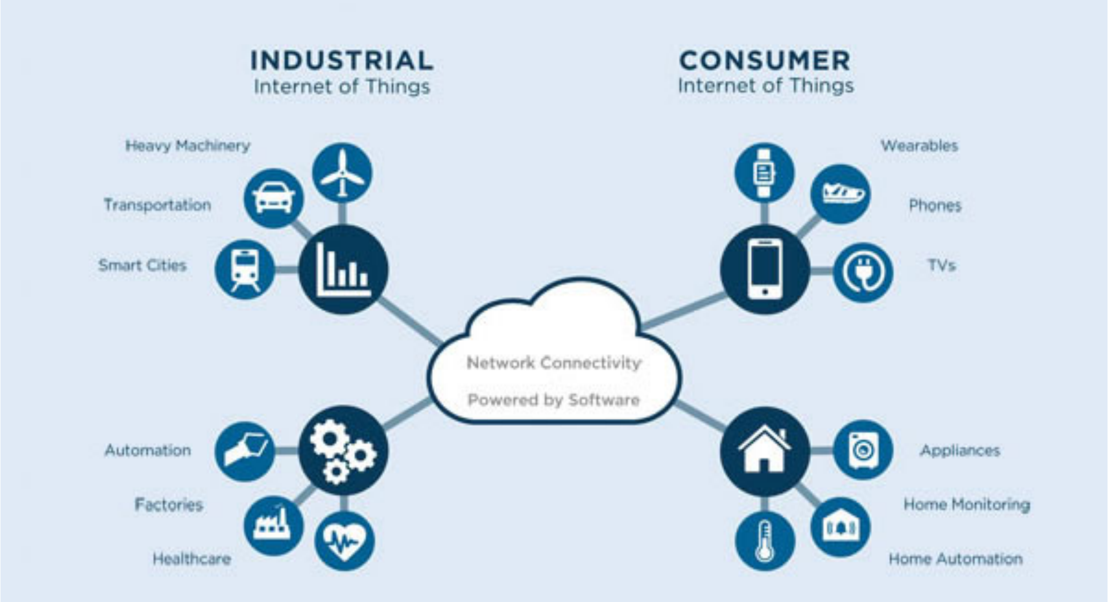
LogTransfer: Cross-System Log Anomaly Detection for Software Systems with Transfer Learning

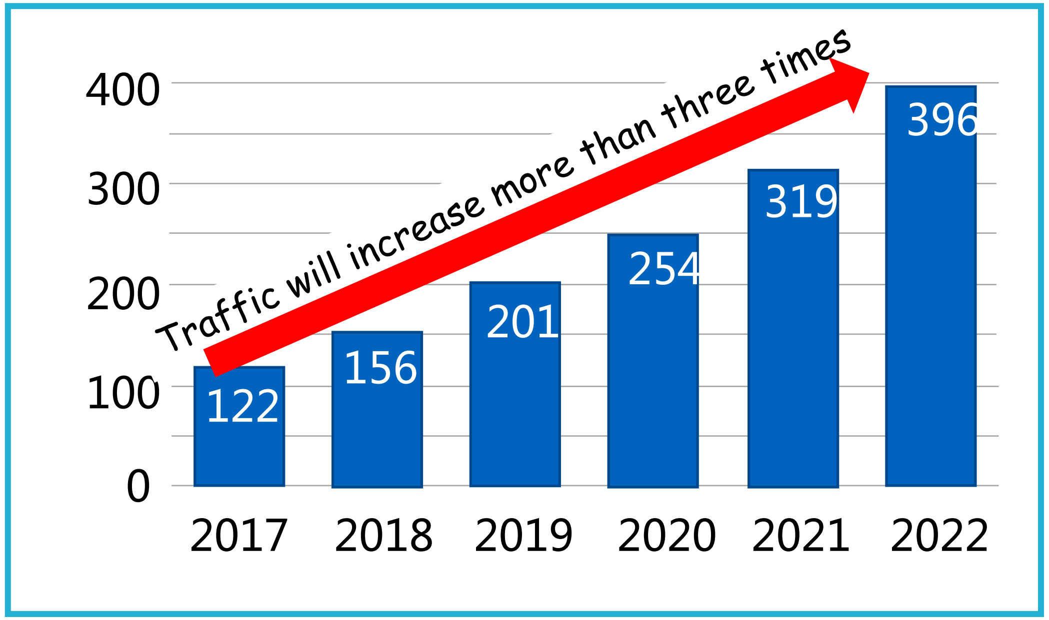
Rui Chen[1], Shenglin Zhang[1], Dongwen Li[1], Yuzhe Zhang[1], Fangrui Guo[1], Weibin Meng[2], Dan Pei[2], Yuzhi Zhang[1], Xu Chen[1], Yuqing Liu[1]

[1]Nankai University, [2]Tsinghua University

Software systems

- Software systems are playing important roles in daily life.
- Reliability as well as availability are highly demanded in modern software systems.





1. cited from: http://www.hidrfid.com/rfid/news/1328.html

System anomaly

System anomalies will cause degradation, impact revenue and user experience.

Delta Says Computer Breakdown Cut Revenue by $100 Million

by Michael Sasso

September 2, 2016 — 9:05 AM EDT *Updated on* September 2, 2016 — 9:17 AM EDT

Delta Air Lines Inc. said the computer failure that caused 2,300 flight cancellations last month cut sales about $100 million and reduced a key revenue figure.

