Cutting Away the Confusion From Crowdtesting

Junjie Wang[1,3], Mingyang Li[1,3], Song Wang[4], Tim Menzies[5], Qing Wang[1,2,3,*]

[1]*Laboratory for Internet Software Technologies,* [2]*State Key Laboratory of Computer Science,*
Institute of Software Chinese Academy of Sciences, Beijing, China
[3]*University of Chinese Academy of Sciences, Beijing, China,* **Corresponding author*
[4]*Electrical and Computer Engineering, University of Waterloo, Canada*
[5]*Department of Computer Science, North Carolina State University, Raleigh, NC, USA*
Email: {wangjunjie,limingyang,wq}@itechs.iscas.ac.cn, song.wang@uwaterloo.ca, tim@menzies.us

This file is the appendix of the paper "Cutting Away the Confusion From Crowdtesting". It provides some results and analysis which could not be included in the paper due to word limit.

1. Research Question

- **RQ4 Alternative**: Can other combination manner of screenshot similarity and textual similarity outperform SETU in detecting duplicate crowdtesting reports?

The motivating examples indicate that the role of screenshots is mainly to demonstrate the context-related information, while the role of textual descriptions is to provide detailed illustration of the reported problem. This is why our proposed approach SETU first uses screenshot similarity to filter the reports to first class and conducts the ranking separately.

However, one may still argue that other combination manners could achieve better performance. RQ4 aims at investigating whether other combination manners could outperform SETU in duplicate detection. We design three new combination manners to investigate this RQ.

2. Experimental Setup

For answering RQ4, we design three new combination manners, i.e., *addCmb*, *multiplyCmb*, and *textFirst*, to investigate whether other combination manner could achieve a better performance. In detail, *addCmb* denotes adding screenshot similarity and textual similarity as one similarity value and ranking the reports based on it, which is a straight-forward manner. *multiplyCmb* denotes multiplying the screenshot similarity with textual similarity as one similarity value and ranking the reports based on it, which is borrowed from [1]. *textFirst* denotes treating the reports with high textual similarity as the first class and ranking them with the screenshot similarity (the second class is treated as SETU does).

3. Results and Analysis

Figure 1 presents the *recall@1*, *recall@5*, *recall@10*, *MAP*, and *MRR* for each experimental project for SETU and other combination manners (i.e., *addCmb*, *multiplyCmb*, *textFirst*.

We can observe that, in most circumstances, performance obtained by SETU is better than or equal with the performance obtained by other combination manners. In very rare cases, the performance obtained by other combination is a little better than the performance of SETU.

We further conduct Mann-Whitney U Test for the five metrics between SETU and *addCmb*, *multiplyCmb*, *textFirst*. Table 2 shows the p-value, the Cliffs delta, and the interpretation of these tests. There are totally 180 tests (12 projects, 3 alternative combination manners, 5 evaluation metrics). Among them, in 73% (131/180) tests, the performance obtained by SETU is significantly (i.e., p-value is less than 0.05) and substantially (i.e., Cliffs delta is not negligible, i.e., small in 30% tests, median in 15% tests, or large in 28% tests) better than the performance of other combination manners. In other 27% (49/180) tests, the performance obtained by SETU demonstrates negligible difference (i.e., Cliff's delta is negligible although some p-value is less than 0.05) with the performance of other combination manners. In none of the 180 tests, the performance of alternative combination manners is significantly and substantially better than the performance of SETU. This further indicates the effectiveness of our designed combination of screenshot similarity and textual similarity for detecting the duplicates.

Among the three alternative combination manners, we can easily see that duplicate detection performance with *multiplyCmb* is the worst. This combination manner is motivated by the state-of-the-art bug report duplicate detection approach [1]. The bad performance in crowdtesting report duplicate detection implies that a method applied successfully in one application field might not be suitable for another application field.

