TAIC

Tencent Academic and Industrial Conference

Changwang ZHANG

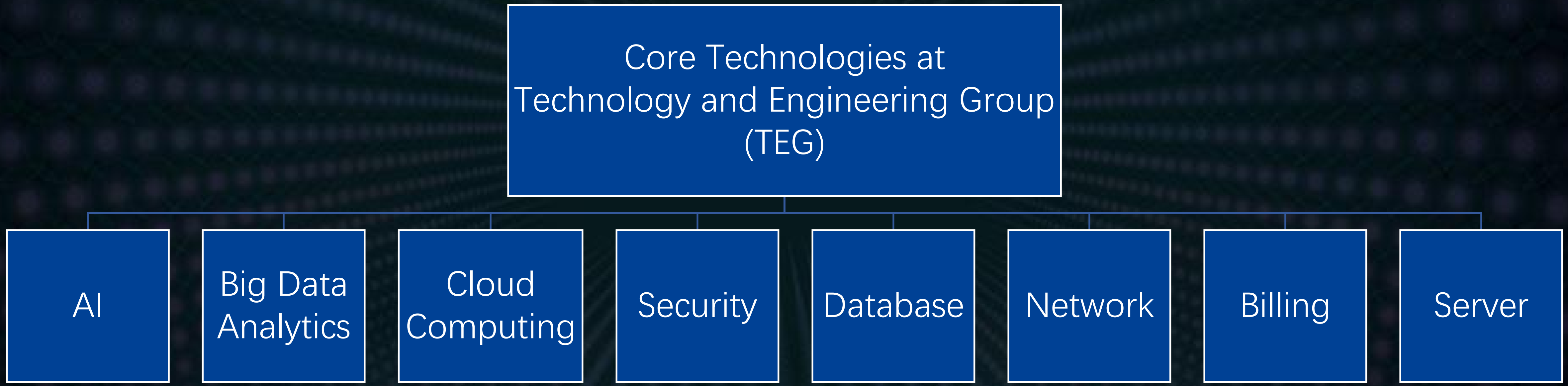
IJCAI-19, Macau SAR, China
August 13, 2019

Tencent 腾讯

Big-data Mining Using Unsupervised Learning and Graph Learning

Technology and Engineering Group (TEG)

- Key values:
  - Develop core/cutting-edge technologies/platforms for the whole company
  - Support a broad range of applications
  - Advance the state of the art and impact the world



Outline

- Unsupervised short text hierarchical classification
- Point-of-interest (POI) matching based on domain knowledge
- Large-scale graph mining and learning

Unsupervised short text hierarchical classification

An user vs. AI Master

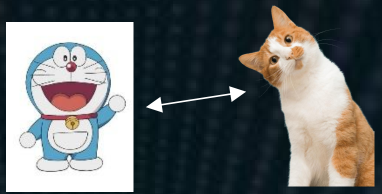
AI is changing the world,
AI is beating humans ...

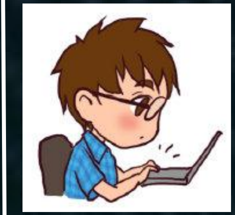
Wow, please help me do a three-level text classification

Ok. You have 800 classes then provide 8 million labelled data to train the model...

8 million.......

AI Master requires too much labelled data!

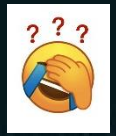




"robot cat" is misclassified as "pets", can I help you rectify that?

You human want to modify a model with millions of parameters? No way! Go and label more data...

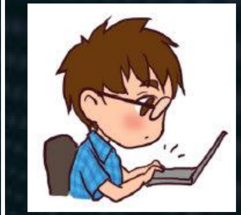
......



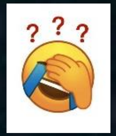
AI Master is not controllable by the user

What can I do if I want to adjust the meaning and levels of the targeted categories?

You can do nothing but to re-label 2 million data to further train the model

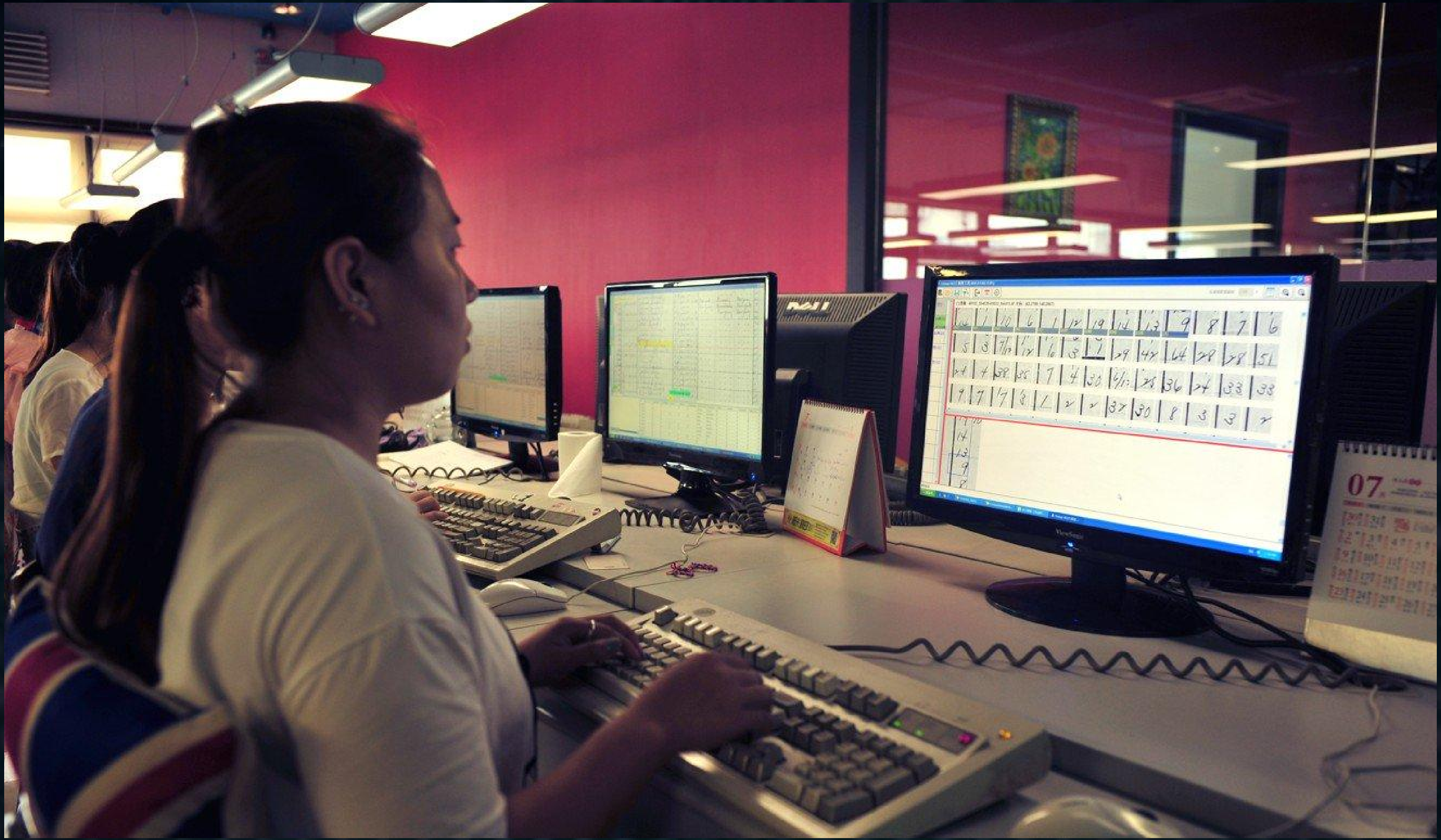


......



AI Master is not easily adaptive

AI master depends on massive human effort in labelling!



What is the problem and the solution?

The Problem:

AI Master requires too much labelled data!