Passenger revenue for each seat flown a mile, an industry benchmark, fell 9.5 percent in August, in part because of the outage and subsequent recovery efforts, the carrier said in a statement Friday. The breakdown reduced unit revenue, as the measure is also known, by two percentage points, Delta said.

The country's second-largest airline earlier forecast that third-quarter unit revenue would fall 4 percent to 6 percent.

A power-control module at Delta's Atlanta computer center failed and caught fire Aug. 8, shutting down electricity to the system. About 300 of the airline's 7,000 servers weren't wired to backup power, the company had said.

Data Center Outages Cost Nearly $9,000 Per Minute

08/24/2016 |

Study concludes cost of downtime has risen 38% in five years

READ MORE ON DATA CENTERS »
RELATED COVERAGE

How to comprehensively and precisely detect anomalies has brought about widespread attention!

1. cited from: https://www.bloomberg.com/news/articles/2016-09-02/delta-says-computer-system-breakdown-cut-revenue-by-100-million
2. cited fromhttps://www.buildings.com/news/industry-news/articleid/20588/title/data-center-outages-cost-nearly-9-000-per-minute-

System anomaly

An instance of a system anomaly

```
Jan 10 06:35:36 192.168.193.40 : 2016 Jan 10 06:35:31 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/40: Ethernet11/40 is suspended
Jan 10 06:35:36 192.168.193.40 : 2016 Jan 10 06:35:31 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/41: Ethernet11/41 is suspended
Jan 10 06:35:58 192.168.193.40 : 2016 Jan 10 06:35:53 GMT: %SYSMGR-SLOT11-2-TMP_DIR_FULL: System temporary directory usage is unexpectedly high at 100%.
Jan 10 06:36:07 192.168.193.40 : 2016 Jan 10 06:36:01 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/40: Ethernet11/40 is suspended
Jan 10 06:36:07 192.168.193.40 : 2016 Jan 10 06:36:01 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/41: Ethernet11/41 is suspended
Jan 10 06:36:10 192.168.193.40 : 2016 Jan 10 06:36:04 GMT: %OSPF-5-ADJCHANGE:  ospf-1 [6940]  Nbr 10.231.36.154 on port-channel102 went DOWN
Jan 10 06:36:37 192.168.193.40 : 2016 Jan 10 06:36:32 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/40: Ethernet11/40 is suspended
Jan 10 06:36:37 192.168.193.40 : 2016 Jan 10 06:36:32 GMT: %ETH_PORT_CHANNEL-5-PORT_SUSPENDED: Ethernet11/41: Ethernet11/41 is suspended
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETH_PORT_CHANNEL-5-PORT_DOWN: port-channel102: Ethernet12/18 is down
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETH_PORT_CHANNEL-5-FOP_CHANGED: port-channel102: first operational port changed from Ethernet12/18 to Ethernet12/19
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETHPORT-5-IF_BANDWIDTH_CHANGE: Interface port-channel102,bandwidth changed to 20000000 Kbit
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETHPORT-5-IF_DOWN_INITIALIZING: Interface Ethernet12/18 is down (Initializing)
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETH_PORT_CHANNEL-5-PORT_DOWN: port-channel102: Ethernet12/19 is down
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETH_PORT_CHANNEL-5-FOP_CHANGED: port-channel102: first operational port changed from Ethernet12/19 to Ethernet12/20
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETHPORT-5-IF_BANDWIDTH_CHANGE: Interface port-channel102,bandwidth changed to 10000000 Kbit
Jan 10 06:36:53 192.168.193.40 : 2016 Jan 10 06:36:48 GMT: %ETHPORT-5-IF_DOWN_INITIALIZING: Interface Ethernet12/19 is down (Initializing)
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETHPORT-5-IF_DOWN_PORT_CHANNEL_MEMBERS_DOWN: Interface port-channel102 is down (No operational members)
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETH_PORT_CHANNEL-5-PORT_DOWN: port-channel102: Ethernet12/20 is down
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETH_PORT_CHANNEL-5-FOP_CHANGED: port-channel102: first operational port changed from Ethernet12/20 to none
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETHPORT-5-IF_BANDWIDTH_CHANGE: Interface port-channel102,bandwidth changed to 100000 Kbit
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETHPORT-5-IF_DOWN_INITIALIZING: Interface Ethernet12/20 is down (Initializing)
Jan 10 06:36:57 192.168.193.40 : 2016 Jan 10 06:36:52 GMT: %ETHPORT-5-IF_DOWN_PORT_CHANNEL_MEMBERS_DOWN: Interface port-channel102 is down (No operational members)
```

System logs

System logs are unstructured text.

A log message is composed of constant part and variable part.

……

[SIF pica_sif] Interface te-1/1/11, changed state to down.

……

Type	Contents
Variable Part	Te-1/1/11
Constant Part	Interface *** changed state to down

Previous methods

Type	Input of Template	Method
Supervised	Template index	Decision tree
		SVM
		CNN-based model
		Linear regression
Unsupervised	Template count	PCA
	Template index	DeepLog
		LogCluster
		Isolation forest
		Invariant mining
	Template embedding	LogAnomaly

Supervised methods: Massive labelling efforts & no semantics information of logs.

Unsupervised methods: Suffering from low accuracy in real-world service systems.

System logs in anomaly detection

Service Type A	Nov 6 20:06:32192.168.190.65 %%10IFNET/3/LINK_UPDOWN(l): GigabitEthernet1/0/10 link status is UP. …… Nov 6 20:06:32 192.168.190.65 %%10IFNET/3/LINK_UPDOWN(l): GigabitEthernet1/0/10 link status is DOWN.
Service Type B	Feb 14 00:37:36 192.168.191.79 2018: [SIF pica_sif]Interface te-1/1/11, changed state to down …… Feb 14 00:37:38 192.168.191.79 2018: [SIF pica_sif]Interface te-1/1/11, changed state to up

Different type of systems/services event generate log sequences with similar semantics.

