Relating Developers´ Concepts and Artefact Vocabulary in a Financial Software Module

by
Tezcan Dilshener
and
Michel Wermelinger

Agenda

- Introduction
  - Research questions
- Related work
  - Inspirations
- Methodology
  - Building blocks
- Results
  - Correlation
  - Precision and recall
- Discussion
- Conclusion

Motivation

- **One of the main challenges in software maintenance is to**
  - **Accurately identify where and how high-level concepts are implemented in code.**
- **Investigation of a financial application module by comparing the vocabulary of**
  - **Change requests, User guide, Source code**
  - **Elicited domain concepts**
- **Our aim: Role of vocabulary in providing a good leverage during maintenance**
  - **Do identifiers reflect domain concepts?**
  - **Can identifier names used to find relevant classes for implementing a given change request?**

Related Work

- **Vocabulary comparison by Haiduc *et al.***
  - **occurrence of domain concepts in identifier names and source comments.**
  - **we compare domain concepts beyond code and include change requests.**
- **Recovery of traceability by Antoniol *et al.***
  - **from source code classes to functional requirements.**
  - **we attempt to recover between change requests and source code.**

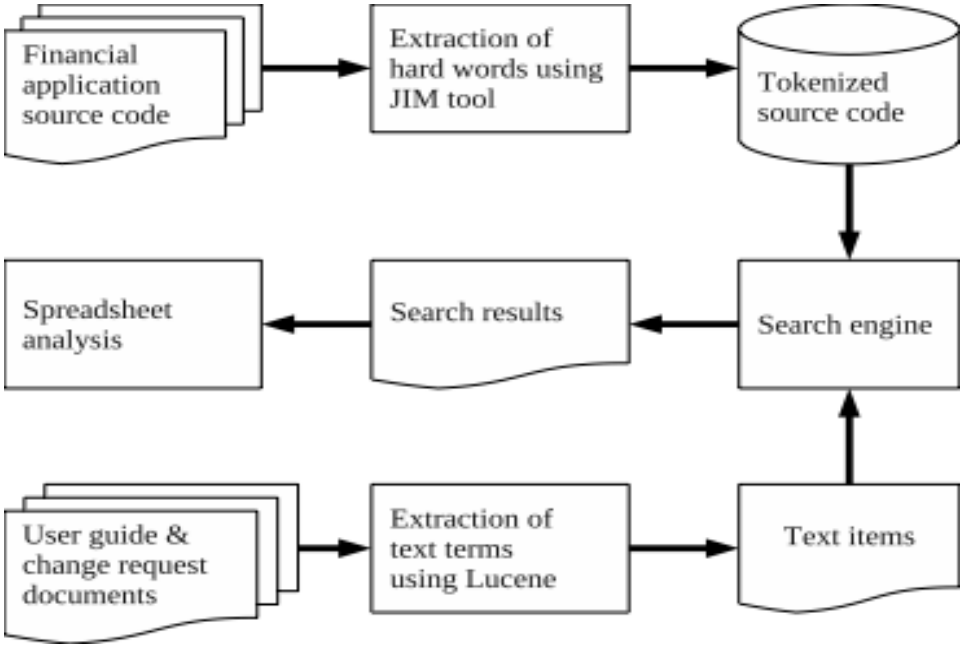
Artefacts

CR	Change Request Description
1088	change the layout of not editable fields in the calculation mask to formatted text.
1090	Allow to edit the market values at the asset level, calculation mask with edit.
2003	Pdlgd export data to an importable excel format.
2010	allow volatility values **greather** than 1.

Identifier names as classes	Declarations in source
K5MarketHelperCalcRiskPfUnit	riskTotal, marketCalcPfUnit
K5MarketProcessorCopyAssets	assetName, copyProperties
K5MarketExportHandler	exportType, exportItems
K5MarketValueKeyVolatility	volaIndexCorr, volaScaling

Methodology

- **Extraction, search and analysis flow**



Results - Domain Concepts

Concept (exact search)	Rank in CR/Defect	Rank in User Guide	Rank in Source Code
market	1	1	2
value	2	7	3
calculation	3	2	14
risk	4	4	4
asset	5	12	8
roundup	6	16	15
diversification	7	15	13
time	8	6	5

Results - Correlation

	Exact Search / Stem Search	**Exact Search / Stem Search**	**Exact Search / Stem Search**
Correlation Between =>	CR & Guide	CR & Code	Guide & Code
Common concepts	17	16	36
Spearman rank correlation	0.32 / 0.52	0.093 / 0.13	0.55 / 0.67
p-value	**0.19 / 0.037**	**0.72 / 0.62**	**0.0016 / 0.0002**

Results - Precision and recall

concepts searched	**recall (%)**	**precision (%)**
calculation, market	100	5.41
calculation, asset, market	90.91	6.76
roundup	0	0.00
pdlgd	0	0.00
volatility, market	100	4.26

CR vocabulary searched (mapping/stop-word)	**recall (%)**	**precision (%)**
calculation, **helper**	100	14.04
calculation, asset, adapter, data, edit, operation, report, version, workflow	54.55	6.12
data, export	100	33.33
pdlgd, data, export	100	13.33
volatility	66.67	20

Discussion - Our first aim

- **Do the identifiers reflect the domain concepts?**
  - **Full business concept coverage.**
    - **All three artefacts include all the domain concepts.**
  - **Only 80% of concepts occur in code and user guide.**
    - **Potential inefficiencies during maintenance.**
  - **Important concepts in user guide also remain in code.**
    - **Good alignment to ease maintenance tasks.**
  - **Weak correlation between CR and other two artifacts.**
    - **Not an issue since CR is specific per unit of work.**

Discussion - Our second aim

- **Can the identifier names used to find the relevant classes for implementing a given change request?**
  - **Using CR vocabulary.**
    - **Achieve very good recall but poor precision.**
  - **Mapping terms to project specific counterparts.**
    - **Improves precision.**
  - **Ignoring frequent concepts by acting as stop-words.**
    - **Drastically reduces false positives.**
  - **Stem search is ineffective when descriptive identifiers used.**
    - **Decreased precision, while not increasing recall.**

Conclusions

- **An efficient approach to relate vocabulary of information sources for maintenance;**
  - **Concepts, change requests, user guide and code.**
- **Application of approach to industrial code that follows good naming conventions.**
  - **Alignment between guide and code could be improved.**
  - **Descriptive identifiers support high recall, but low precision.**
  - **Applied simple techniques and improved precision.**
- **Further research is required.**
  - **Incrementally adding mappings and stop-words.**
  - **Automatic heuristics, like looking for very frequent words.**