When using *textFirst* for duplicate detection, we noticed that, in some cases, *recall@1* can be a little higher than our proposed approach SETU, e.g., in P3 to P6. This might because, for the filtered reports in the first class, *textFirst* would rank them based on the image similarity, and in this case, reports with the same screenshots can be identified and ranked ahead to make a larger *recall@1*. But for SETU, the reports in the first class is ranked based on the textual similarity, in which case we can hardly find the reports with exact the same textual descriptions. Nevertheless, for other evaluation metrics, *textFirst* does not achieve as high performance as *SETU*. This might because, as illustrated in

Table 1: **Performance comparison between SETU and other combination manners (RQ4)**

	SETU	**add Cmb**	**mult. Cmb**	**text First**	**Improvement**	**SETU**	**add Cmb**	**mult. Cmb**	**text First**	**Improvement**	**SETU**	**add Cmb**	**mult. Cmb**	**text First**	**Improvement**
	P1					**P2**					**P3**				
recall@1	**0.792**	0.760	0.528	0.768	*3% - 50%*	**0.649**	0.610	0.548	0.524	*6% - 23%*	**0.715**	0.668	0.542	0.726	*-1% - 31%*
recall@5	**0.872**	0.872	0.632	0.850	*0% - 37%*	**0.836**	0.816	0.806	0.790	*2% - 5%*	**0.894**	0.866	0.657	0.844	*3% - 36%*
recall@10	**0.944**	0.936	0.672	0.902	*0% - 40%*	**0.875**	0.864	0.844	0.834	*1% - 4%*	**0.915**	0.890	0.678	0.875	*2% - 34%*
MAP	**0.280**	0.262	0.234	0.258	*6% - 19%*	**0.570**	0.500	0.463	0.496	*13% - 23%*	**0.452**	0.428	0.406	0.436	*3% - 11%*
MRR	**0.831**	0.818	0.584	0.800	*1% - 42%*	**0.736**	0.719	0.693	0.674	*2% - 9%*	**0.794**	0.749	0.600	0.760	*4% - 32%*
	P4					**P5**					**P6**				
recall@1	**0.729**	0.709	0.708	0.730	*0% - 2%*	**0.553**	0.500	0.255	0.536	*3% - 116%*	**0.722**	0.682	0.702	0.718	*0% - 5%*
recall@5	**0.927**	0.906	0.895	0.906	*2% - 3%*	**0.815**	0.779	0.619	0.759	*4% - 31%*	**0.871**	0.851	0.803	0.855	*1% - 8%*
recall@10	**0.958**	0.937	0.937	0.917	*2% - 4%*	**0.922**	0.904	0.648	0.886	*1% - 42%*	**0.915**	0.911	0.867	0.896	*0% - 5%*
MAP	**0.584**	0.550	0.563	0.541	*3% - 7%*	**0.288**	0.262	0.184	0.259	*9% - 56%*	**0.543**	0.521	0.450	0.518	*4% - 20%*
MRR	**0.815**	0.769	0.792	0.783	*2% - 5%*	**0.663**	0.624	0.390	0.641	*3% - 70%*	**0.792**	0.742	0.756	0.787	*0% - 6%*
	P7					**P8**					**P9**				
recall@1	**0.542**	0.380	0.240	0.450	*20% - 125%*	**0.647**	0.613	0.370	0.624	*3% - 74%*	**0.549**	0.487	0.128	0.530	*3% - 328%*
recall@5	**0.666**	0.665	0.395	0.605	*0% - 68%*	**0.849**	0.811	0.501	0.806	*4% - 69%*	**0.768**	0.715	0.334	0.691	*7% - 129%*
recall@10	**0.693**	0.706	0.467	0.686	*-1% - 48%*	**0.877**	0.849	0.526	0.832	*3% - 66%*	**0.811**	0.765	0.416	0.731	*6% - 94%*
MAP	**0.259**	0.181	0.119	0.189	*37% - 117%*	**0.321**	0.298	0.199	0.292	*7% - 61%*	**0.307**	0.268	0.095	0.244	*14% - 223%*
MRR	**0.565**	0.502	0.352	0.512	*10% - 60%*	**0.738**	0.704	0.436	0.715	*3% - 69%*	**0.644**	0.561	0.218	0.598	*7% - 195%*
	P10					**P11**					**P12**				
recall@1	**0.440**	0.377	0.378	0.465	*-5% - 16%*	**0.773**	0.715	0.542	0.779	*0% - 42%*	**0.719**	0.619	0.571	0.729	*-1% - 25%*
recall@5	**0.695**	0.626	0.632	0.626	*9% - 11%*	**0.887**	0.873	0.626	0.796	*1% - 41%*	**0.927**	0.867	0.707	0.820	*6% - 31%*
recall@10	**0.726**	0.691	0.670	0.651	*5% - 11%*	**0.924**	0.912	0.653	0.842	*1% - 41%*	**0.948**	0.918	0.808	0.898	*3% - 17%*
MAP	**0.219**	0.185	0.188	0.214	*2% - 18%*	**0.450**	0.423	0.279	0.348	*6% - 61%*	**0.564**	0.504	0.364	0.514	*9% - 54%*
MRR	**0.552**	0.494	0.484	0.528	*4% - 14%*	**0.828**	0.737	0.585	0.775	*6% - 41%*	**0.805**	0.705	0.705	0.735	*9% - 14%*

