=0.7
Venue Cosine Sim d=0.88
≤
>
false
true
HIF: Heterogeneous Information Fusion
KAT Matching Decision

# Overview


EMB
AGG
PROP
CFC
KAT Induction
EMB(e1)[A1]
EMB(e1)[A2]
EMB(e1)[A3]
EMB(e1)[A4]
EMB(e2)[A1]
EMB(e2)[A2]
EMB(e2)[A3]
EMB(e2)[A4]
HIF(e1)[A1]
HIF(e1)[A2]
HIF(e1)[A3]
HIF(e1)[A4]
HIF(e2)[A1]
HIF(e2)[A2]
HIF(e2)[A3]
HIF(e2)[A4]
Year
Venue
Author
Title
2002
sigmod conference
by jiawe
data mining ……
ACM SIGMOD
J Han,
Data Mining : Concepts ……
e1[A1]
e2[A1]
A1
A2
A3
A4
Comparison Metrics
Comparison Features
Author Cosine Sim d=0.9
≤
>
false
Title Jaccard Sim d=0.7
true
Venue Cosine Sim d=0.88
false
true
HIF: Heterogeneous Information Fusion
KAT Matching Decision

# Overview


EMB
AGG
PROP
CFC
KAT Induction
Year
Venue
Author
Title
2002
sigmod conference
data mining …… by jiawei
ACM SIGMOD
J Han,
Data Mining : Concepts ……
EMB($e_1$)[$\mathcal{A}_1$]
EMB($e_1$)[$\mathcal{A}_2$]
EMB($e_1$)[$\mathcal{A}_3$]
EMB($e_1$)[$\mathcal{A}_4$]
EMB($e_2$)[$\mathcal{A}_1$]
EMB($e_2$)[$\mathcal{A}_2$]
EMB($e_2$)[$\mathcal{A}_3$]
EMB($e_2$)[$\mathcal{A}_4$]
HIF($e_1$)[$\mathcal{A}_1$]
HIF($e_1$)[$\mathcal{A}_2$]
HIF($e_1$)[$\mathcal{A}_3$]
HIF($e_1$)[$\mathcal{A}_4$]
HIF($e_2$)[$\mathcal{A}_1$]
HIF($e_2$)[$\mathcal{A}_2$]
HIF($e_2$)[$\mathcal{A}_3$]
HIF($e_2$)[$\mathcal{A}_4$]
$e_1$[$\mathcal{A}_1$]
$e_2$[$\mathcal{A}_1$]
$\mathcal{A}_1$
$\mathcal{A}_2$
$\mathcal{A}_3$
$\mathcal{A}_4$
Comparison Metrics
Comparison Features
Author Cosine Sim d=0.9
Title Jaccard Sim d=0.7
Venue Cosine Sim d=0.88
≤
>
false
true
HIF: Heterogeneous Information Fusion
KAT Matching Decision

# EMB – Word Embedding

- Segmentation and padding for each attribute value.

$$\underbrace{(\langle \text{BEG} \rangle, w_1, w_2, \cdots \langle \text{END} \rangle, \langle \text{PAD} \rangle, \cdots, \langle \text{PAD} \rangle)}_{\text{length} = l}$$

- Static embedding as look up table operation.
  - Attribute value $e[\mathcal{A}_i] \Longrightarrow l$ embedding vectors.
  - $\text{EMB}(e)[\mathcal{A}_i] \in \mathbb{R}^{l \times d_2}$

# AGG – Word Information Aggregation

- Attribute value $e[\mathcal{A}_i] \Longrightarrow l$ embedding vectors $\Longrightarrow$ 1 embedding vector

# AGG – Word Information Aggregation

- Attribute value $e[\mathcal{A}_i] \Longrightarrow l$ embedding vectors $\Longrightarrow$ 1 embedding vector
- Aggregation weight is learned from attention
  - Attention vector for the $i^{th}$ attributes: $\alpha_i$

$$\alpha_i = \mathrm{Softmax}(\mathbf{EMB}(e)[\mathcal{A}_i]\ a_i)^\top \in \mathbb{R}^{1\times l}$$

# AGG – Word Information Aggregation

- Attribute value $e[\mathcal{A}_i] \Longrightarrow l$ embedding vectors $\Longrightarrow$ 1 embedding vector
- Aggregation weight is learned from attention
  - Attention vector for the $i^{th}$ attributes: $\alpha_i$

$$\alpha_i = \mathrm{Softmax}(\mathbf{EMB}(e)[\mathcal{A}_i]\, a_i)^\top \in \mathbb{R}^{1\times l}$$

