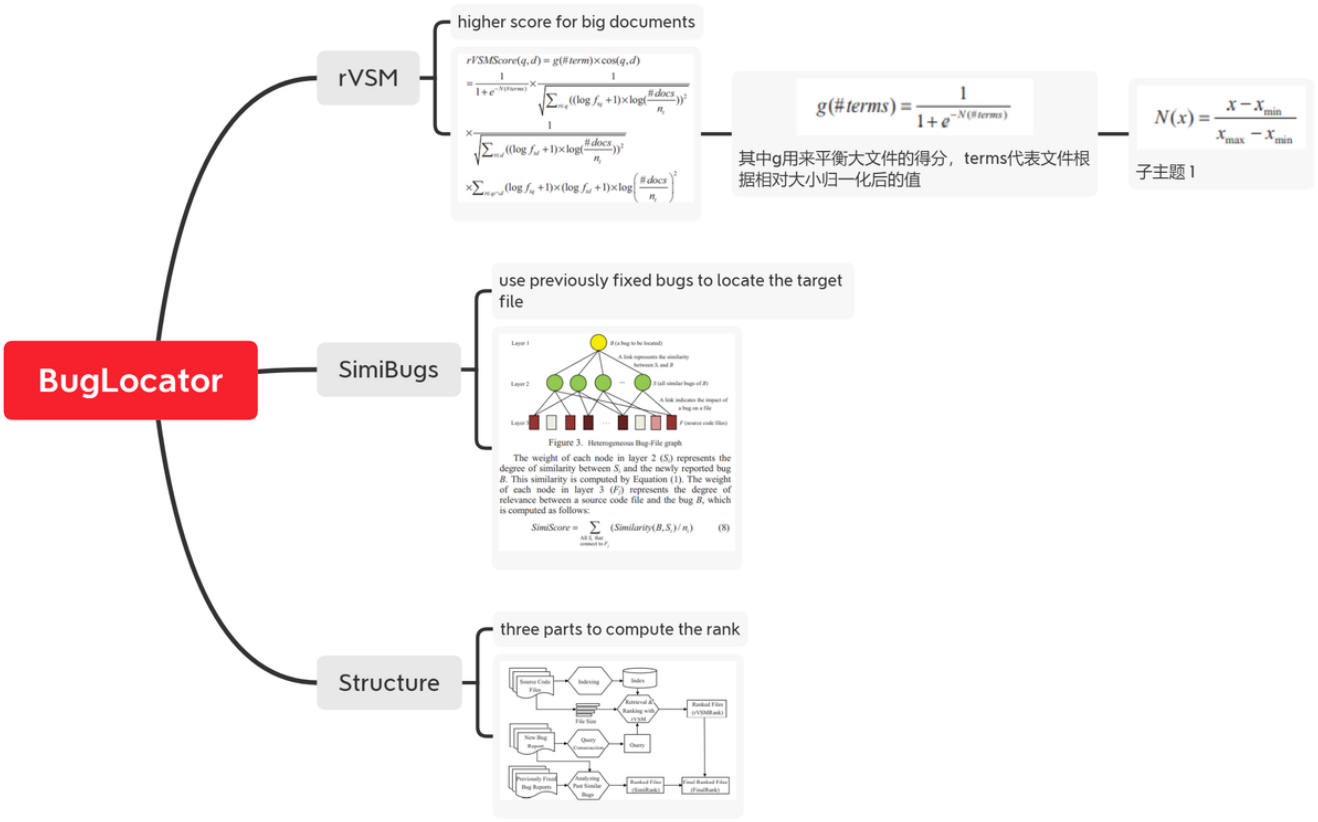
BugLocator阅读报告

作者：韩禧

1. 概述



2. 过程

1. rVSM

由于通过VSM计算的精确度不够，主要是缺少对大文件的支持导致较大的文件权值偏低，因而提出了rVSM。从统计学的角度看，更大的文件意味着更高的发生bug的概率，因此更有可能成为包含缺陷的代码。通过 1/e^-x 函数进行平衡后，大文件将具有更高的排名。

tf-idf的计算过程也有变更，传统计算过程为：

$$tf(t,d)=\frac{f_{td}}{\#terms},idf(t)=\log(\frac{\#docs}{n_t})$$

其中“ #terms represents the total number of terms in document d, and #docs represents the total number of documents in the corpus ”。

新的计算过程为：

$$tf(t,d)=\log(f_{td})+1$$

同时对每个term的权重计算也变为：

$$w_{t\in d}=tf_{td}\times idf_t=(\log f_{td}+1)\times\log(\frac{\#docs}{n_t})$$

$$\left|\overrightarrow{V_d}\right|=\sqrt{\sum_{t\in d}((\log f_{td}+1)\times\log(\frac{\#docs}{n_t}))^2}$$

2. SimiBugs

通过将新的bug报告与近期解决的bug报告进行比对，通过加权得出最有可能与当前bug相关联的文件。这个结构的计算方式类似 于PageRank。

3. 总体结构

最终的结果由rVSM结果与bug报告相似度结果加权得出，

FinalScore = (1-α)*N(rVSMScore)+ α* N(SimiScore)

3. 实验及结果

使用了Eclipse、SWT、AspectJ、ZXing四个开源项目作为测试数据：

TABLE II. THE PERFORMANCE OF BUGLOCATOR

System	α	Top 1	Top 5	Top 10	MRR	MAP
ZXing	0.2	8 (40%)	12 (60%)	14 (70%)	0.50	0.44
SWT	0.2	39 (39.80%)	66 (67.35%)	80 (81.63%)	0.53	0.45
AspectJ	0.3	88 (30.77%)	146 (51.05%)	170 (59.44%)	0.41	0.22
Eclipse	0.3	896 (29.14%)	1653 (53.76%)	1925 (62.60%)	0.41	0.30

并将最后的结果与SUM进行了比对：