Which Similarity Metric to Use for Software Documents?
A study on Information Retrieval based Software Engineering Tasks

Md Masudur Rahman
masud@virginia.edu
Saikat Chakraborty
saikatc@virginia.edu
Baishakhi Ray
rayb@virginia.edu

Different IR-based Software Engineering (SE) tasks operate on different document artifacts.
SE artifacts are heterogeneous in nature and different in characteristics than natural language.
We investigate the impact of IR models on different SE artifacts and analyze how such informed choice could lead to an improved performance.

Representative Software Engineering Task

**Project Recommendation**
Given a project as a query, the task is to find functionally similar projects from GitHub. For example, given a query with a Video Recorder project, the system tries to return a list of Video Recorder projects.

**Bug Localization**
Given a bug report as query, this task ranks all the source files in the project repository based on their relevance with the query. The files that top the ranking are more likely to contain the cause of the bug.

**Dataset:** 1590 GitHub projects
**Text:** Description, Readme
**Code:** Method-Class name, Package name, API

**Dataset:** 200 bug reports of JDT
**Mixture of text and code:** Bug reports, source code

Performance of Different Models on Various SE Artifacts (MAP@10)

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Readme</th>
<th>Code</th>
<th>Method</th>
<th>Class</th>
<th>Package Name</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSM</td>
<td>0.51</td>
<td>0.37</td>
<td></td>
<td>0.37</td>
<td>0.29</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>LSI</td>
<td>0.57</td>
<td>0.39</td>
<td></td>
<td>0.36</td>
<td>0.24</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>BM25</td>
<td>0.51</td>
<td>0.26</td>
<td></td>
<td>0.16</td>
<td>0.007</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>WMD</td>
<td>0.51</td>
<td>0.29</td>
<td></td>
<td>0.25</td>
<td>0.20</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation Metric
❖ Mean Average Precision (MAP)
❖ Mean Reciprocal Rank (MRR)

Can an informed choice of IR model improve the performance of SE tasks?

**Project Recommendation**
We replace the ranking algorithm from LSI to VSM of CLAN[1], a project recommendation tool.
- CLAN leverages code - code document artifacts.
- Our experiments show VSM performs best for such code - code artifacts.

Modified model, vsmCLAN achieves up to 35% improved performance.

**Bug Localization**
We replace the ranking algorithm from VSM to BM25 of LR[2], a bug localization tool.
- LR leverages mixture (text - code) document artifacts.
- Our experiments show BM25 performs best for such mixture artifacts.

Modified model, LR-BM25 achieves 43% improved performance.

References:

Department of Computer Science
University of Virginia