

一般的な語句の除去は、2段階で実施する。まず Google Ngram Viewer²に記載された英語の検索語の最頻出語 10000 件を除去する。さらに、既に検索語に含む単語を除去する。

次に、残った単語について、頻出度により検索結果との関連性をスコアリングする。文書に存在する各単語について、最も多く出現した文書における出現数 $\max(tf)$ と文書全体での出現数 df から次式により t_{score} を算出する。

$$t_{score} = \max(tf) \times df$$

ソースコードや設定ファイルの例示などで頻出する単語のように、多くの文書に現れかつ一文書中に頻出する単語については、 t_{score} が大きくなると期待できる。そこで、この t_{score} によって文書中の単語を順序付け、上位 5 件を特徴語としてユーザに提示する。

3.2 フィードバック受付部

ユーザは提示された特徴語に対して、自分の検索意図と関連するか判断し、その結果を画面上で指定する。そしてフィードバック受付部では、関連すると指定された特徴語をユーザからのフィードバックとして受け付ける。

フィードバック受付部は、関連すると指定された特徴語を検索語に追加し再度検索を行うことで、適合フィードバックを実現する。関連すると指定されなかった特徴語については、今回の適合フィードバックのセッションの以降の検索で除外するように、除外リストに加える。

4. 実験

被験者が提案手法を利用して、障害の解決方法を導出できるかを観察することにより評価実験を行った。対象とする障害には、Stack Overflow に投稿され、解決方法が記載されている質問内で言及された障害事象を用いた。利用した質問は次の条件を満たすものである。

- 客観的に有用な解決方法が存在することを担保するため、解決済みかつ解決方法のコメントに対するスコアが 1 票以上のもの。
- 解決のためにはコード読解の必要性があるため、被験者のプログラミングの理解スキルに従い、“Java”または“C”タグが付いたもの。
- 最新更新順に上位 50 件以内に存在するもの。

被験者は情報工学を専攻する学生 10 名である。57 個の障害ケースを作成し、5 分割した上で 2 人ずつの被験者に割り当てた。提案手法を実装し

たシステムを利用し、各障害ケースに対する解決方法を 2 人の被験者に導出させ、その様子を観察した。なお、10 分間で導出できない場合は実験を打ち切った。同一の障害ケースを対象に実験した 2 人の被験者による特徴語利用有無と、解決方法の導出可否を表 1 に示す。

表 1 実験結果

解決方法導出 \ 特徴語利用	2人利用	1人利用	利用無
2人とも可能	7	8	1
1人のみが可能	3	12	5
2人とも不可能	9	7	5
計	19	27	11

全体の 57 件中 46 件 (約 81%) で、どちらかの被験者が特徴語を利用した適合フィードバックを実施していた。このことから、被験者の意図に沿った特徴語を抽出できていたと言える。また、どちらの被験者も特徴語を利用しなかった 11 件では 2 人とも解決方法を導出できたケースが 1 件 (約 9%) だったが、2 人とも特徴語を利用した 19 件では 7 件 (約 37%) に増えた。このことから、特徴語の追加が解決方法の導出に貢献できていることがわかった。

5. 結論

単語の重み付けと適合フィードバックにより障害対処のための Web 検索を支援する手法を提案した。今後の課題として、言語別のソースコード解析による特徴語抽出、フィードバックによって追加した特徴語に対する矛盾する検索語や類似検索語の除去が挙げられる。

謝辞

本研究は、JSPS 科研費 JP17K00110 の助成を受けたものである。

参考文献

- [1] Vaidyanathan, R., Das, S. and Srivastava, N.: Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual, International Journal of Computer Applications, Vol. 105, No. 8, pp.1-6 (2014)
- [2] 廣瀬 賢幸, 鵜林 尚靖, 亀井 靖高, 佐藤 亮介: Stack Overflow を利用した自動バグ修正の検討, ソフトウェア工学の基礎 XXIV (日本ソフトウェア科学会 FOSE2017), pp33-42 (2017)

² <https://github.com/first20hours/google-10000-english>