论文汇报

《Duplicate Bug Report Detection- How Far Are We?》 2022 TOSEM

TING ZHANG, Singapore Management University, Singapore
DONGGYUN HAN, Royal Holloway, University of London, United Kingdom
VENKATESH VINAYAKARAO, Chennai Mathematical Institute, India
IVANA CLAIRINE IRSAN, Singapore Management University, Singapore
BOWEN XU*, Singapore Management University, Singapore
FERDIAN THUNG, Singapore Management University, Singapore
DAVID LO, Singapore Management University, Singapore
LINGXIAO JIANG, Singapore Management University, Singapore

Ting Zhang
Singapore Management University
在 phdcs.smu.edu.sg 的电子邮件经过验证 - 首页
Software Engineering

2023.4.20

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

Motivation

Despite the many research works and practitioners' adoption of DBRD, unfortunately, it is unclear which DBRD technique can recommend the duplicate BR most accurately overall.

01

Create a benchmark that addresses the limitations of existing evaluation datasets

The most recent work by Rodrigues et al. [43] shows that SABD [43] outperforms REP [50] and Siamese Pair [18]. However, their experiments are only limited to a collection of old BRs from Bugzilla ITSs, in which the latest data used belongs to the year 2008.

02

Compare research tools on the same dataset

Concurrent with the work by Rodrigues et al. [43], Xiao et al. [57] and He et al. [27] have proposed other DBRD solutions. They have not been compared to each other. Besides, they did not compare with the tools used in practice.

03

Compare research and industrial tools

Motivation

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

Age Bias

State Bias

ITS Bias

RQ1: *How significant are the potential biases on the evaluation of DBRD techniques?*

RQ2: *How do state-of-the-art DBRD research tools perform on recent data from diverse ITSs?*

RQ3: *How do the DBRD approaches proposed in research literature compare to those used in practice?*

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

Table 2. Comparison between different approaches

Approach	Type	Feature Engineering		Distance Measurement
		Embedding	Modeling	
REP [50]	Categorical	-	handcrafted	linear combination
	Textual	-	$BM25F_{ext}$	
Siamese Pair [18]	Categorical	customized	single-layer	Cosine Similarity
	Textual	GloVe	bi-LSTM + CNN	
SABD [43]	Categorical	customized	ReLU	fully-connected layer
	Textual	GloVe	bi-LSTM + attention	fully-connected layer
HINDBR [57]	Categorical	HIN2vec	MLP	Manhattan Distance
	Textual	Word2vec	RNN	
DC-CNN [27]	Categorical	Word2vec	dual-channel CNN	Cosine Similarity
	Textual			

Table 3. Textual and categorical fields that are leveraged by the approaches

Fields		REP [50]	Siamese-Pair [18]	SABD [43]	HINDBR [57]	DC-CNN [27]
Textual	summary	✓	✓	✓	✓	✓
	description	✓	✓	✓	✓	✓
Categorical	product	✓	✓	✓	✓	✓
	component	✓	✓	✓	✓	✓
	priority	✓	✓	✓	✓	
	severity		✓	✓	✓	
	type	✓				
	version	✓			✓	

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

Dataset

Table 9. Statistics of training and testing data

ITS	Project	Train		Test	Total
		# BRs (% Dup)	# Dup Pairs	# BRs (% Dup)	# BRs (% Dup)
Bugzilla	Eclipse	19,607 (4.7%)	1,725	7,976 (6.5%)	27,583 (5.2%)
	Mozilla	137,886 (10.1%)	35,474	55,701 (11.2%)	193,587 (10.4%)
Jira	Hadoop	10,276 (2.8%)	328	3,740 (2.5%)	14,016 (2.7%)
	Spark	6,738 (4%)	414	2,841 (3%)	9,579 (3.7%)
GitHub	Kibana	9,849 (2.9%)	376	7,167 (2.6%)	17,016 (2.8%)
	VSCode	40,801 (7.2%)	9,008	21,291 (6.8%)	62,092 (7%)

