A THE CALCULATION OF FILE PATH SIMILARITY FEATURE

Algorithm 1 Calculate the value of the file path similarity feature

INPUT:
PR_{new}: The new PR.
RE: A reviewer in the project
OUTPUT: file path similarity feature of RE to PR_{new}: FpSim(PR_{new}, RE)
Method:

1: historyReviews = A list of previous PRs reviewed by RE before PR_{new} submission
2: **for** Review $PR_{prev} \in$ historyReviews **do**
3: Sim = 0
4: $FilePaths_{prev}$ = getFilePaths(PR_p)
5: $FilePaths_{new}$ = getFilePaths(PR_{new})
6: # Compute file path similarity between PR_{new} and PR_{prev}
7: **for** $fp_{new} \in FilePaths_{new}$ **do**
8: **for** $fp_{prev} \in FilePaths_{prev}$ **do**
9: Sim = Sim + $\frac{prefixCommon(fp_{new}, fp_{prev})}{Max(Length(fp_{new}), Lenght(fp_{prev}))}$
10: **end for**
11: **end for**
12: Scores[PR_{prev}] = Sim
13: **end for**
14: FpSim(PR_{new}, RE) = aggregate(Scores)
15: **return** FpSim(PR_{new}, RE)

Algorithm 1 shows the detailed steps to compute the file path similarity feature denoted as $FpSim(PR_{new}, RE)$. It takes a new PR (i.e., PR_{new}) and an active reviewer (i.e., RE) as input. First, the algorithm searches the previous PRs that are reviewed by the reviewer before the submission time of the new PR (i.e., PR_{new}) (Line 1). For each previous PR (i.e., PR_{prev}) reviewed by the reviewer, the algorithm computes the file path similarity between the previous PR (i.e., PR_{prev}) and the new PR (i.e., PR_{new}) (Lines 2 to 13). The algorithm retrieves the file paths (i.e., $FilePaths_{prev}$) of the previous PR (i.e., PR_{prev}) (Line 4) and the file paths (i.e., $FilePaths_{new}$) of the new PR (i.e., PR_{new}) (Line 5). Next, For each file path (i.e., fp_{new}) from the new PR (i.e., PR_{new}) and each file path (i.e., fp_{prev}) from the previous PR (i.e., PR_{prev}), the algorithm computes the similarity. Then, the similarity between the previous PR (i.e., PR_{prev}) and the new PR (i.e., PR_{new}) is the sum of the similarities of all possible pairs of file paths (Line 7 to 11). After obtaining the file path similarities between all previous reviewed PRs and the new PR (i.e., PR_{new}), we aggregate the similarities to compute the file path similarity feature for the active reviewer (Lines 14 to 15).

To compute the similarity between two file paths, i.e., a file path (i.e., fp_{new}) of the new PR (i.e., PR_{new}) and a file path (i.e., fp_{prev}) of the previous PR (i.e., PR_{prev}), the algorithm first splits the file paths into components (i.e., directories and file name) using the slash character ('/') as the delimiter. Then, the algorithm computes the longest common prefix components between the two file paths. The longest common prefix is the number of common components that occur in both file paths from the root directory to the file name. Next, the value of the longest common prefix is normalized by the maximum number of components of the file paths (i.e., the longer one of fp_{new} and fp_{prev}). For example, given fp_{new} = "$lib/chef/fs/file\_system/cookbooks\_dir.rb$" and fp_{prev} = "$lib/chef/formatters/doc.rb$". The common directories of the two file paths are "$lib/chef/$", which signifies that the longest common prefix is 2. The number of components in the file paths fp_{new} and fp_{prev} are 5 and 4, respectively. Thus, the maximum number of components is 5. Finally, the similarity between the file paths fp_{new} and fp_{prev} is $\frac{2}{max(5,4)} = 0.4$.