# AGG – Word Information Aggregation

- Attribute value $e[\mathcal{A}_i] \Longrightarrow l$ embedding vectors $\Longrightarrow$ 1 embedding vector
- Aggregation weight is learned from attention
  - Attention vector for the $i^{th}$ attributes: $\alpha_i$

$$\alpha_i = \textbf{Softmax}(\textbf{EMB}(e)[\mathcal{A}_i]\ a_i)^\top \in \mathbb{R}^{1\times l}$$

  - Aggregation as weighted sum: $\mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i]) \in \mathbb{R}^{d_a}$

$$\mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i]) = \mathrm{ReLU}\left(\alpha_i\ \mathrm{EMB}(e)[\mathcal{A}_i]\ \mathbf{W}_{ai}\right)$$

# PROP – Attribute Information Propagation

- Recover noisy attribute value by information propagation
  - Learn propagation weight with Scaled-Dot-Product

$$\mathbf{q}_i = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{Q}$$
$$\mathbf{k}_j = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{K}$$
$$\mathbf{v}_i = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{V}_i$$

$$\mathbf{A}_{ij} = \mathrm{Softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{m}}\right)$$

# PROP – Attribute Information Propagation

- Recover noisy attribute value by information propagation
    - Learn propagation weight with Scaled-Dot-Product

$$\mathbf{q}_i = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{Q}$$
$$\mathbf{k}_j = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{K}$$
$$\mathbf{v}_i = \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{V}_i$$

$$\mathbf{A}_{ij} = \mathrm{Softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{m}}\right)$$

# PROP – Attribute Information Propagation

- Recover noisy attribute value by information propagation
  - Learn propagation weight with Scaled-Dot-Product
  - Keep identity information with Residual connection

$$
\begin{aligned}
\mathbf{q}_i &= \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{Q} \\
\mathbf{k}_j &= \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{K} \\
\mathbf{v}_i &= \mathrm{AGG}(\mathrm{EMB}(e)[\mathcal{A}_i])\ \mathbf{V}_i
\end{aligned}
$$

$$
\mathbf{A}_{ij} = \mathrm{Softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{m}}\right)
$$

$$
\mathrm{PROP}(\mathrm{AGG}(e))[\mathcal{A}_i] = \mathrm{ReLU}\left(\mathbf{v}_i \middle\| \sum_{j \neq i} \mathbf{A}_{ij}\mathbf{v}_j\right)
$$

# CFC – Comparison Feature Computation

- Embedded feature comparison
  - Cosine Similarity
  - $L_2$ Distance
  - Pearson Coefficient

# CFC – Comparison Feature Computation

- Embedded feature comparison
  - Cosine Similarity
  - $L_2$ Distance
  - Pearson Coefficient

# CFC – Comparison Feature Computation

- Embedded feature comparison
- Original attribute value comparison

| Attribute Type | Comparison Metrics |
|---|---|
| boolean | Exact matching distance |
| number | Exact matching distance, Absolute distance, Levenshtein distance, Levenshtein similarity |
| string of length 1 | Levenshtein distance, Levenshtein similarity, Jaro similarity, Jaro Winkler similarity, Exact matching distance, Jaccard similarity with QGram tokenizer, |
| string of length $[2, 5]$ | Jaccard similarity with QGram tokenizer, Jaccard similarity with delimiter tokenizer, Levenshtein distance, Levenshtein similarity Cosine similarity with delimiter tokenizer, Monge Elkan similarity, Smith Waterman similarity, |
| string of length $[6, 10]$ | Jaccard similarity with QGram tokenizer, Cosine similarity with delimiter tokenizer, Levenshtein distance, Levenshtein similarity, Monge Elkan similarity |
| string of length $[10, \infty]$ | Jaccard similarity with QGram tokenizer, Cosine similarity with delimiter tokenizer |

# KAT Induction

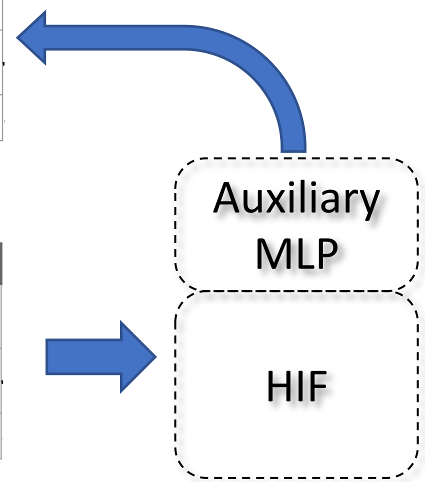
- Key Attribute heuristic
  - Entity records can be determined to be a match with few key attributes
  - Some attributes are more important than others for EM
- Key Attribute Tree
  - Inducted with decision tree algorithm
  - Input: CFC features
  - Output: True for matching and False for non-matching