Table 2: **Results of Mann-Whitney U Test between SETU and other combination manners (RQ4)**

	SETU vs. addCmb	**SETU vs. multiplyCmb**	**SETU vs. textFirst**	**SETU vs. addCmb**	**SETU vs. multiplyCmb**	**SETU vs. textFirst**	**SETU vs. addCmb**	**SETU vs. multiplyCmb**	**SETU vs. textFirst**
	P1			**P2**			**P3**		
recall@1	0.00 (0.18 S)	0.00 (0.95 L)	0.00 (0.27 S)	0.00 (0.24 S)	0.00 (0.64 L)	0.00 (0.71 L)	0.00 (0.28 S)	0.00 (0.82 L)	0.99 (-0.1 N)
recall@5	0.59 (-0.0 N)	0.00 (0.94 L)	0.15 (0.07 N)	0.00 (0.13 N)	0.00 (0.13 N)	0.00 (0.25 S)	0.00 (0.14 N)	0.00 (0.85 L)	0.00 (0.19 S)
recall@10	0.84 (-0.0 N)	0.00 (0.90 L)	0.17 (0.06 N)	0.13 (0.06 N)	0.00 (0.13 N)	0.00 (0.25 S)	0.03 (0.10 N)	0.00 (0.92 L)	0.00 (0.19 S)
MAP	0.39 (0.02 N)	0.00 (0.21 S)	0.03 (0.13 N)	0.00 (0.50 L)	0.00 (0.67 L)	0.00 (0.48 L)	0.03 (0.11 N)	0.00 (0.37 M)	0.02 (0.11 N)
MRR	0.01 (0.15 S)	0.00 (0.88 L)	0.00 (0.25 S)	0.01 (0.11 N)	0.00 (0.14 N)	0.00 (0.37 M)	0.00 (0.27 S)	0.00 (0.81 L)	0.00 (0.24 S)
	P4			**P5**			**P6**		
recall@1	0.03 (0.14 N)	0.11 (0.09 N)	0.71 (-0.0 N)	0.00 (0.33 M)	0.00 (0.99 L)	0.00 (0.18 S)	0.00 (0.20 S)	0.10 (0.06 N)	0.60 (-0.0 N)
recall@5	0.11 (0.09 N)	0.00 (0.25 S)	0.02 (0.16 S)	0.00 (0.20 S)	0.00 (0.80 L)	0.00 (0.23 S)	0.46 (0.00 N)	0.00 (0.39 M)	0.09 (0.06 N)
recall@10	0.05 (0.13 N)	0.23 (0.05 N)	0.00 (0.21 S)	0.00 (0.19 S)	0.00 (0.95 L)	0.00 (0.17 S)	0.02 (0.09 N)	0.00 (0.30 S)	0.00 (0.13 N)
MAP	0.10 (0.10 N)	0.08 (0.11 N)	0.00 (0.23 S)	0.00 (0.26 S)	0.00 (0.80 L)	0.00 (0.36 M)	0.00 (0.12 N)	0.00 (0.53 L)	0.00 (0.16 S)
MRR	0.00 (0.28 S)	0.08 (0.11 N)	0.01 (0.18 S)	0.00 (0.25 S)	0.00 (0.95 L)	0.14 (0.06 N)	0.00 (0.25 S)	0.00 (0.20 S)	0.13 (0.05 N)
	P7			**P8**			**P9**		
recall@1	0.00 (0.84 L)	0.00 (0.99 L)	0.00 (0.54 L)	0.00 (0.22 S)	0.00 (0.98 L)	0.00 (0.13 N)	0.00 (0.39 M)	0.00 (1.0 L)	0.00 (0.13 N)
recall@5	0.18 (0.10 N)	0.00 (0.98 L)	0.00 (0.34 M)	0.00 (0.23 S)	0.00 (0.97 L)	0.00 (0.21 S)	0.00 (0.27 S)	0.00 (0.99 L)	0.00 (0.38 M)