Can we transfer anomalous patterns from one software system to another one?

Challenges

Different types of systems are different in log syntax.

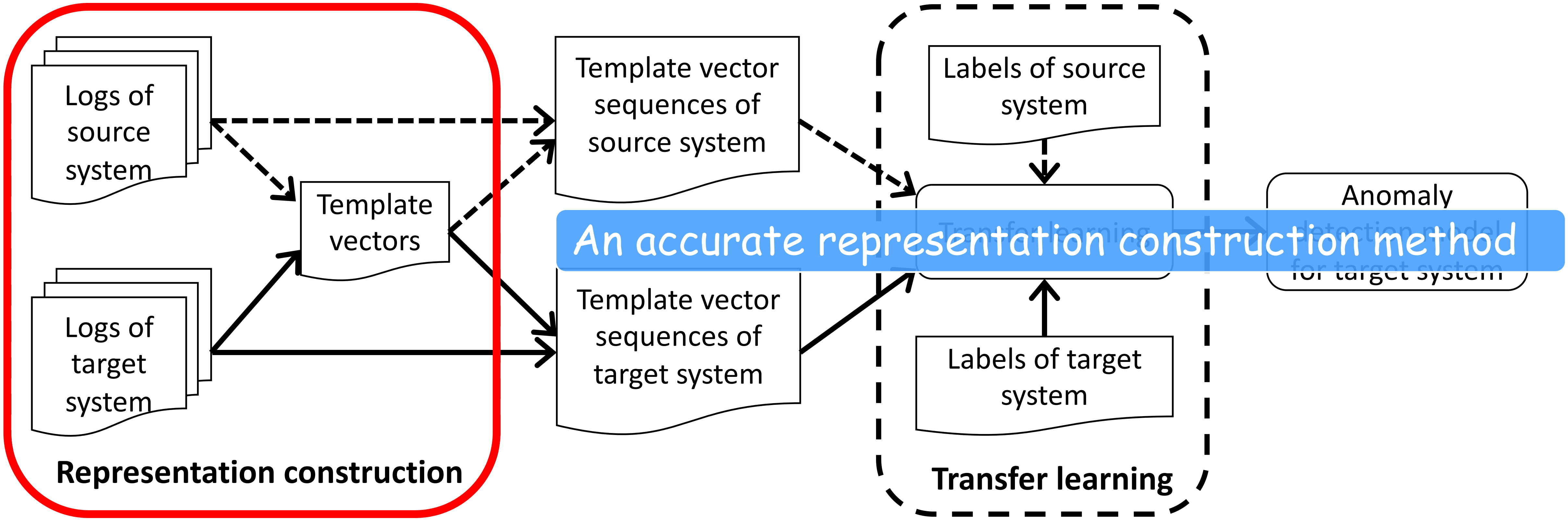
Service Type A	[SIF pica_sif]Interface te-1/1/11, changed state to down [SIF pica_sif]Interface te-1/1/11, changed state to up [OSPF]Neighbour(rid:, addr:) on vlan20, changed state from Init to ExStart [OSPF]Neighbour(rid:, addr:) on vlan20, changed state from ExStart to Exchange [OSPF]Neighbour(rid:, addr:) on vlan20, changed state from Exchange to Loading [OSPF]Neighbour(rid:, addr:) on vlan20, changed state from Loading to Full [OSPF]Neighbour(rid:, addr:) on vlan20, changed state from Full to Down [SIF]Vlan-interface vlan20, changed state to down [SIF]Vlan-interface vlan20, changed state to up
Service Type B	%%10IFNET/3/LINK_UPDOWN(l): GigabitEthernet1/0/10 link status is DOWN. %%10IFNET/3/LINK_UPDOWN(l): GigabitEthernet1/0/10 link status is UP. %%10OSPF/3/OSPF_NBR_CHG(l): OSPF 1 Neighbor (Vlan-interface20) from Loading to Full. %%10OSPF/3/OSPF_NBR_CHG(l): OSPF 1 Neighbor (Vlan-interface20) from Full to ExStart. %%10OSPF/3/OSPF_NBR_CHG(l): OSPF 1 Neighbor (Vlan-interface20) from Full to Down. %%10OSPF/3/OSPF_NBR_CHG(l): OSPF 1 Neighbor (Vlan-interface20) from Full to Init. %%10IFNET/3/LINK_UPDOWN(l): Vlan-interface20 link status is DOWN. %%10IFNET/3/LINK_UPDOWN(l): Vlan-interface20 link status is UP.

Challenges

Noises in anomalous log sequences

Service Type A	*** logined the switch
	*** logouted from the switch
	PICALIBCOMM pica_login Fan is plugged in
	PICALIBCOMM pica_login RPSU is plugged in serial number ***
	Redundancy power supply unit RPSU is plugged in serial number ***
	Receive SFP_PRE message module plugged into port ***
Service Type B	10SHELL SHELL_LOGINFAIL TELNET user *** failed to log in from ***
	10DEVM POWER REMOVED Trap cPowerRemoved power ID is ***
	10SHELL LOGIN Trap cLogIn *** login from ***
	10SHELL LOGOUT Trap cLogOut *** logout from ***
	10SHELL SHELL_CMD Task IPAddr User *** Command is ***
	10LLDP LLDP_CREATE_NEIGHBOR New neighbor created on Port *** ID is ***

Framework of LogTransfer



An accurate representation construction method

Offline model training process

Representation construction

- Building the template embedding:

Containing semantic and syntactic information of a log entry.