# Mask Attribute Values

- Auxiliary MLP layers
- Training objective: Cross entropy
    - Auxiliary MLP output
    - Weighted Bag of Word vector

| title | manf./modelno | price |
|---|---|---|
| *instant immersion spanish deluxe 2.0* | topics entertainment | 49.99 |
| *adventure workshop 4th-6th grade 7th edition* | encore software | 19.99 |
| *sharp printing calculator* | sharp el1192bl | 37.63 |

| title | manf./modelno | price |
|---|---|---|
| *instant immersion spanish deluxe 2.0* | topics entertainment | MASK |
| *adventure workshop 4th-6th grade 7th edition* | MASK | 19.99 |
| *sharp printing calculator* | MASK | 37.63 |

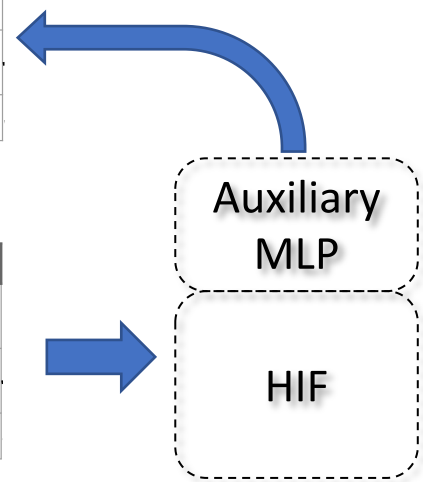


# Mask Attribute Values

- Auxiliary MLP layers
- Training objective: Cross entropy
  - Auxiliary MLP output
  - Weighted Bag of Word vector

| title | manf./modelno | price |
|---|---|---|
| *instant immersion spanish deluxe 2.0* | topics entertainment | 49.99 |
| *adventure workshop 4th-6th grade 7th edition* | encore software | 19.99 |
| *sharp printing calculator* | sharp el1192bl | 37.63 |

| title | manf./modelno | price |
|---|---|---|
| *instant immersion spanish deluxe 2.0* | topics entertainment | MASK |
| *adventure workshop 4th-6th grade 7th edition* | MASK | 19.99 |
| *sharp printing calculator* | MASK | 37.63 |



# Index

1. Background
2. Methodology
   2.1. Heterogeneous Information Fusion
   2.2. Key Attribute Tree
   2.3. Self-Supervised Training
3. **Experiments**
4. Conclusion & Future Work

# Datasets

- Structured: Attribute values are complete
- Dirty: Attribute values are noisy with missing and misplacement
- Real: Industrial dataset from Taobao

| Type | Dataset | #Attr. | #Rec. | #Pos. | #Neg. | Rate |
|---|---|---|---|---|---|---|
| Structured | I-$A_1$ | 8 | 2,908 | 132 | 407 | 10% |
| | D-$A_1$ | 4 | 4,739 | 2,220 | 10,143 | 1% |
| | D-$S_1$ | 4 | 13,270 | 5,347 | 23,360 | 1% |
| Dirty | I-$A_2$ | 8 | 2,908 | 132 | 407 | 10% |
| | D-$A_2$ | 4 | 4,739 | 2,220 | 10,143 | 1% |
| | D-$S_2$ | 4 | 13,270 | 5,347 | 23,360 | 1% |
| Real | Phone | 36 | 940 | 1,099 | 2,241 | 10% |
| | Skirt | 20 | 9,708 | 6,371 | 18,202 | 1% |
| | Toner | 13 | 7,065 | 4,551 | 13,481 | 1% |