Workflow

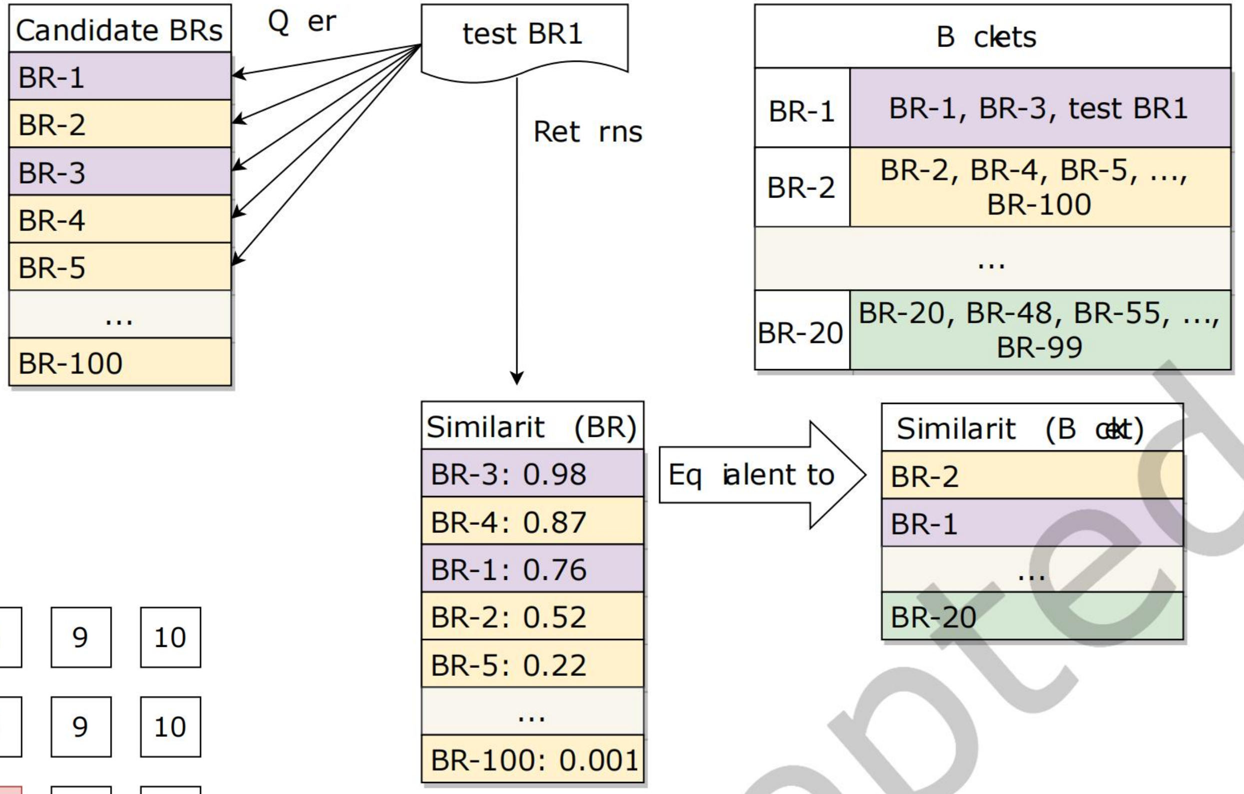


Fig. 4. The workflow of retrieving the correct bucket.

Metric

$$RR@k = \frac{n_k}{m},$$

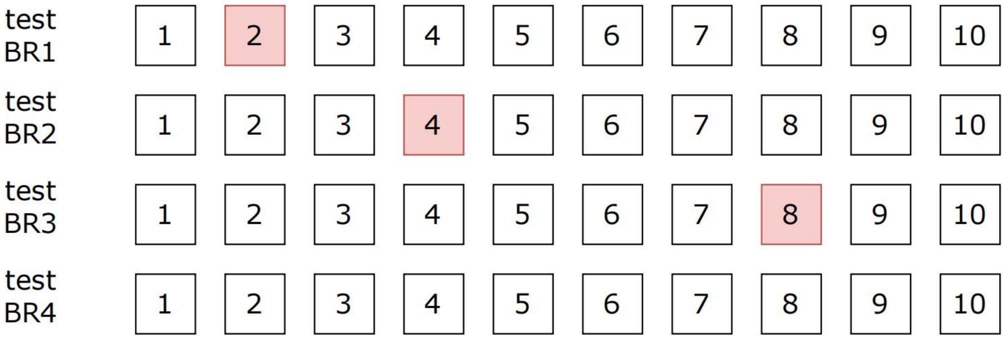


Fig. 3. Examples of the predictions in the top-10 positions for 4 test BRs.