AI Master is not controllable by the user

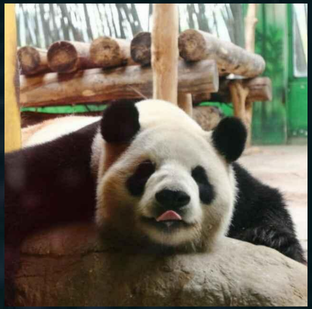
AI Master is not easily adaptive

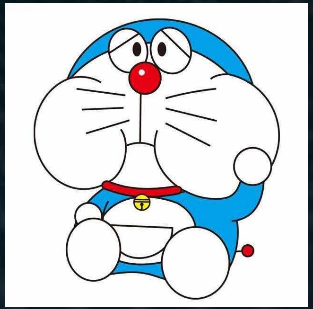
AI Master is learning patterns but not knowledge.

AI Master does not really understand the meaning of the targeted categories, let alone consider the meaning while doing the classification.

The Solution:

Unsupervised short text hierarchical classification algorithm based on keyword and category knowledge



The knowledge

Keyword knowledge:

- Keyword web search context
- Keyword encyclopedia context
- Keyword to category word posterior correlation probability

Unlabeled Item

Item
烟台红富士 (Yantai Red Fuji)
…

Labeled Item

Item	label
烟台红富士 (Yantai Red Fuji)	生鲜>>水果>>苹果 (Fresh>>Fruit>>Apple)
…	…

Posterior probability

$$P(w_c \mid w_k) = \frac{P(w_c)P(w_k \mid w_c)}{P(W_k)}$$

$$= \frac{\text{The number of items containing } w_k \text{ in category } C}{\text{The total number of items containing } w_k}$$

Keyword	Category word	Correlation probability
Red Fuji	Fresh	0.85
Red Fuji	Fruit	0.90
Red Fuji	Apple	0.98
Walmart	Pigeon egg	0.18
…	…	…

No dependence on massive human labelled data

Category knowledge:

Category name	Category expression	Category priority
Pet-Cat	Pet>>Cat	1
Pet-Dog	Pet>>Dog	1
Pet-Other	Pet>>$other$	1

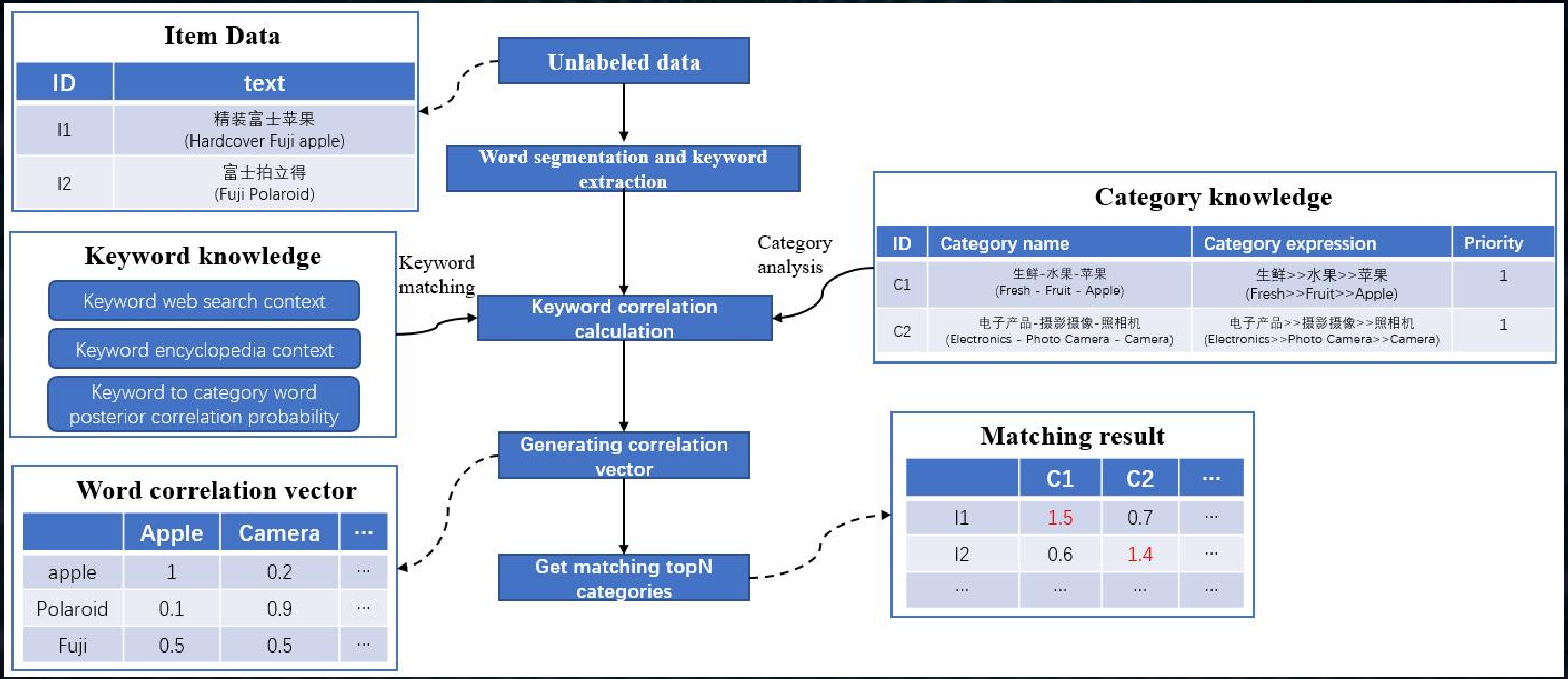
No cartoon characters in pets
Leopards are big cats
No husky in dogs
Prefer dogs over other pets

Category name	Category expression	Category priority
Pet-Cat	Pet-Cartoon>>Cat+Leopard	1
Pet-Dog	Pet-Cartoon>>Dog-Husky	1.5
Pet-Other	Pet-Cartoon>>$other$	1

Understandable, adaptable and controllable by the user

The algorithm

Item
Keyword + Category knowledge } ⊗ ⇒ Item category



Experimental setup

Datasets:

Point of interest (POI) name classification

Items	Labels
经典发型沙龙 (Classic hair salon)	生活服务:美容美发:美发 (Life service: beauty salon: hairdressing)
森马 (Semir)	购物:服饰鞋包 (Shopping: dress shoes bag)
华美通科技 (Huameitong Technology)	公司企业:公司企业 (Enterprise: company)

Total classes: 351 / Total levels: 3

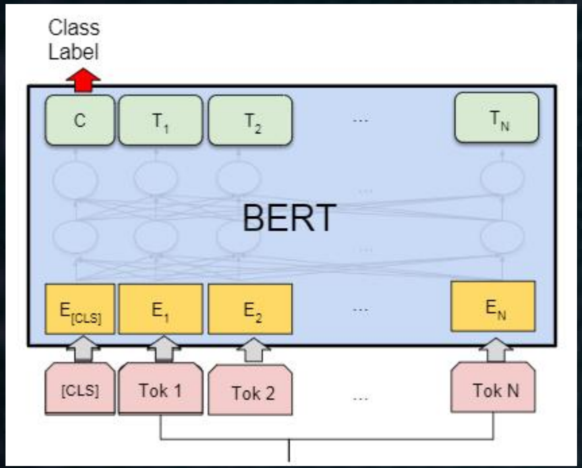
eCommerce item name classification

Items	Labels
华为mate 20 pro钢化膜全屏彩膜贴膜 (Huawei mate 20 pro tempered film full screen color)	手机:手机配件:手机贴膜 (Mobile phone: mobile phone accessories: mobile phone film)
YONEX 尤尼克斯YY羽毛球长裤 (YONEX Yonex YY badminton trousers)	运动户外:体育用品:羽毛球服 (Sports Outdoor: Sporting Goods: Badminton Wear)
盐焗腰果 400克 (Salted cashew nuts 400g)	食品饮料:休闲食品:坚果炒货 (Food Beverage: Snack Food: Nuts Roasted)