# Low Resource Entity Matching

- Evaluation metrics: $F_1$ measure

| Methods | I-A$_1$ | D-A$_1$ | D-S$_1$ | I-A$_2$ | D-A$_2$ | D-S$_2$ | Phone | Skirt | Toner |
|---|---|---|---|---|---|---|---|---|---|
| DM-RNN | 63.6 | 85.4 | 74.8 | 42.3 | 45.7 | 39.0 | 90.0 | 67.6 | 68.6 |
| DM-ATT | 55.8 | 82.5 | 79.0 | 46.5 | 45.2 | 57.8 | 80.3 | 54.4 | 48.8 |
| DM-HYB | 60.9 | 86.6 | 78.0 | 49.5 | 46.2 | 60.4 | 91.9 | 64.2 | 67.4 |
| HierMatcher | 61.9 | 37.5 | 68.2 | 37.8 | 32.6 | 45.8 | 86.2 | 61.7 | 55.2 |
| Magellan | 92.3 | 93.7 | 85.1 | 50.6 | 65.6 | 71.1 | 93.6 | 96.6 | **97.2** |
| HIF+DT | **96.0** | 96.4 | 87.5 | **54.9** | 80.1 | 74.2 | **94.9** | **96.7** | **97.2** |
| HIF+KAT$_{ID3}$ | 95.8 | **96.6** | **88.2** | 51.6 | 79.0 | **79.5** | 94.5 | **96.7** | **97.2** |
| HIF+KAT$_{XGB}$ | 90.6 | 93.3 | 87.9 | 41.5 | **80.3** | **79.5** | 94.4 | 96.2 | **97.2** |
| HIF+LN | 77.9 | 21.0 | 54.7 | 41.6 | - | 78.5 | 72.2 | 62.8 | 86.0 |
| HIF+LR | 84.2 | 87.1 | 84.6 | 46.5 | - | 68.1 | 87.5 | 41.7 | 62.0 |
| HIF-WBOW | 93.0 | 92.7 | 75.4 | 43.2 | 47.9 | 43.7 | 91.6 | 66.3 | 74.0 |
| HIF-EMB | 91.1 | 90.9 | 76.6 | 30.8 | 53.9 | 46.8 | 89.9 | 65.7 | 79.8 |
| HIF-ALONE | 94.6 | 96.1 | 82.9 | 45.6 | 73.5 | 63.2 | 91.8 | 63.0 | 72.9 |

# Low Resource Entity Matching

- Evaluation metrics: $F_1$ measure

| Methods | I-A$_1$ | D-A$_1$ | D-S$_1$ | I-A$_2$ | D-A$_2$ | D-S$_2$ | Phone | Skirt | Toner |
|---|---|---|---|---|---|---|---|---|---|
| DM-RNN | 63.6 | 85.4 | 74.8 | 42.3 | 45.7 | 39.0 | 90.0 | 67.6 | 68.6 |
| DM-ATT | 55.8 | 82.5 | 79.0 | 46.5 | 45.2 | 57.8 | 80.3 | 54.4 | 48.8 |
| DM-HYB | 60.9 | 86.6 | 78.0 | 49.5 | 46.2 | 60.4 | 91.9 | 64.2 | 67.4 |
| HierMatcher | 61.9 | 37.5 | 68.2 | 37.8 | 32.6 | 45.8 | 86.2 | 61.7 | 55.2 |
| Magellan | 92.3 | 93.7 | 85.1 | 50.6 | 65.6 | 71.1 | 93.6 | 96.6 | **97.2** |
| HIF+DT | **96.0** | 96.4 | 87.5 | **54.9** | 80.1 | 74.2 | **94.9** | **96.7** | **97.2** |
| HIF+KAT$_{ID3}$ | 95.8 | **96.6** | **88.2** | 51.6 | 79.0 | **79.5** | 94.5 | **96.7** | **97.2** |
| HIF+KAT$_{XGB}$ | 90.6 | 93.3 | 87.9 | 41.5 | **80.3** | **79.5** | 94.4 | 96.2 | **97.2** |
| HIF+LN | 77.9 | 21.0 | 54.7 | 41.6 | - | 78.5 | 72.2 | 62.8 | 86.0 |
| HIF+LR | 84.2 | 87.1 | 84.6 | 46.5 | - | 68.1 | 87.5 | 41.7 | 62.0 |
| HIF-WBOW | 93.0 | 92.7 | 75.4 | 43.2 | 47.9 | 43.7 | 91.6 | 66.3 | 74.0 |
| HIF-EMB | 91.1 | 90.9 | 76.6 | 30.8 | 53.9 | 46.8 | 89.9 | 65.7 | 79.8 |
| HIF-ALONE | 94.6 | 96.1 | 82.9 | 45.6 | 73.5 | 63.2 | 91.8 | 63.0 | 72.9 |