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

RQ1: *How significant are the potential biases on the evaluation of DBRD techniques?*

Age Bias

State Bias

ITS Bias

Table 5. Statistics of old (2012–2014) and recent (2018–2020) data for RQ1

Project	Age	Train		Test	Total	
		# BRs (% Dup)	# Dup Pairs	# BRs (% Dup)	# BRs (% Dup)	# Master BRs
Mozilla	Old	198,653 (9.9%)	35,474	139,502 (9.9%)	338,155 (9.9%)	21,554
	Recent	137,886 (10.1%)	60,498	55,701 (11.2%)	193,587 (10.4%)	10,702
Eclipse	Old	49,355 (5.5%)	4,482	25,021 (12.1%)	74,376 (7.7%)	3,254
	Recent	19,607 (4.7%)	1,725	7,976 (6.5%)	27,583 (5.2%)	959

Table 6. The percentage of BRs changed the corresponding state in 2018–2020

Platform	Summary	Description	Product	Component	Priority	Severity	Version
Eclipse	10.8%	-	7.8%	11.7%	1.2%	5.6%	8.6%
Mozilla	11.8%	-	21.4%	24.5%	24.5%	5.4%	4.2%

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

RQ1: *How significant are the potential biases on the evaluation of DBRD techniques?*

0.05 / 6 = 0.0083

d ∈ [-1, 1], 0表示两个样本没有差异

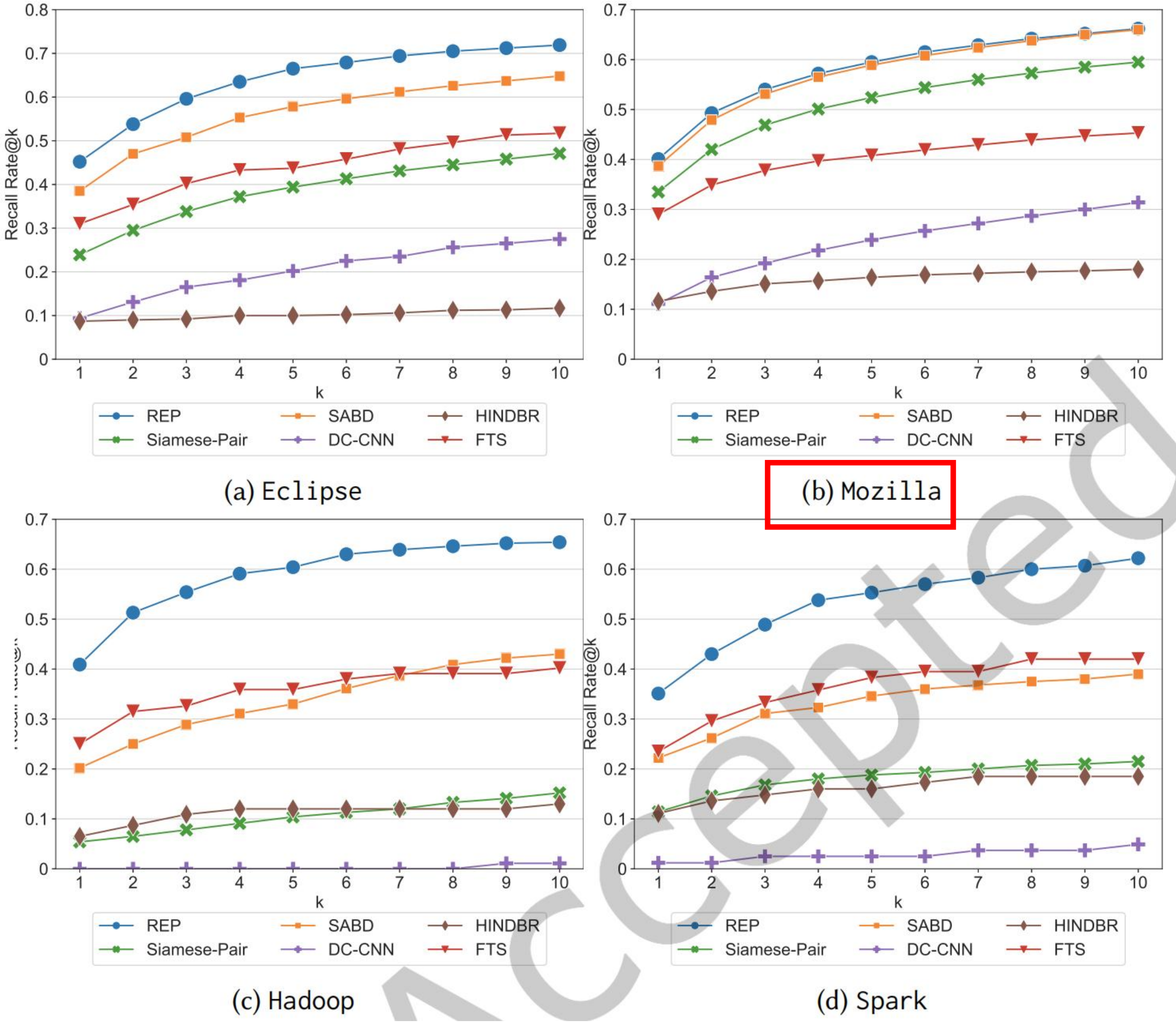
Table 8. Mann-Whitney-U with Cliff's Delta Effect Size $|d|$ on RQ1