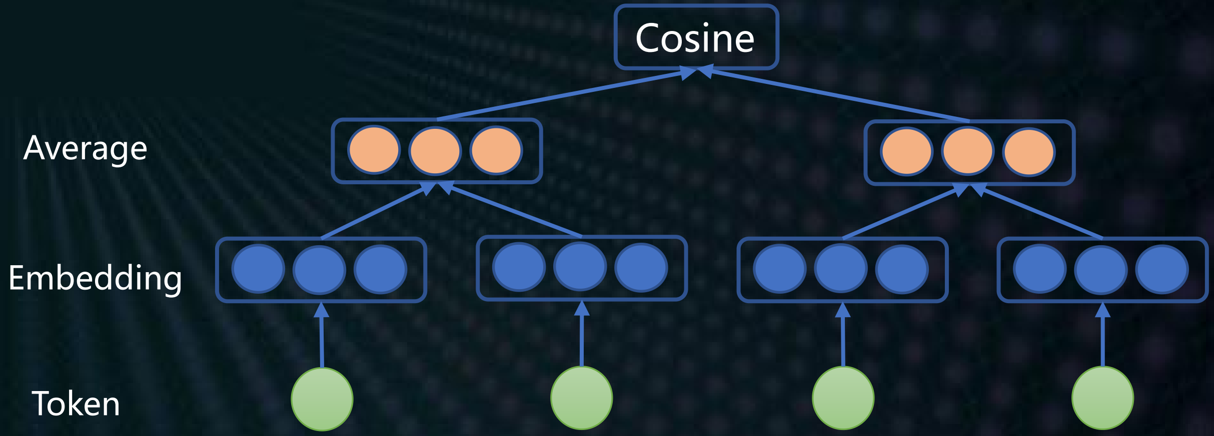
Total classes: 4234 / Total levels: 3

Algorithms:

Supervised methods: BERT fine-tuning



Unsupervised methods: word2Vec+cosine similarity



Experimental results

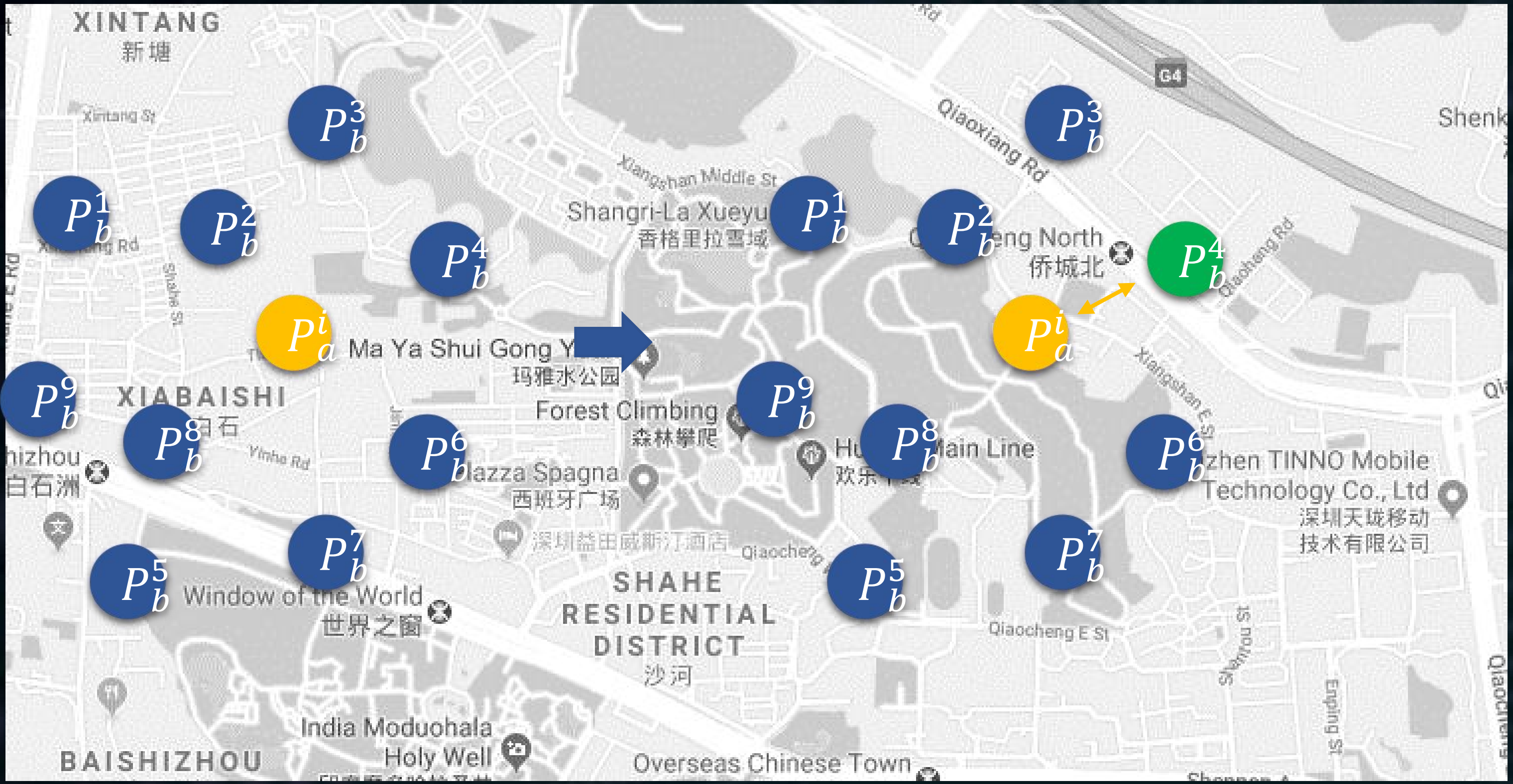
Dataset	Model	First level category			Second level category			Third level category		
		P	R	F	P	R	F	P	R	F
Point of interest (POI) name	Bert word vector+Cosine	60%	58%	59%	28%	27%	27%	23%	16%	19%
	Tencent AI Lab word vector+Cosine	65%	65%	65%	40%	40%	40%	29%	33%	31%
	Our method	81%	28%	42%	59%	20%	30%	50%	15%	23%
	Our method (Knowledge on 1k training data)	85%	42%	56%	61%	31%	41%	59%	29%	39%
	Our method (Knowledge on 10k training data)	87%	55%	67%	61%	39%	48%	59%	39%	48%
	BERT fine-tuning (1k training data)	17%	17%	17%	2.4%	2.4%	2.4%	0.5%	1.7%	0.8%
	BERT fine-tuning (10k training data)	83%	83%	83%	66%	66%	66%	61%	58%	60%
eCommerce item name	Bert word vector+Cosine	49%	49%	49%	30%	30%	30%	7%	7%	7%
	Tencent AI Lab word vector+Cosine	68%	68%	68%	46%	46%	46%	15%	15%	15%
	Our method	80%	80%	80%	66%	65%	66%	47%	46%	46%
	Our method (Knowledge on 1k training data)	81%	81%	81%	66%	65%	66%	47%	46%	46%
	Our method (Knowledge on 10k training data)	84%	84%	84%	70%	70%	70%	50%	49%	50%
	BERT fine-tuning (1k training data)	31%	31%	31%	14%	14%	14%	2.2%	2.2%	2.2%
	BERT fine-tuning (10k training data)	80%	80%	80%	71%	71%	71%	53%	53%	53%

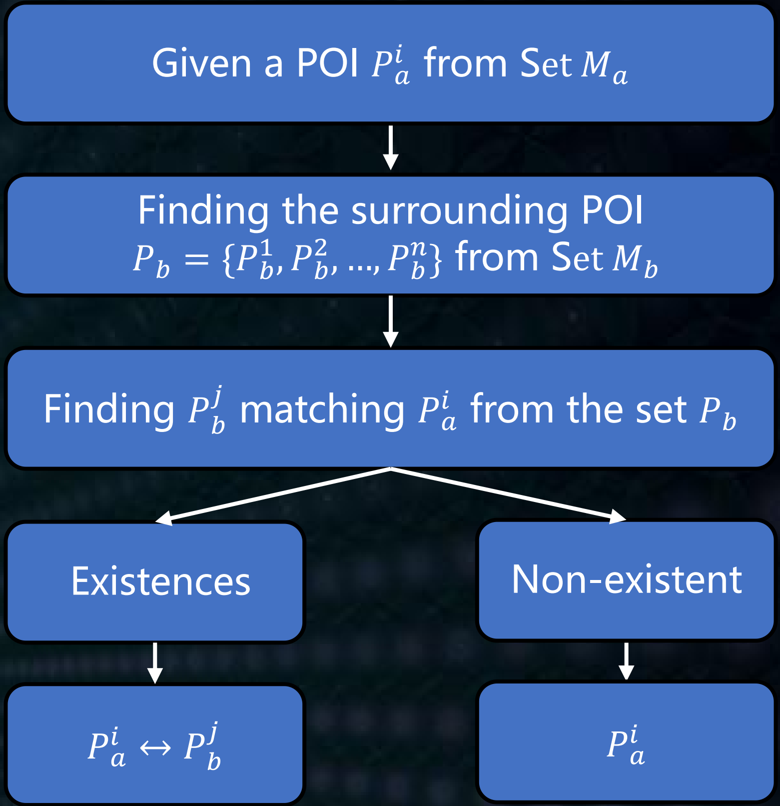
Some final thoughts

- Category knowledge can help effectively filter inaccurate instances and improve classification accuracy.
- Keyword web search and encyclopedia context is helpful in understanding the semantics of short text.
- Keyword to category word posterior correlation probability supplements the classifier with the domain knowledge in item samples.