# Low Resource Entity Matching

- Evaluation metrics: $F_1$ measure

| Methods | I-A$_1$ | D-A$_1$ | D-S$_1$ | I-A$_2$ | D-A$_2$ | D-S$_2$ | Phone | Skirt | Toner |
|---|---|---|---|---|---|---|---|---|---|
| DM-RNN | 63.6 | 85.4 | 74.8 | 42.3 | 45.7 | 39.0 | 90.0 | 67.6 | 68.6 |
| DM-ATT | 55.8 | 82.5 | 79.0 | 46.5 | 45.2 | 57.8 | 80.3 | 54.4 | 48.8 |
| DM-HYB | 60.9 | 86.6 | 78.0 | 49.5 | 46.2 | 60.4 | 91.9 | 64.2 | 67.4 |
| HierMatcher | 61.9 | 37.5 | 68.2 | 37.8 | 32.6 | 45.8 | 86.2 | 61.7 | 55.2 |
| Magellan | 92.3 | 93.7 | 85.1 | 50.6 | 65.6 | 71.1 | 93.6 | 96.6 | **97.2** |
| HIF+DT | **96.0** | 96.4 | 87.5 | **54.9** | 80.1 | 74.2 | **94.9** | **96.7** | **97.2** |
| HIF+KAT$_{ID3}$ | 95.8 | **96.6** | **88.2** | 51.6 | 79.0 | **79.5** | 94.5 | **96.7** | **97.2** |
| HIF+KAT$_{XGB}$ | 90.6 | 93.3 | 87.9 | 41.5 | **80.3** | **79.5** | 94.4 | 96.2 | **97.2** |
| HIF+LN | 77.9 | 21.0 | 54.7 | 41.6 | - | 78.5 | 72.2 | 62.8 | 86.0 |
| HIF+LR | 84.2 | 87.1 | 84.6 | 46.5 | - | 68.1 | 87.5 | 41.7 | 62.0 |
| HIF-WBOW | 93.0 | 92.7 | 75.4 | 43.2 | 47.9 | 43.7 | 91.6 | 66.3 | 74.0 |
| HIF-EMB | 91.1 | 90.9 | 76.6 | 30.8 | 53.9 | 46.8 | 89.9 | 65.7 | 79.8 |
| HIF-ALONE | 94.6 | 96.1 | 82.9 | 45.6 | 73.5 | 63.2 | 91.8 | 63.0 | 72.9 |

# Low Resource Entity Matching

- Evaluation metrics: $F_1$ measure

| Methods | I-$A_1$ | D-$A_1$ | D-$S_1$ | I-$A_2$ | D-$A_2$ | D-$S_2$ | Phone | Skirt | Toner |
|---|---|---|---|---|---|---|---|---|---|
| DM-RNN | 63.6 | 85.4 | 74.8 | 42.3 | 45.7 | 39.0 | 90.0 | 67.6 | 68.6 |
| DM-ATT | 55.8 | 82.5 | 79.0 | 46.5 | 45.2 | 57.8 | 80.3 | 54.4 | 48.8 |
| DM-HYB | 60.9 | 86.6 | 78.0 | 49.5 | 46.2 | 60.4 | 91.9 | 64.2 | 67.4 |
| HierMatcher | 61.9 | 37.5 | 68.2 | 37.8 | 32.6 | 45.8 | 86.2 | 61.7 | 55.2 |
| Magellan | 92.3 | 93.7 | 85.1 | 50.6 | 65.6 | 71.1 | 93.6 | 96.6 | **97.2** |
| HIF+DT | **96.0** | 96.4 | 87.5 | **54.9** | 80.1 | 74.2 | **94.9** | **96.7** | **97.2** |
| HIF+$KAT_{ID3}$ | 95.8 | **96.6** | **88.2** | 51.6 | 79.0 | **79.5** | 94.5 | **96.7** | **97.2** |
| HIF+$KAT_{XGB}$ | 90.6 | 93.3 | 87.9 | 41.5 | **80.3** | **79.5** | 94.4 | 96.2 | **97.2** |
| HIF+LN | 77.9 | 21.0 | 54.7 | 41.6 | - | 78.5 | 72.2 | 62.8 | 86.0 |
| HIF+LR | 84.2 | 87.1 | 84.6 | 46.5 | - | 68.1 | 87.5 | 41.7 | 62.0 |
| HIF-WBOW | 93.0 | 92.7 | 75.4 | 43.2 | 47.9 | 43.7 | 91.6 | 66.3 | 74.0 |
| HIF-EMB | 91.1 | 90.9 | 76.6 | 30.8 | 53.9 | 46.8 | 89.9 | 65.7 | 79.8 |
| HIF-ALONE | 94.6 | 96.1 | 82.9 | 45.6 | 73.5 | 63.2 | 91.8 | 63.0 | 72.9 |