Bias	**Approach**	**Data**	***p*-value**	$|d|$
Age	REP	Eclipse	0.003	0.78 (large)
		Mozilla	0.005	0.72 (large)
	Siamese Pair	Eclipse	< 0.001	1 (large)
		Mozilla	0.003	0.76 (large)
	SABD	Eclipse	0.001	0.82 (large)
		Mozilla	0.012	0.66 (large)
State	REP	Eclipse	0.105	0.44 (medium)
		Mozilla	0.190	0.36 (medium)
	Siamese Pair	Eclipse	0.063	0.5 (large)
		Mozilla	0.190	0.36 (medium)
	SABD	Eclipse	0.315	0.28 (small)
		Mozilla	0.315	0.28 (small)
ITS	REP	Jira	0.056	0.36 (medium)
		GitHub	< 0.001	0.66 (large)
	Siamese Pair	Jira	< 0.001	1 (large)
		GitHub	< 0.001	0.97 (large)
	SABD	Jira	< 0.001	0.97 (large)
		GitHub	< 0.001	0.77 (large)

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

RQ2: *How do state-of-the-art DBRD research tools perform on recent data from diverse ITSs?*



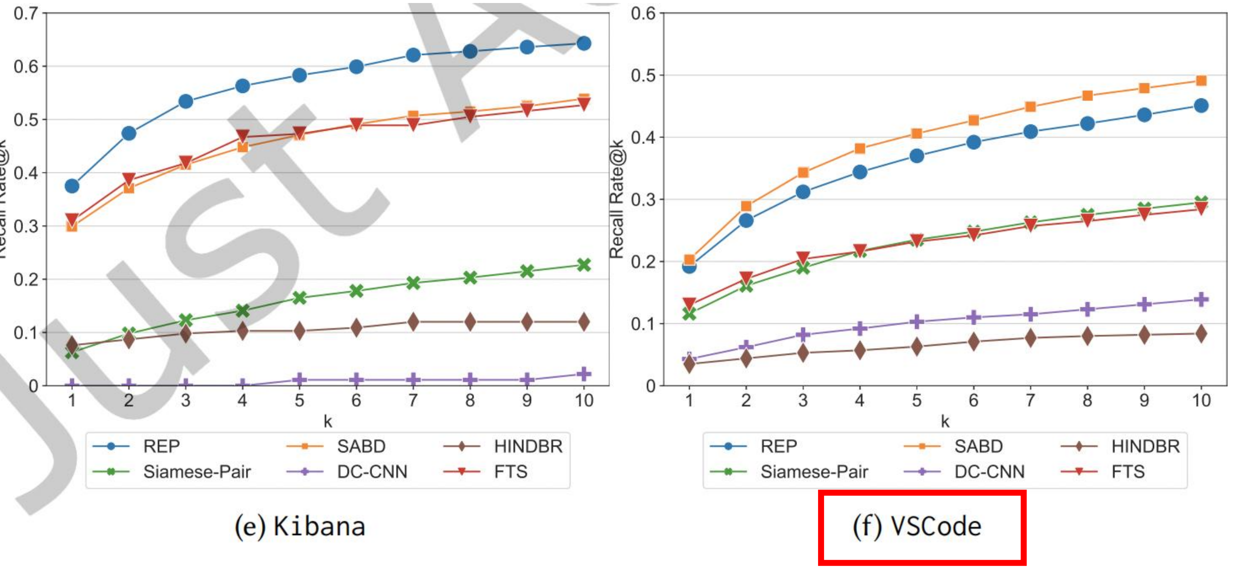
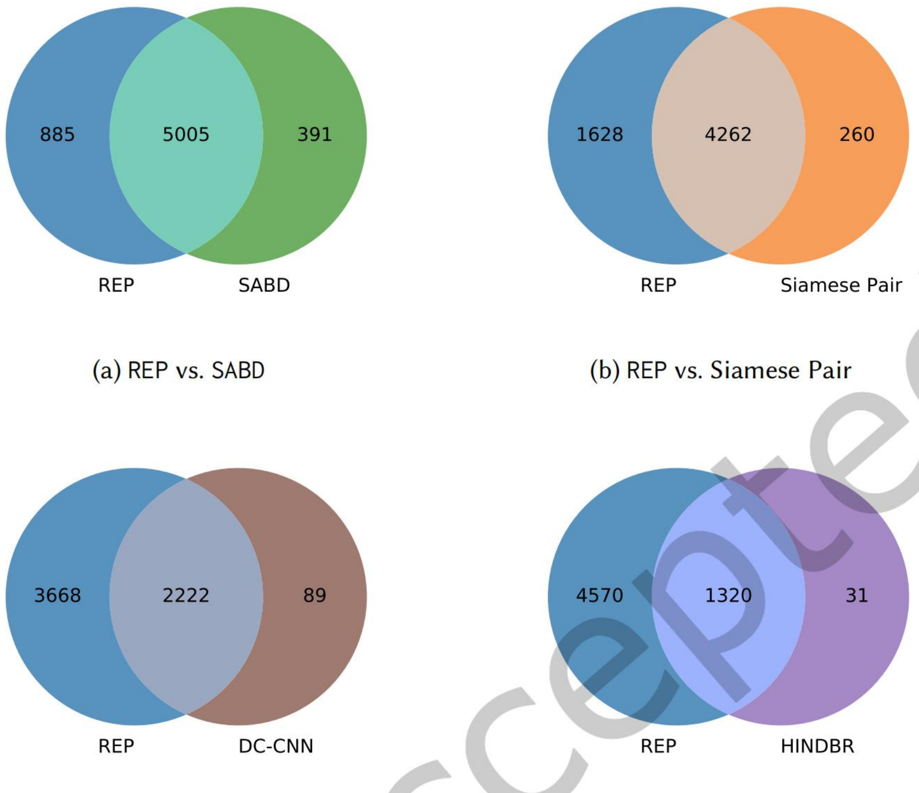


Fig. 5. Recall Rate@k in the test data of Eclipse, Mozilla, Hadoop, Spark, Kibana, and VSCode

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

RQ2: *How do state-of-the-art DBRD research tools perform on recent data from diverse ITSs?*



(a) REP vs. SABD

(b) REP vs. Siamese Pair

(c) REP vs. DC-CNN

(d) REP vs. HINDBR

Fig. 6. REP compared to the other four approaches in terms of successful predictions

Table 10. Investigation of which component benefits REP

RR@k	All	w/o				
		description	short_desc	product	component	priority
1	0.460	0.327	0.350	0.456	0.458	0.450
2	0.544	0.415	0.458	0.527	0.554	0.540
3	0.610	0.456	0.494	0.575	0.598	0.602
4	0.646	0.490	0.510	0.617	0.633	0.637
5	0.673	0.515	0.533	0.644	0.658	0.665
6	0.690	0.531	0.552	0.662	0.663	0.679
7	0.704	0.544	0.569	0.671	0.671	0.694
8	0.706	0.556	0.579	0.681	0.683	0.706
9	0.715	0.565	0.600	0.687	0.683	0.712
10	0.721	0.573	0.613	0.698	0.692	0.717

Duplicate Bug Report Detection- How Far Are We?