Raw system log:
[SIF pica_sif] Interface te-1/1/11, changed state to down

Log template (with FT-tree):
Interface *, changed state to down

Word embeddings (with Glove):
Interface: $V_1 \rightarrow (v_{11}, v_{12}, v_{13}, \dots, v_{1n})$
changed: $V_2 \rightarrow (v_{21}, v_{22}, v_{23}, \dots, v_{2n})$
state: $V_3 \rightarrow (v_{31}, v_{32}, v_{33}, \dots, v_{3n})$
to: $V_4 \rightarrow (v_{41}, v_{42}, v_{43}, \dots, v_{4n})$
down: $V_5 \rightarrow (v_{51}, v_{52}, v_{53}, \dots, v_{5n})$

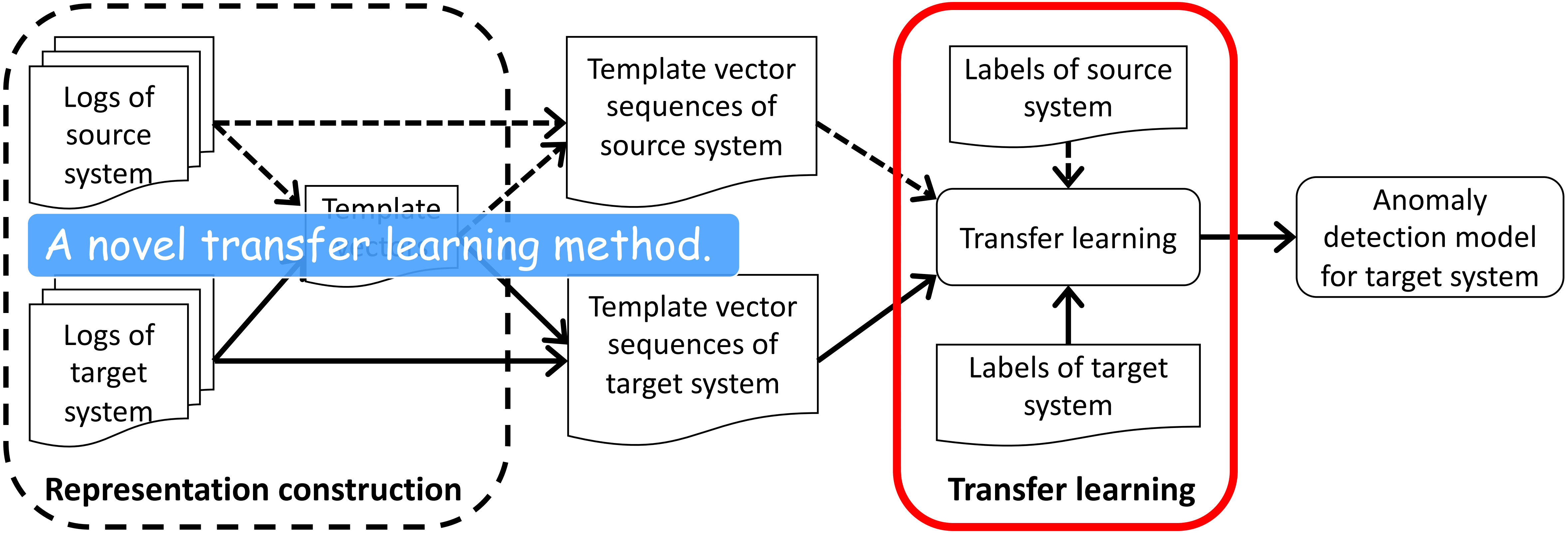
Template embedding:
Interface *, changed state to down:
t $\rightarrow$ average($V_1 + V_2 + V_3 + V_4 + V_5$)

the objective function of Glove:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

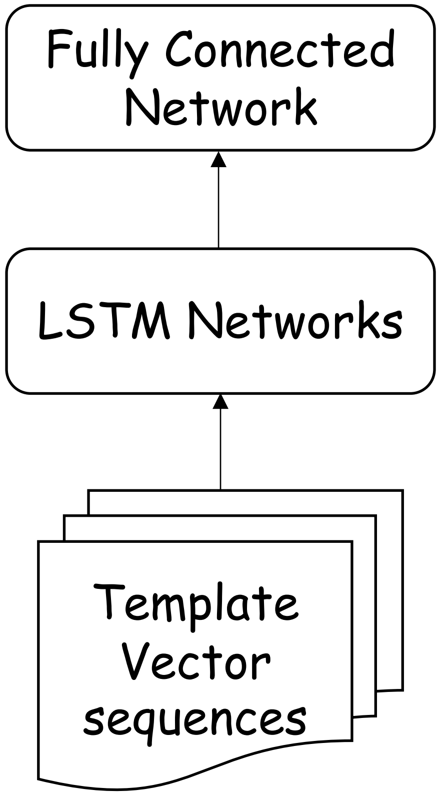
- Example of how to generate a template embedding.

Transfer learning



A novel transfer learning method.