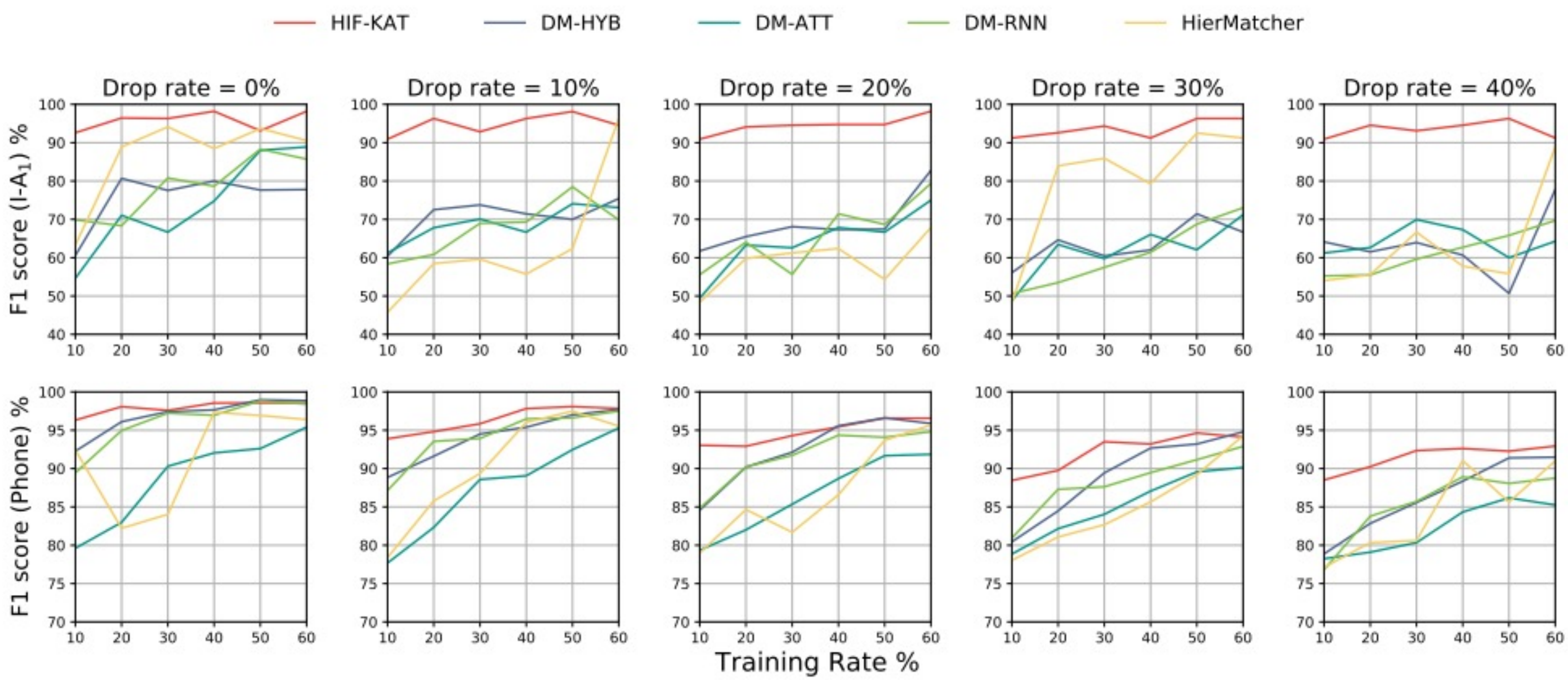
# Low Resource Entity Matching

- Evaluation metrics: $F_1$ measure

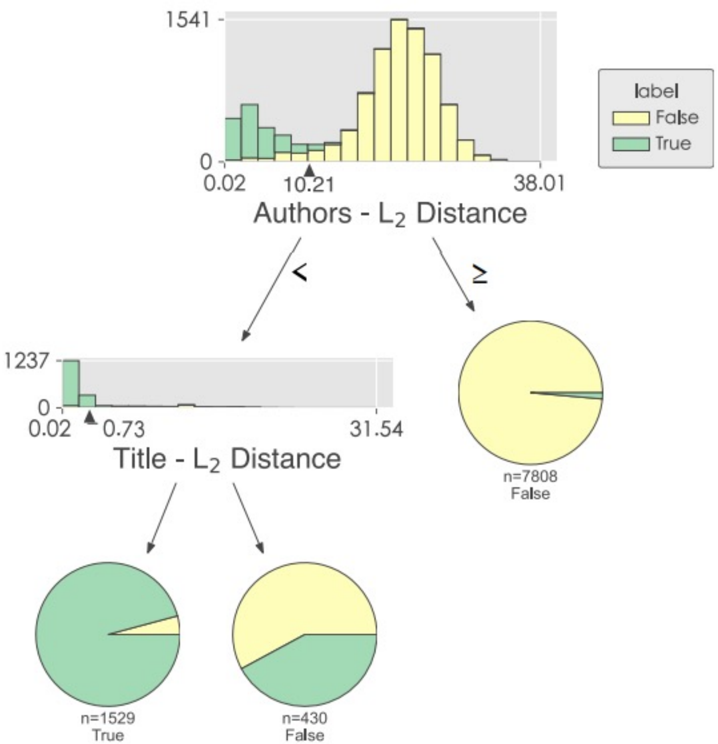
| Methods | I-$A_1$ | D-$A_1$ | D-$S_1$ | I-$A_2$ | D-$A_2$ | D-$S_2$ | Phone | Skirt | Toner |
|---|---|---|---|---|---|---|---|---|---|
| DM-RNN | 63.6 | 85.4 | 74.8 | 42.3 | 45.7 | 39.0 | 90.0 | 67.6 | 68.6 |
| DM-ATT | 55.8 | 82.5 | 79.0 | 46.5 | 45.2 | 57.8 | 80.3 | 54.4 | 48.8 |
| DM-HYB | 60.9 | 86.6 | 78.0 | 49.5 | 46.2 | 60.4 | 91.9 | 64.2 | 67.4 |
| HierMatcher | 61.9 | 37.5 | 68.2 | 37.8 | 32.6 | 45.8 | 86.2 | 61.7 | 55.2 |
| Magellan | 92.3 | 93.7 | 85.1 | 50.6 | 65.6 | 71.1 | 93.6 | 96.6 | **97.2** |
| HIF+DT | **96.0** | 96.4 | 87.5 | **54.9** | 80.1 | 74.2 | **94.9** | **96.7** | **97.2** |