Point-of-interest (POI) matching based on domain knowledge

Point-of-interest (POI) matching

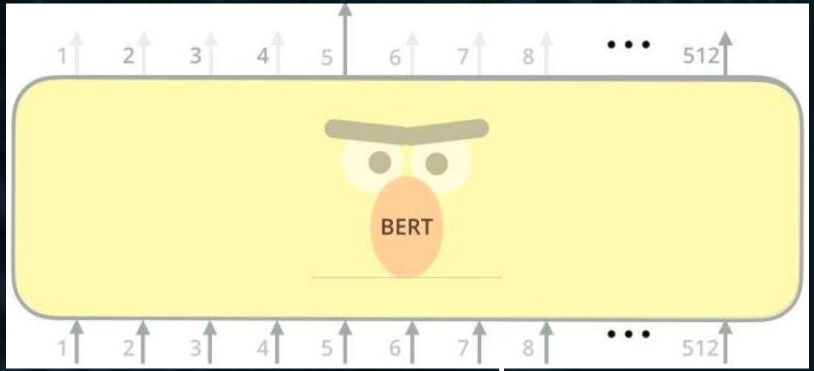




Poi Set1				Poi Set2				Match
Name	Type	Longitude	Latitude	Name	Type	Longitude	Latitude	
海岸城购物中心 (Coastal City Shopping Center)	购物:综合商场 (Shopping: Shopping Malls)	113.9352	22.5173	海岸城 (Coastal City)	购物服务:商场:普通商场 (Shopping Service: Shopping Mall: Ordinary Shopping Mall)	113.9347	22.5184	1
海岸城购物中心 (Coastal City Shopping Center)	购物:综合商场 (Shopping: Shopping Malls)	113.9352	22.5173	海岸城东座 (Coastal City East Block)	商务住宅:楼宇:商务写字楼 (Business Residence: Building: Business Office Building)	113.9361	22.5170	0

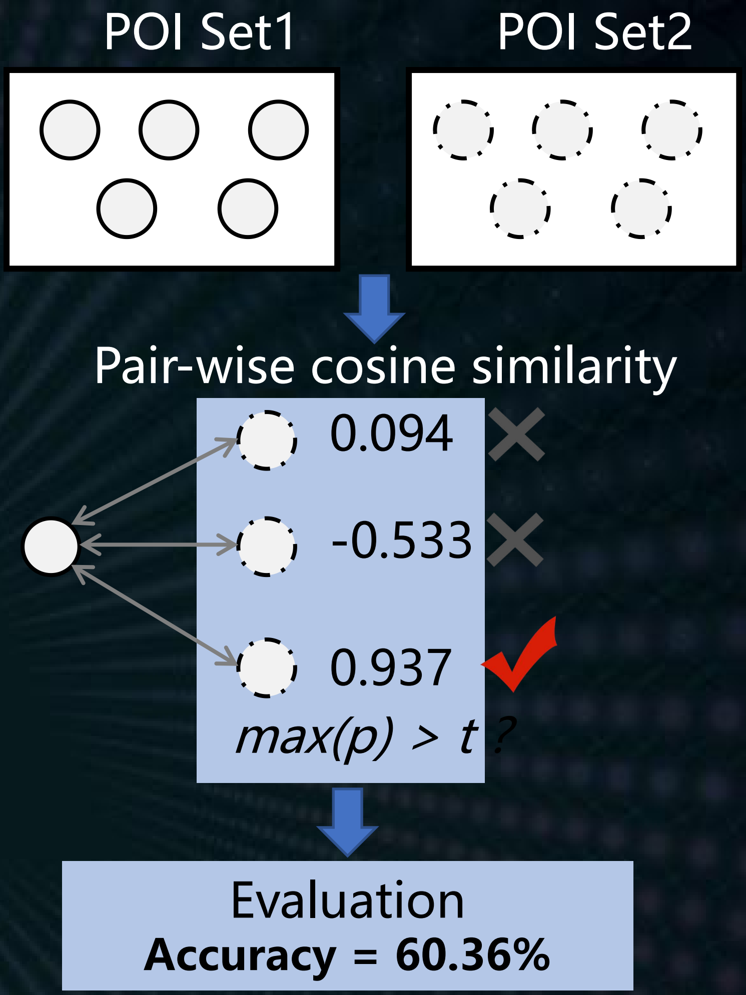
Classic methods-Unsupervised matching

P1: 0.4837323, 0.055763606, ……, -0.37044808
P2: 0.064815, -0.121840894, ……, 0.18656228
……



name, type

id	name	type	longitude	latitude
P1	Fornet Laundry (Hyde 3rd Road)	Life Service; Laundry; Laundry	114.93517	22.099756
P2	Shenzhen Kempinski Hotel	Accommodation service; Hotel; Five-star Hotel	113.44935	22.529044
		……		

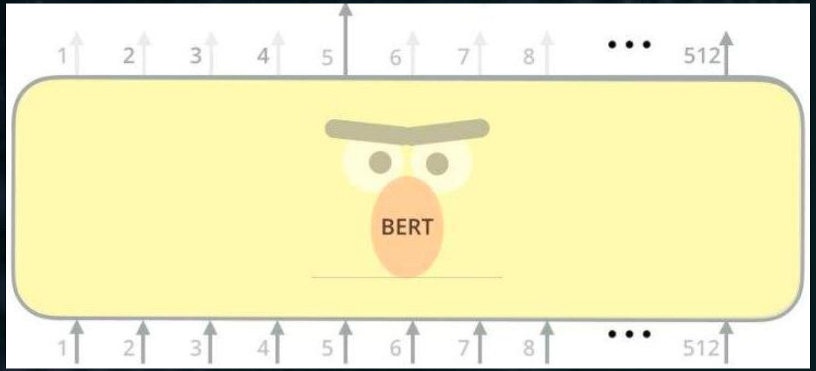


Experimental datasets:

- Set1: 8,071 POIs Set2: 4,799 POIs / Matched POI pairs: 2,631
- 75% used for training, 25% used for the test