Transfer learning

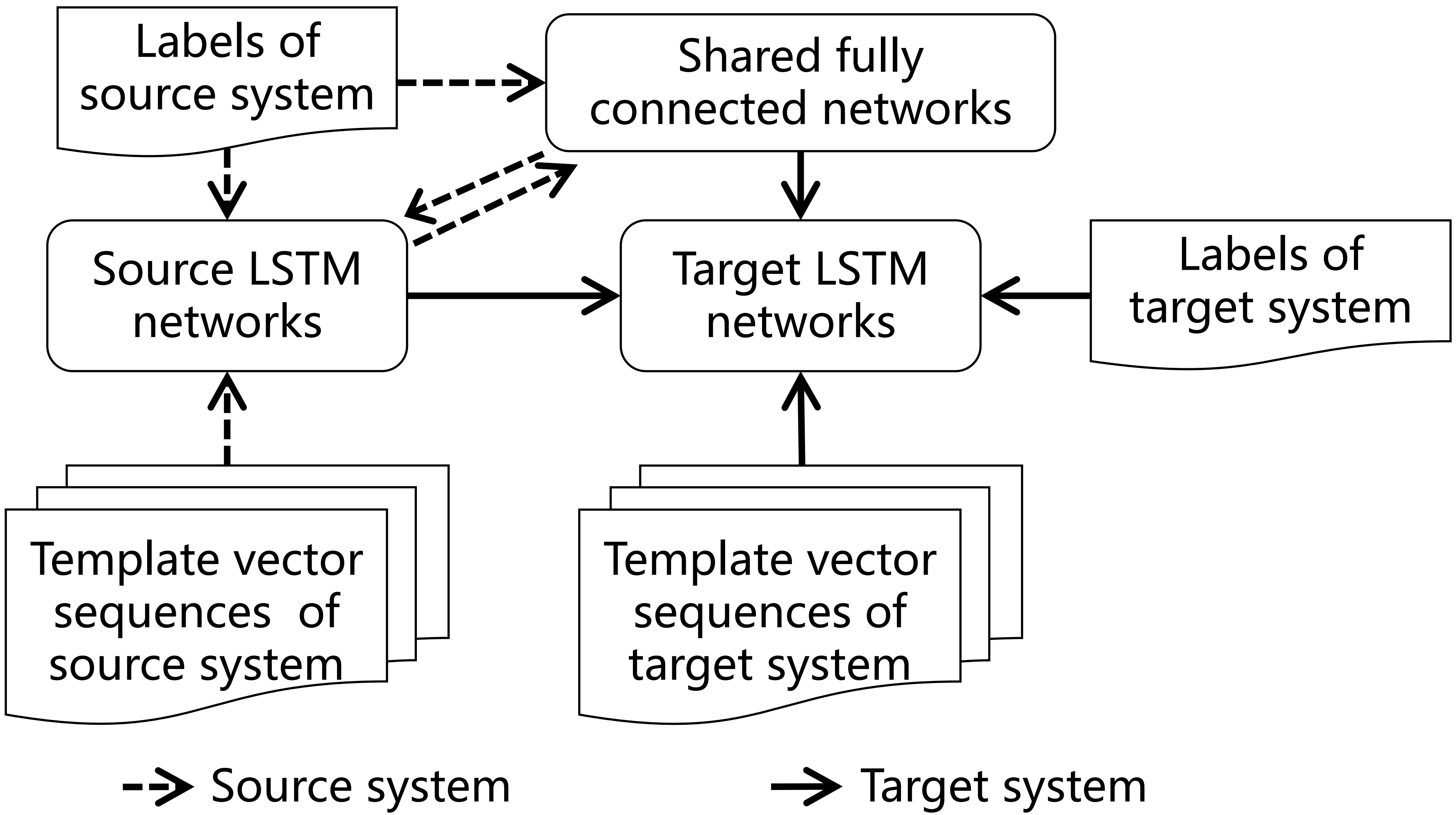


Fully Connected Network for anomaly detection

LSTM networks to extract the pattern of log sequences

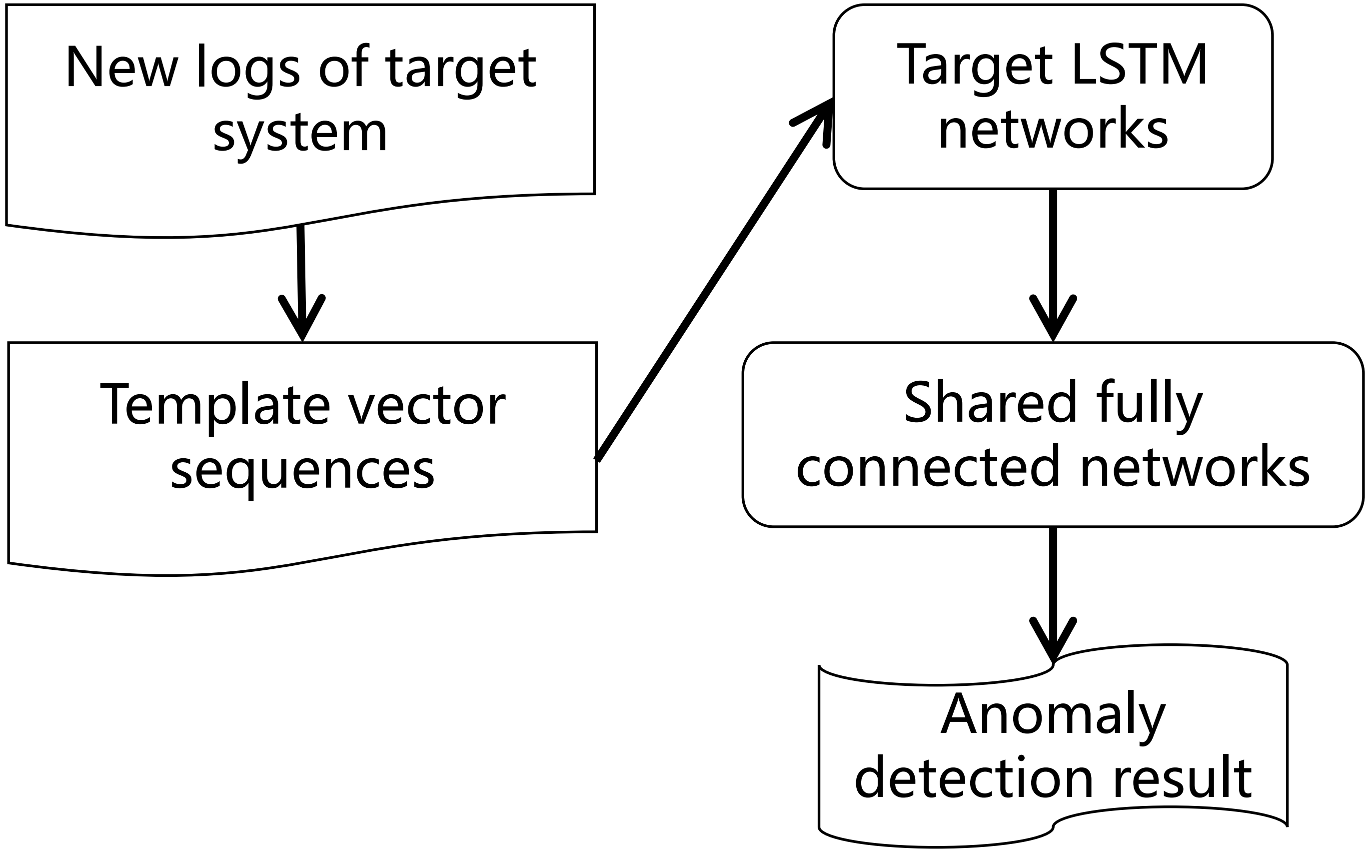