| HIF+$KAT_{ID3}$ | 95.8 | **96.6** | **88.2** | 51.6 | 79.0 | **79.5** | 94.5 | **96.7** | **97.2** |
| HIF+$KAT_{XGB}$ | 90.6 | 93.3 | 87.9 | 41.5 | **80.3** | **79.5** | 94.4 | 96.2 | **97.2** |
| HIF+LN | 77.9 | 21.0 | 54.7 | 41.6 | - | 78.5 | 72.2 | 62.8 | 86.0 |
| HIF+LR | 84.2 | 87.1 | 84.6 | 46.5 | - | 68.1 | 87.5 | 41.7 | 62.0 |
| HIF-WBOW | 93.0 | 92.7 | 75.4 | 43.2 | 47.9 | 43.7 | 91.6 | 66.3 | 74.0 |
| HIF-EMB | 91.1 | 90.9 | 76.6 | 30.8 | 53.9 | 46.8 | 89.9 | 65.7 | 79.8 |
| HIF-ALONE | 94.6 | 96.1 | 82.9 | 45.6 | 73.5 | 63.2 | 91.8 | 63.0 | 72.9 |

# Sensitivity Test

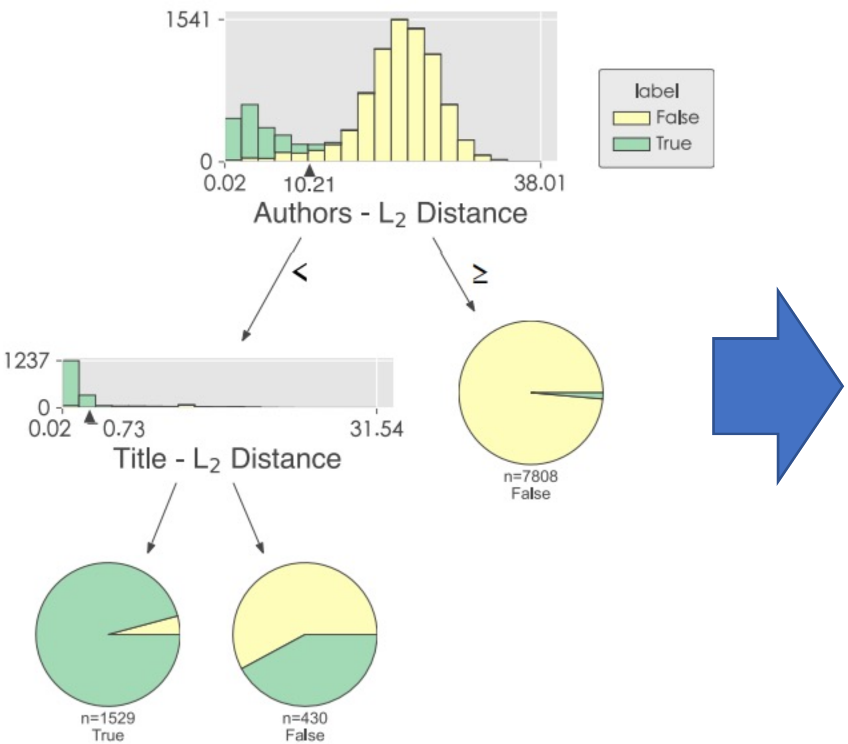
- Control variables:
  - Training set size
  - Missing rate