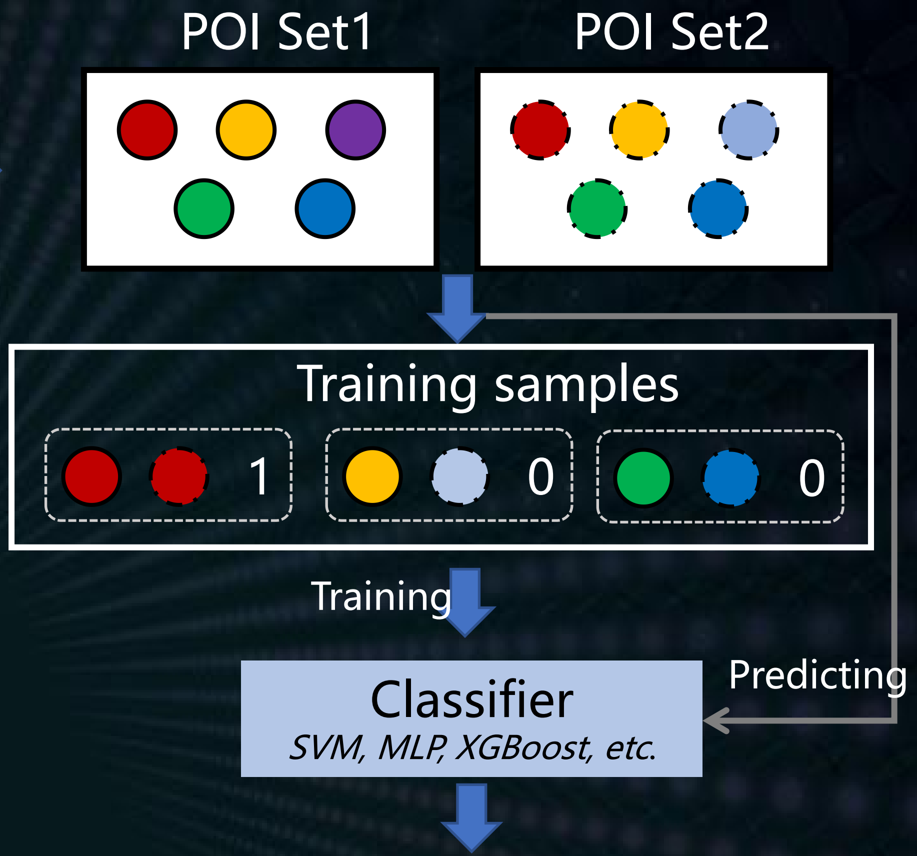
Classic methods-Supervised classification

P1: 0.4837323, 0.055763606, ……, -0.37044808
P2: 0.064815, -0.121840894, ……, 0.18656228
……



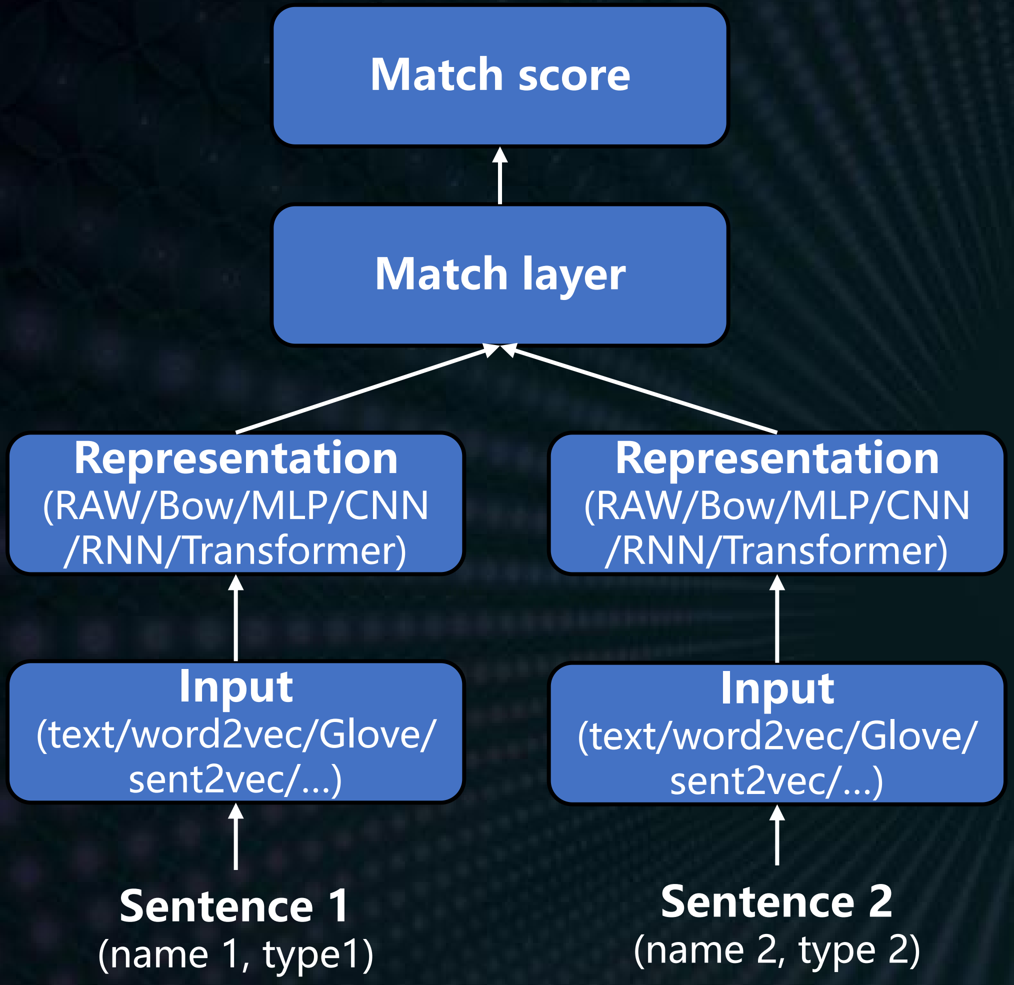
name, type

id	name	type	longitude	latitude
P1	Fornet Laundry (Hyde 3rd Road)	Life Service; Laundry; Laundry	114.93517	22.099756
P2	Shenzhen Kempinski Hotel	Accommodation service; Hotel; Five-star Hotel	113.44935	22.529044
		……		



Method	Accuracy
MLP	55.71%
SVM	72.00%
XGBoost	55.53%

More recent methods-Match net



Text matching

Extracting meaningful matching patterns from words, phrases, and sentences to produce the matching score.

Sentences of POI: "name, type"

- Sentence 1: Coastal City Shopping Center, Shopping: Shopping Malls
- Sentence 2: Coastal City, Shopping Service: Shopping Mall: Ordinary Shopping Mall

Method	Input layer	Representation layer	Match layer	Accuracy
DSSM	text	MLP	Cosine	57.66%
BERT fine-tuning	text	Transformer	-	73.92%

Problems and our solution

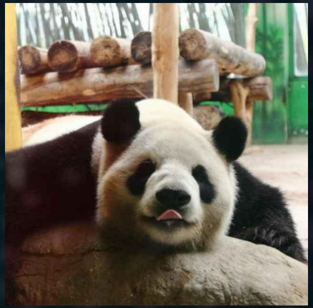
Problems:

- Classic unsupervised methods have low accuracy.
- Existing supervised methods have low accuracy due to inadequate training samples.
- BERT fine-tuning is way too slow in training and prediction.

Our solution:

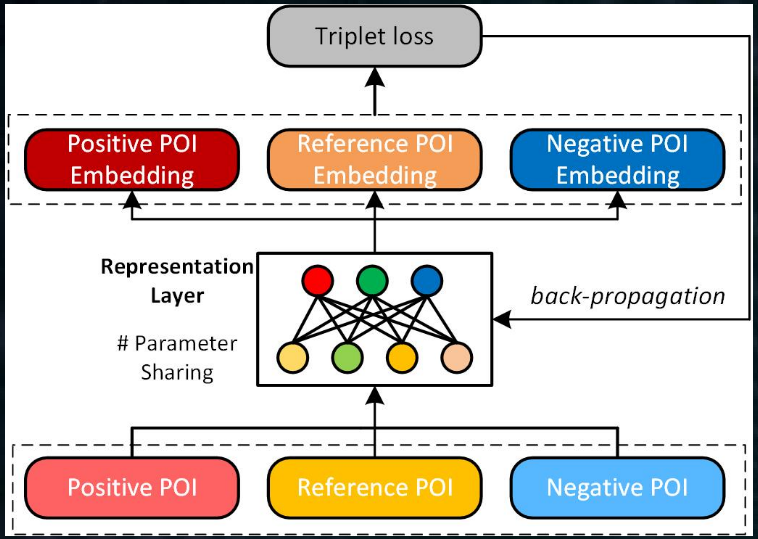
- Reducing reliance on massive training samples by incorporating domain knowledge in the model.
- Using pre-trained BERT vectors and triplet loss for the model instead of BERT fine-tuning.