Template vector sequences with syntactic and semantic info.

Transfer learning



Framework of LogTransfer

Online anomaly detection process



Dataset introduction

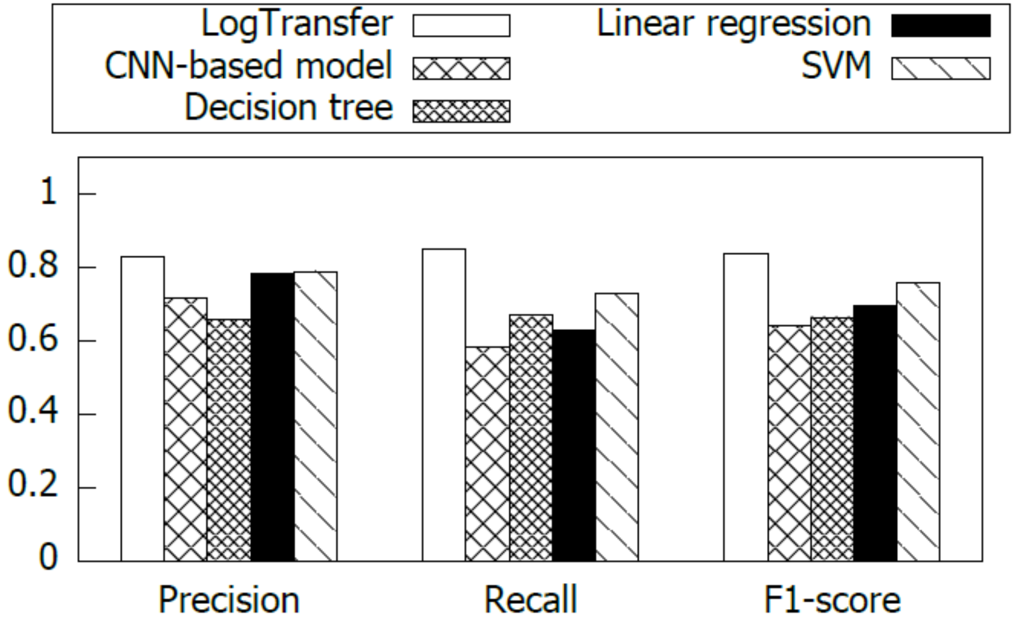
Three switch system log from a top cloud service provider