# Case Study & Interpretability

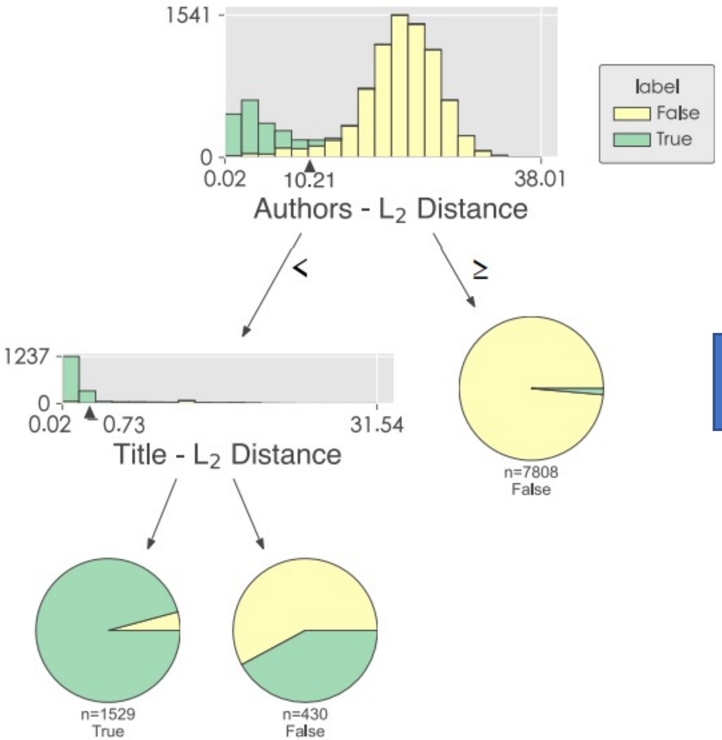


# Case Study & Interpretability




Rule 1: if $L_2\left(\mathrm{HIF}(e_1), \mathrm{HIF}(e_2)\right)[\mathrm{Authors}] \geq 10.21$
then $e_1, e_2$ are not a match;
Rule 2: if $L_2\left(\mathrm{HIF}(e_1), \mathrm{HIF}(e_2)\right)[\mathrm{Authors}] < 10.21$
$\wedge\, L_2\left(\mathrm{HIF}(e_1), \mathrm{HIF}(e_2)\right)[\mathrm{Title}] < 0.73$
then $e_1, e_2$ are a match;
Rule 3: if $L_2\left(\mathrm{HIF}(e_1), \mathrm{HIF}(e_2)\right)[\mathrm{Authors}] < 10.21$
$\wedge\, L_2\left(\mathrm{HIF}(e_1), \mathrm{HIF}(e_2)\right)[\mathrm{Title}] \geq 0.73$
then $e_1, e_2$ are not a match

# Case Study & Interpretability




*Rule 1: if two records have different authors, they will be different publications.*
*Rule 2: if two records have similar authors and similar titles, they will be the same publication.*
*Rule 3: if two records have similar authors and dissimilar titles, they will not be the same publication.*
The soundness of such rules can be examined by our experience.

# Index

# Conclusion and Future Work

- Conclusion
  - The decoupled framework provides a paradigm for utilizing unlabeled data and providing interpretable EM process.
- Future Work
  - Leveraging extra entity records
  - Incorporating pre-trained language models
  - Incorporating HIF with other EM models

# Thanks!

- We thank the reviewers, organizers and audiences.
- Codes & Datasets: https://github.com/THU-KEG/HIF-KAT
- Contact: yaozj20@mails.tsinghua.edu.cn