Method	Input layer	Representation layer	Match layer	Accuracy
Cosine	BERT pre-trained vectors	-	Cosine	60.35%
MLP	BERT pre-trained vectors	MLP	Softmax	55.71%
SVM	BERT pre-trained vectors	-	-	72.00%
XGBoost	BERT pre-trained vectors	-	-	55.53%
DSSM	text	MLP	Cosine	57.66%
BERT fine-tuning	text	Transformer	-	73.92%



Our method

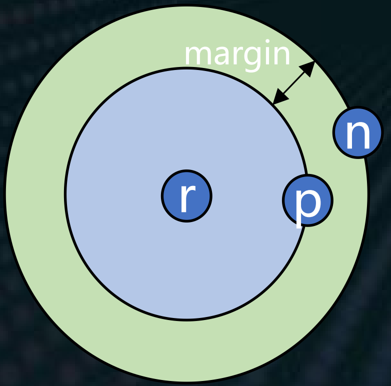
POI Representation Net v1



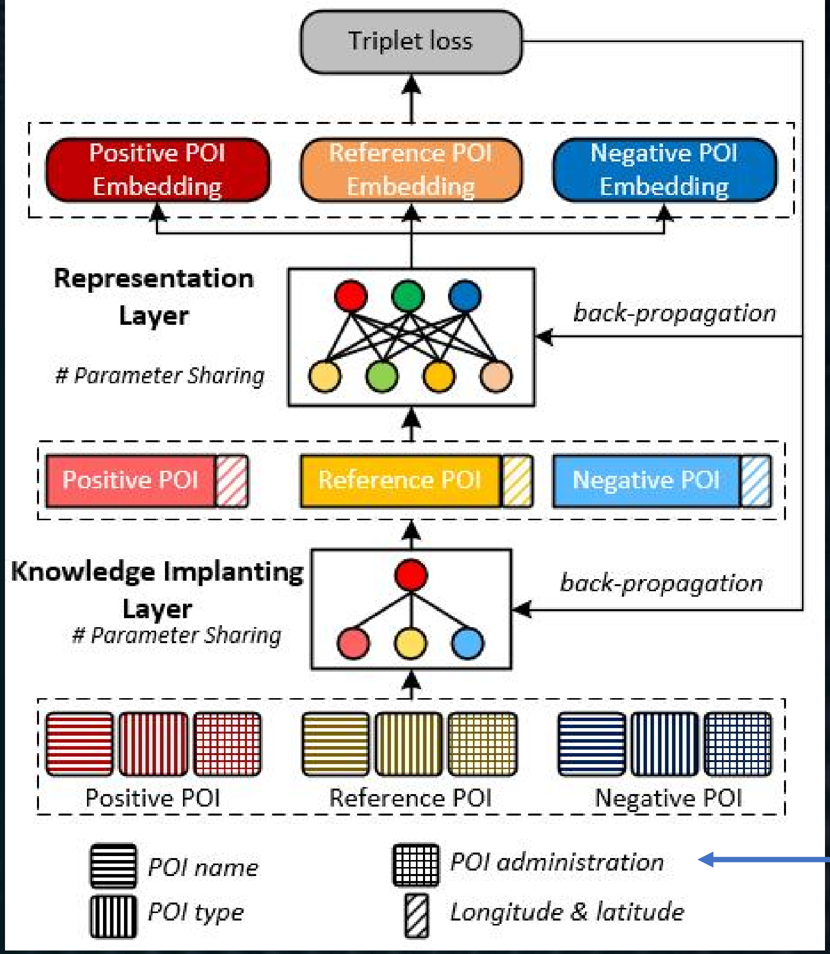
Triplet loss

$$L = \max(0,\ margin - (d(r,n) - d(r,p)))$$

Directly compare the distance between the embedding rather than the matched results.



POI Representation Net v2 (+Domain Knowledge)



Administrative area name knowledge
深圳市 (Shenzhen City)
南山区 (Nanshan District)
广州市 (Guangzhou City)
……

- *The spatial attribute is a key characteristic of geospatial big data.*

Results and Conclusion

Method	Input Layer	Representation Layer	Matching Layer	Accuracy
Cosine Matching	BERT pre-trained vectors	-	Cosine	60.35%
Cosine Matching (+Domain Knowledge)	BERT pre-trained vectors	-	Cosine	70.54%
MLP	BERT pre-trained vectors	MLP	Softmax	55.71%
SVM	BERT pre-trained vectors	-	-	72.00%
XGBoost	BERT pre-trained vectors	-	-	55.53%
DSSM	text	MLP	Cosine	57.66%
BERT fine-tuning	text	Transformer	-	73.92%
Our method v1	BERT pre-trained vectors	MLP	Cosine	65.92%
Our method v2 (+Domain Knowledge)	BERT pre-trained vectors + Knowledge implanting	MLP	Cosine	**85.64%**

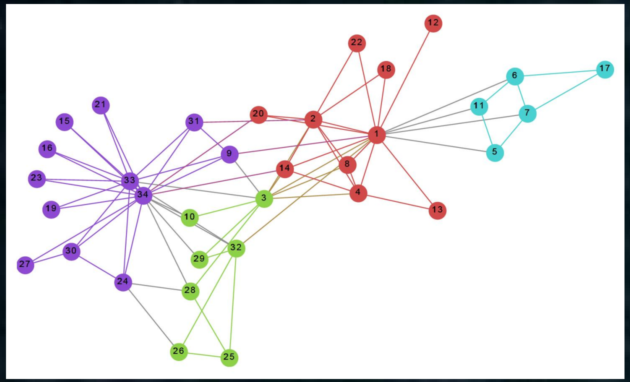
Conclusions

- The proposed model using pre-trained BERT vectors and triplet loss is an efficient and accurate solution.
- The introduction of domain knowledge can improve accuracy for both supervised and unsupervised methods.