The Hadoop application dataset & the HDFS dataset

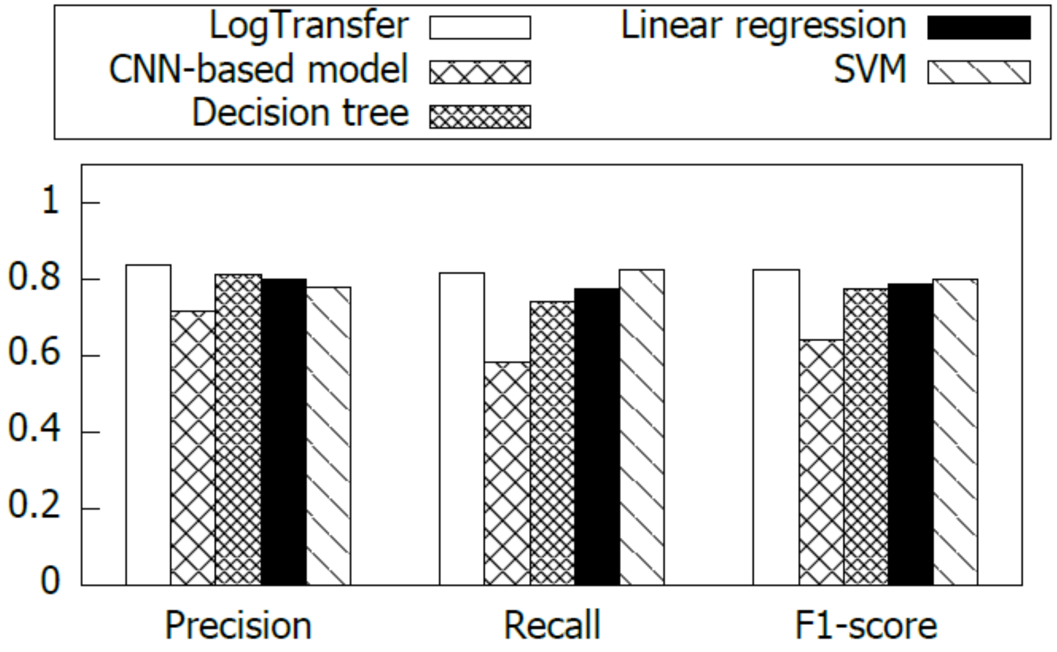
Source of dataset	Type of system	Source of labels	# chunks	# anomalous chunks	# switches/log files
Switch	Type A	Real-world anomalies	2,345,646	6,406	22
	Type B		49,946	1,096	14
	Type C		525,427	4,939	21
Hadoop application	PageRank & WordCount	Manually injected	121,878	73,936	1008
Hadoop file system	HDFS	Real-world anomalies	3,725,203	108,024	575,061

Overall Performance

Evaluation results on switch logs dataset with supervised learning-based model



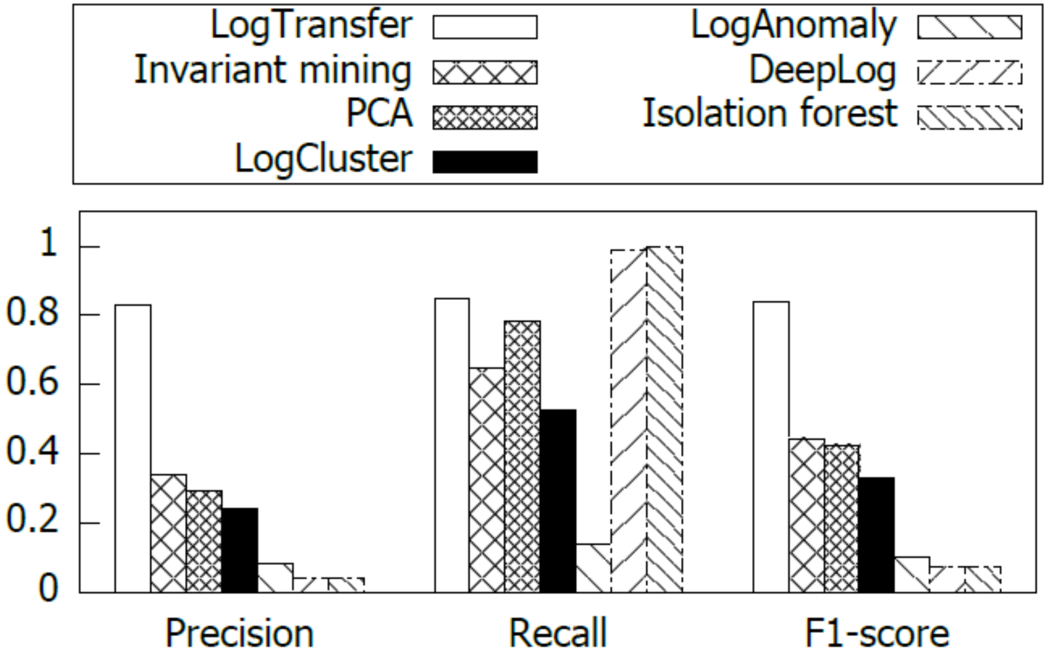
Switch log A -> Switch log B accuracy comparison



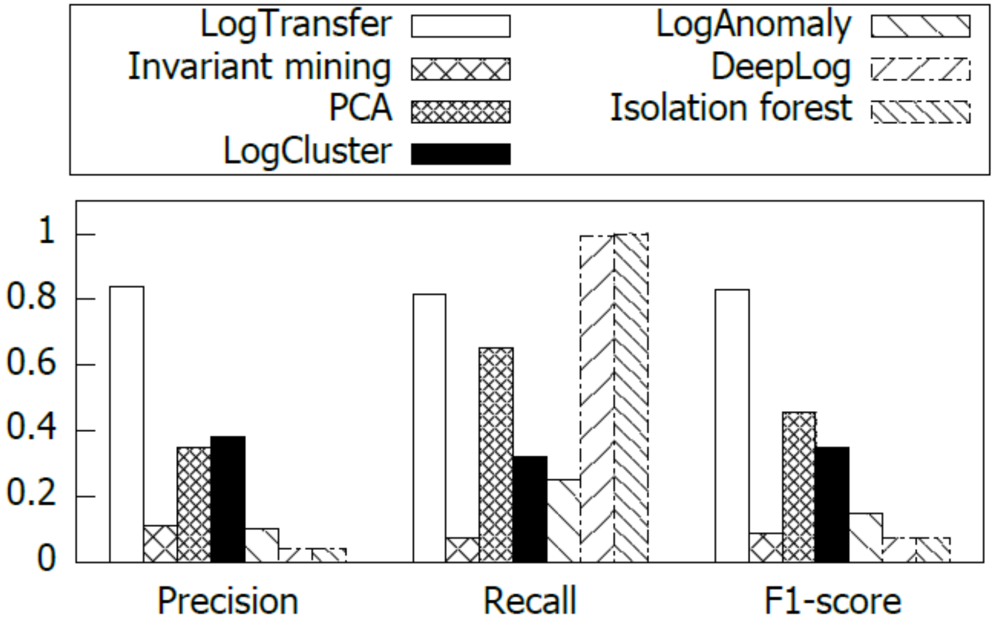
Switch log C -> Switch log B accuracy comparison