2022 TOSEM

RQ3: *How do the DBRD approaches proposed in research literature compare to those used in practice?*

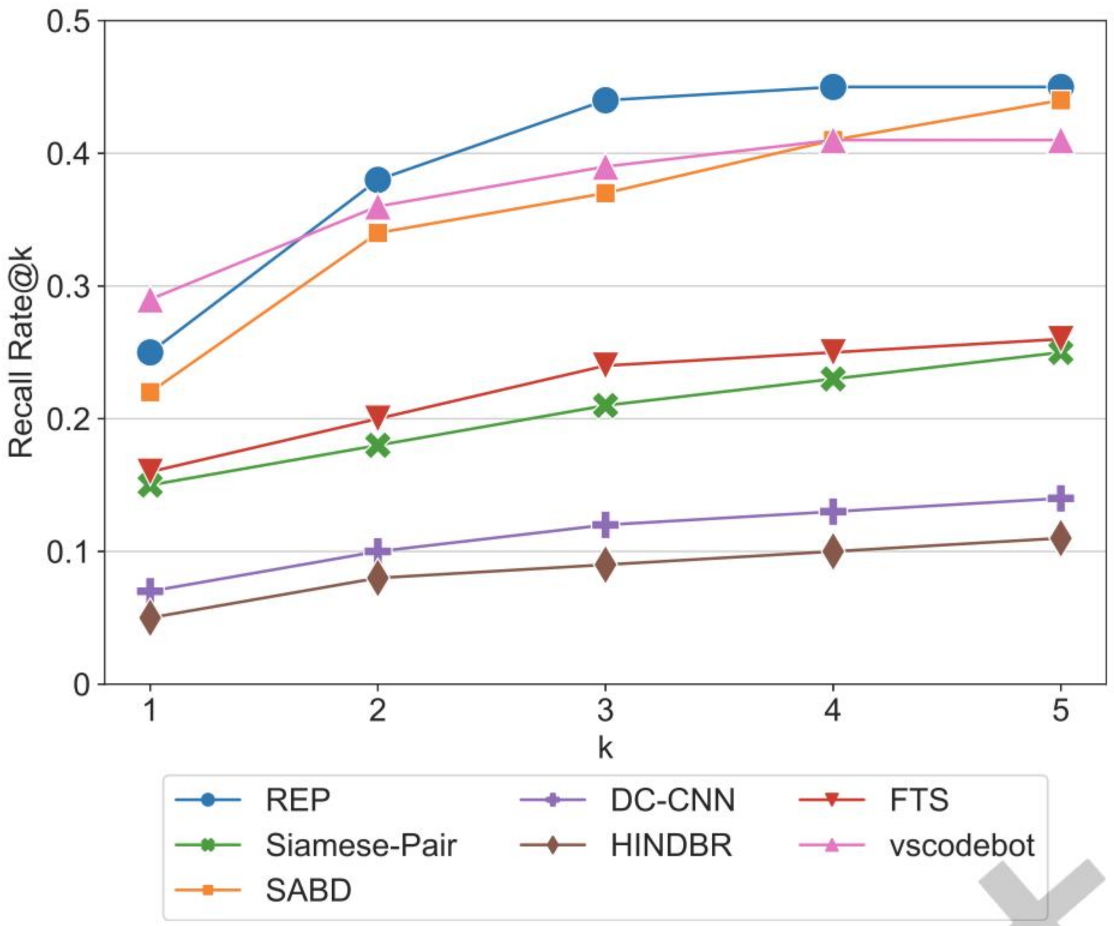


Fig. 7. Recall Rate@k comparing the tools in research and in practice on the VSCode data

谢谢聆听

演讲完毕