recall@10	0.70 (-0.0 N)	0.00 (0.87 L)	0.79 (-0.0 N)	0.00 (0.13 N)	0.00 (0.96 L)	0.00 (0.21 S)	0.00 (0.21 S)	0.00 (0.99 L)	0.00 (0.38 M)
MAP	0.00 (0.70 L)	0.00 (0.95 L)	0.00 (0.62 L)	0.00 (0.26 S)	0.00 (0.86 L)	0.00 (0.27 S)	0.00 (0.26 S)	0.00 (0.99 L)	0.00 (0.45 M)
MRR	0.00 (0.50 L)	0.00 (0.90 L)	0.00 (0.43 M)	0.00 (0.18 S)	0.00 (0.98 L)	0.00 (0.15 S)	0.00 (0.42 M)	0.00 (0.99 L)	0.00 (0.23 S)
	P10			**P11**			**P12**		
recall@1	0.00 (0.50 L)	0.00 (0.42 M)	0.99 (-0.1 N)	0.00 (0.33 M)	0.00 (0.89 L)	0.43 (0.00 N)	0.00 (0.54 L)	0.00 (0.63 L)	0.98 (-0.1 N)
recall@5	0.00 (0.35 M)	0.00 (0.32 S)	0.00 (0.40 M)	0.65 (-0.0 N)	0.00 (0.93 L)	0.00 (0.45 M)	0.00 (0.21 S)	0.00 (0.77 L)	0.00 (0.38 M)
recall@10	0.00 (0.18 S)	0.00 (0.29 S)	0.00 (0.43 M)	0.37 (0.02 N)	0.00 (0.96 L)	0.00 (0.46 M)	0.00 (0.12 N)	0.00 (0.67 L)	0.00 (0.30 S)
MAP	0.00 (0.18 S)	0.00 (0.24 S)	0.53 (-0.0 N)	0.00 (0.20 S)	0.00 (0.85 L)	0.00 (0.66 L)	0.00 (0.43 M)	0.00 (0.89 L)	0.00 (0.29 S)
MRR	0.00 (0.30 S)	0.00 (0.33 M)	0.00 (0.13 N)	0.00 (0.37 M)	0.00 (0.89 L)	0.00 (0.19 S)	0.00 (0.46 M)	0.00 (0.47 M)	0.00 (0.35 M)

Note that: The figures X (Y Z) respectively denotes p-value, Cliffs delta, and interpretation of Cliffs delta (i.e., Large (L), Medium (M), Small (S), and Negligible (N))

Motivation Section, the role of screenshots is to demonstrate the context-based information, and the role of textual descriptions is to help accurately identify the reported problem. Therefore, SETU is designed to filter the first class based on reports' image similarity, and the comparison with other combination manners also proves its effectiveness.

The combination manner proposed in SETU can achieve relatively highest performance than other alternative combination manners.

4. Reference

[1] X. Yang, D. Lo, X. Xia, L. Bao, J. Sun, Combining word embedding with information retrieval to recommend similar bug reports, in: ISSRE'16, pp. 127–137.