Overall Performance

Evaluation results on switch logs dataset with unsupervised learning-based model



Switch log A -> Switch log B accuracy comparison



Switch log C -> Switch log B accuracy comparison

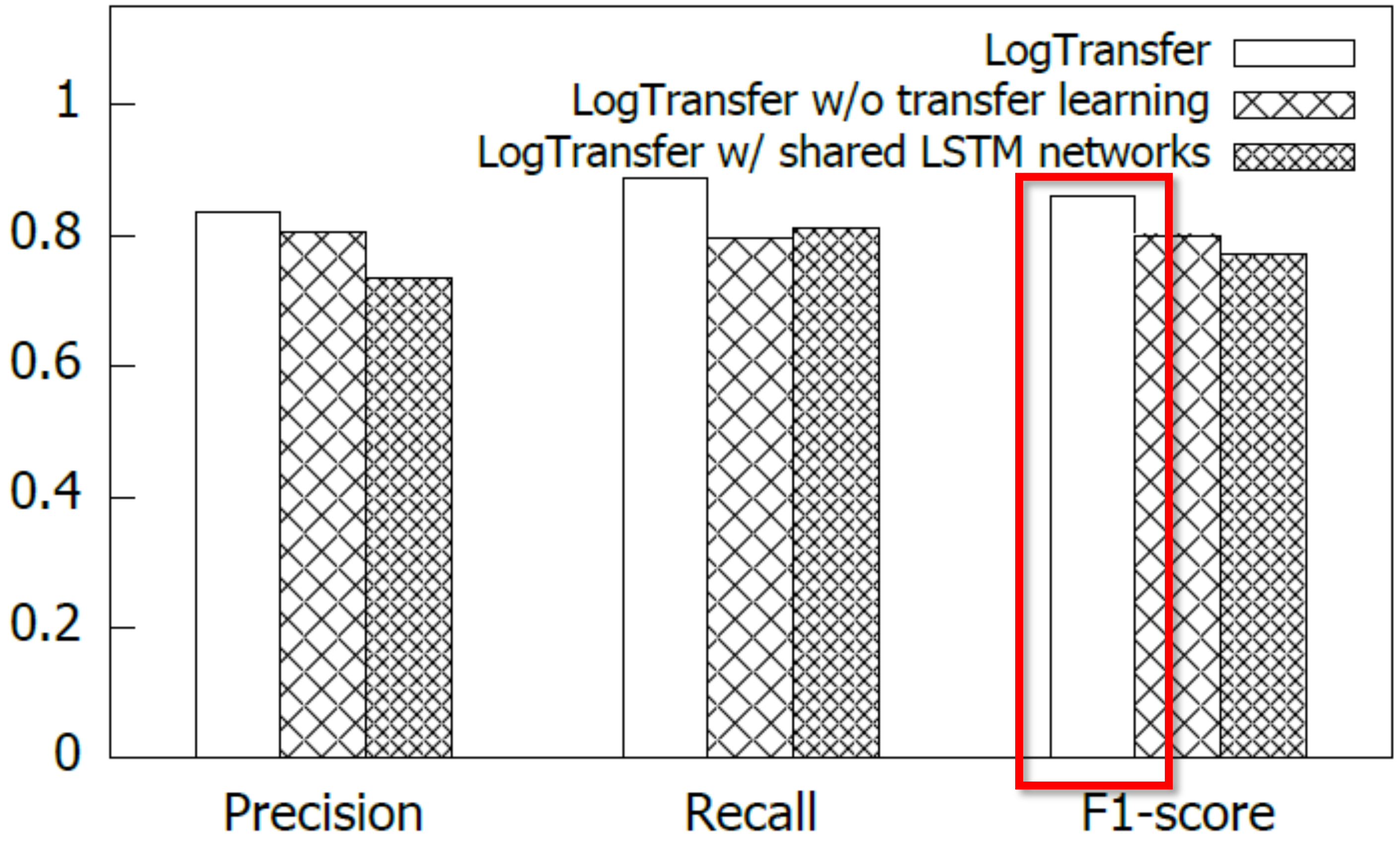
Word embedding method evaluation

Comparison between LogTransfer with word2Vec & LogTransfer

Method	F1-score	AUC score	#False positive	#False negative
LogTransfer w/ word2Vec	0.8368	0.8881	88	84
LogTransfer	0.8606	0.9243	80	59

Transfer learning evaluation

Evaluation of the transfer learning method



Contributions

A new anomaly detection method

- We are the first to apply transfer learning for log anomaly detection
- We identify the challenges lying in that for a large software service.

A new log representation method

- We propose to use Glove to construct logs' representations to accurately measure the similarities of cross-system logs.

A new transfer learning approach

- We propose a novel transfer learning approach that shares fully connected networks between source and target systems, addressing the impact induced by the noises in log sequences.

An extensive evaluation

- We have conducted extensive evaluation experiments using real-world logs to demonstrate LogTransfer's performance.

Thanks

Q&A