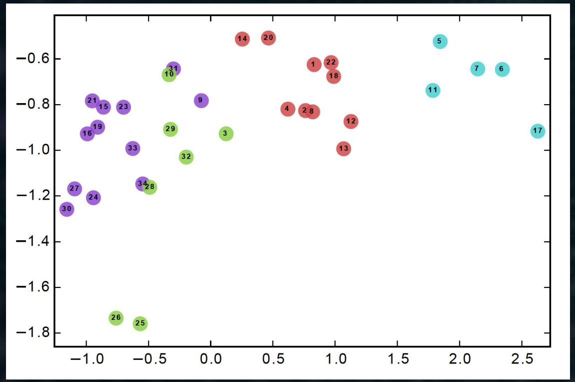
Large-scale graph mining and learning

Graph Representation Learning

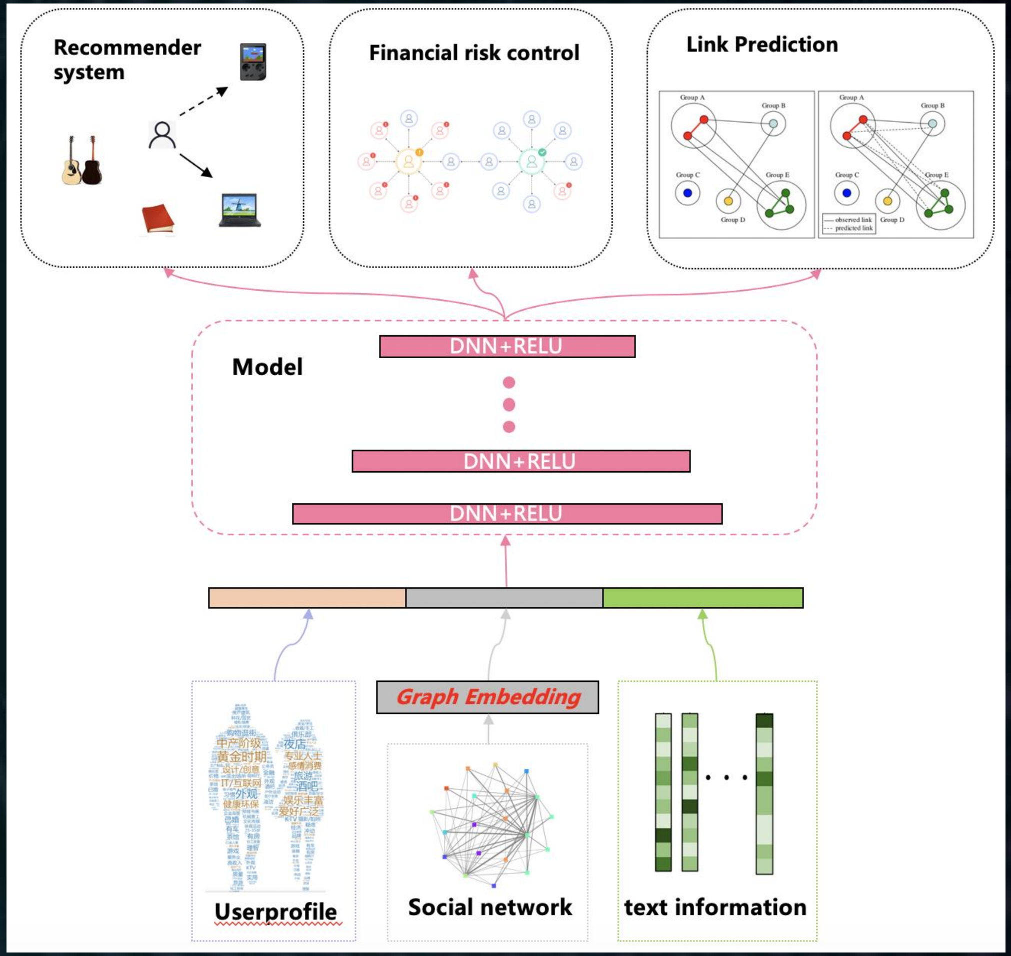
Definition



Learning



Application



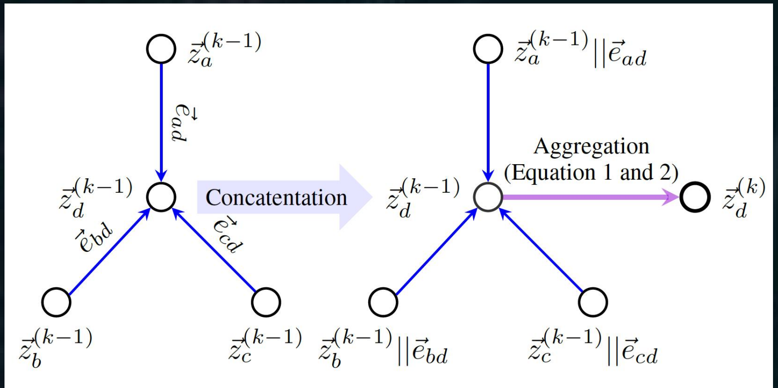
Challenge

- Large-Scale network: million nodes with Ten million edges
- Rich information
  - Node attributes
  - Edge attributes
  - Dynamic
  - Directed
  - Heterogeneous

EdgeProp

- Motivation
  - A research blind zone for edge attributes
  - Not enough research for heterogeneous network
- Contribution
  - **A new message passing mechanism** that allows edge information to propagate into node representation
  - **A scalable realization** of the proposed algorithm

$$h_v^{(k)} = \rho^{(k)}(h_v^{(k-1)} || h_{N(v)}^{(k)}) = \rho^{(k)}(h_v^{(k-1)} || \sum_{u \epsilon N(v)} \emptyset^{(k)}(h_u^{k-1} || e_{uv}))$$



Algorithm 1: Embedding Generation (i.e. forward propagation) Algorithm

input : *Graph* $\mathcal{G}(\mathcal{V}, \mathcal{E})$;
input features $\vec{x}_v, \forall v \in \mathcal{B}$;
edge features $\vec{e}_{ij}, \forall (i,j) \in \mathcal{B}$;
depth K;
differentiable functions ρ, ϕ (e.g. a MLP);
neighborhood sampling function
$\mathcal{N}_k : v \mapsto 2^N$

output: Vector representations $\vec{z}_v, \forall v \in \mathcal{V}$

$\mathcal{B}^K \leftarrow \mathcal{B}$;
for $k = K...1$ **do**
  $B^{k-1} \leftarrow \mathcal{B}^k$;
  for $u \in \mathcal{B}^k$ **do**
    $B^{(k-1)} \leftarrow B^{(k-1)} \cup \mathcal{N}_k(u)$
  end for
end for
$\vec{z}_v^0 \leftarrow \vec{x}_v, \forall v \in \mathcal{B}^0$;
for $k = 1...K$ **do**
  for $v \in \mathcal{B}^k$ **do**
    $\vec{z}_{\mathcal{N}(v)}^{(k)} \leftarrow \sum_{u \in \mathcal{N}(v)} \phi^{(k)}(\vec{z}_u^{(k-1)} || \vec{e}_{uv})$;
    $\vec{z}_v^{(k)} \leftarrow \rho^{(k)}(\vec{z}_v^{k-1} || \vec{z}_{\mathcal{N}(v)}^{(k)})$;
  end for
end for
return $\{\vec{z}_v^{(k)} \forall v \in \mathcal{V}\}$;

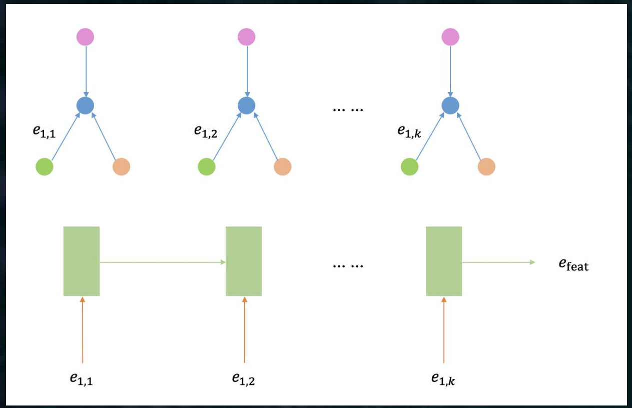
Tencent transaction dataset

Algorithm	Accuracy	Precision	recall	f1 measure
DeepWalk	0.6847	0.6648	0.6286	0.6462
Line	0.6841	0.6636	0.6293	0.6460
Logistic Regression	0.6822	0.6253	0.6621	0.6432
Random Forest	0.7767	0.7474	0.7490	0.7482
GBDT	0.7634	0.7284	0.7483	0.7382
GraphSAGE	0.8110	0.8138	0.8053	0.8095
EdgeProp (w/o nft. + incoming)	0.8303	0.8305	0.8330	0.8317
EdgeProp (w/o nft. + directed)	0.8365	0.8375	0.8399	0.8387
EdgeProp (w/ nft. + incoming)	0.8639	0.8633	0.8622	0.8627
EdgeProp (w/ nft. + directed)	**0.8698**	**0.8686**	**0.8696**	**0.8690**

Feature work

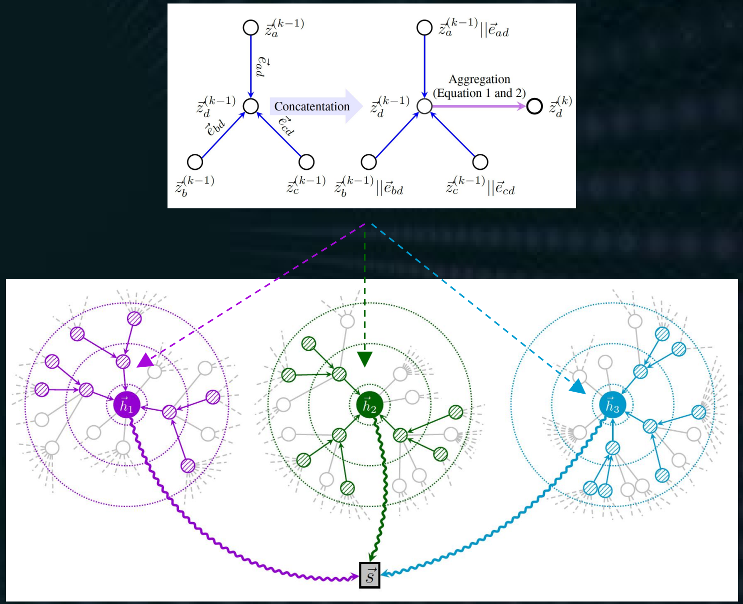
- Make use of time information
- Deal with situation without labels

EdgeProp With LSTM



Method	Accuracy
EdgeProp	91.44%
EdgeProp with LSTM	93.18%

EdgeProp With DGI



Platform Support

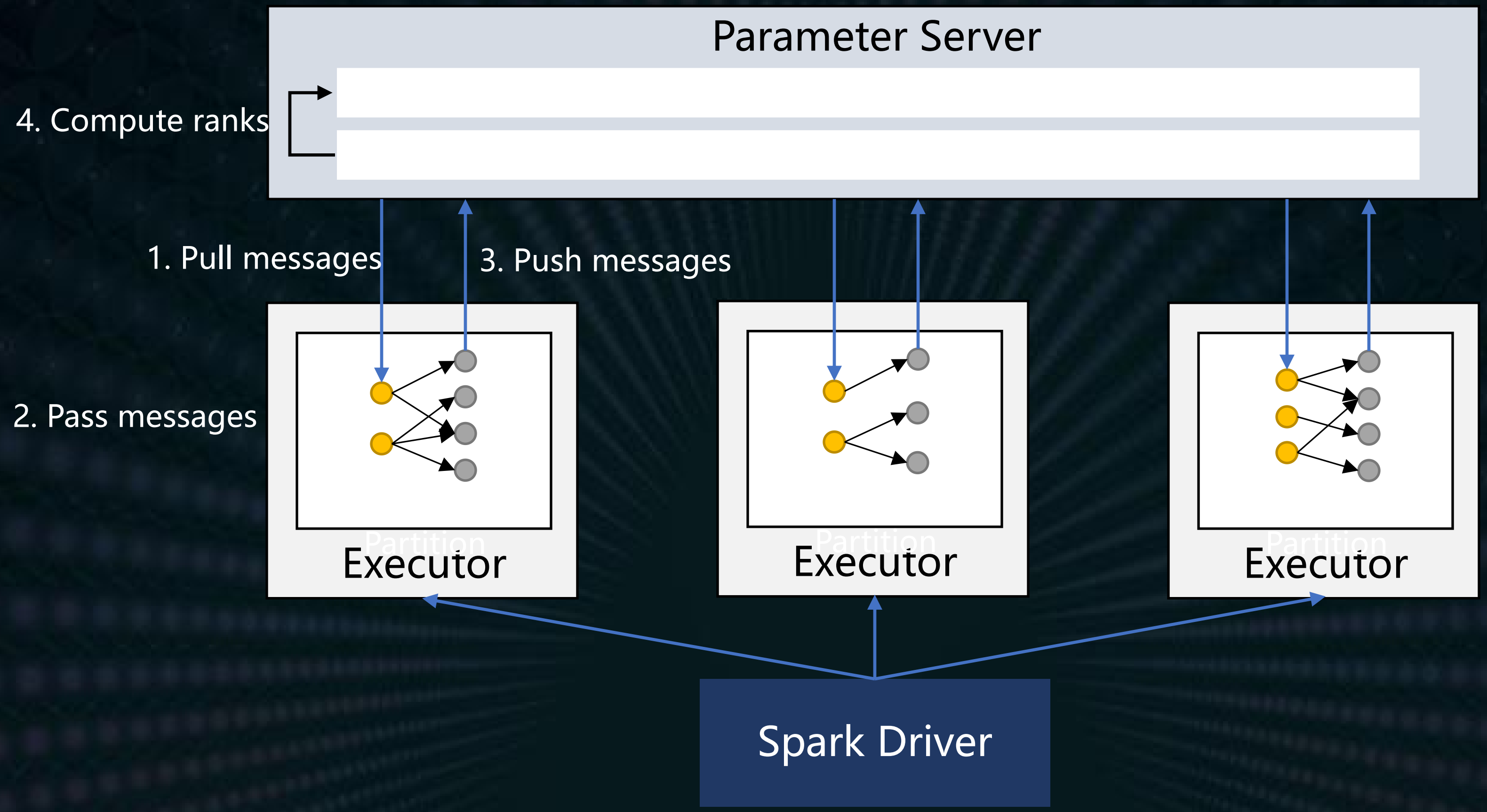


Figure 1. Architecture of Angel's Graph Computing Module: PageRank as an Example

- Enables efficient data sharing (which is essential for complex graph computing) via Parameter Server
- Leverages performance by adopting PS’ s computing power

Platform Support

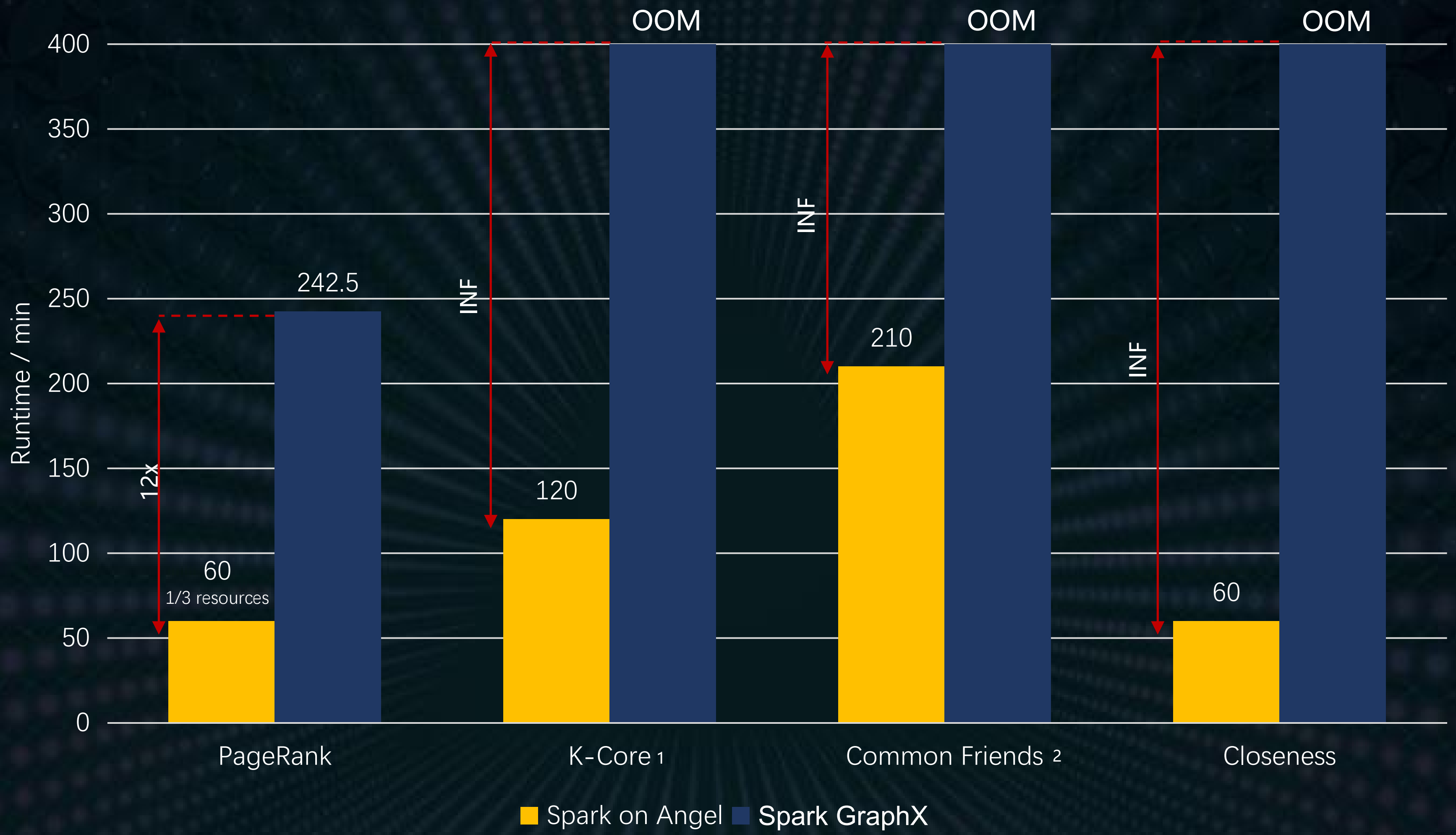


Figure 2. Performance Comparison of Angel and Spark GraphX on Benchmark Graph Mining Algorithms

[1] Iter = 400

[2] Common Friends is tested on dataset with ~150 billion edges, while the other three algorithms are tested on dataset with ~1 billion vertices and ~10 billion edges

Thank you