

Characterizing search activities on stack overflow

Jiakun Liu, Sebastian Baltes, Christoph Treude,
David Lo, Yun Zhang, and Xin Xia,

ESEC/FSE 2021: Proceedings of the 29th ACM Joint Meeting on European
Software Engineering Conference and Symposium on the Foundations of
Software Engineering, August 2021, pp. 919–931

発表者: 井町 瑠貴

背景

Stack Overflow (SO) は質問を投稿する前に
「検索して調べる」ことを推奨している^[1]

しかし、開発者の求める情報と、検索された結果
との間にはギャップがある

[1] How do I ask a good question? <https://stackoverflow.com/help/how-to-ask>

背景

先行研究ではクエリの再定型化^[2,3]や
要約された回答の生成技術が提案されている^[4]

検索ログを用いた大規模な分析を行い、SOでの
開発者の検索方法を特徴づけた研究は少ない

[2] Kaibo Cao, Chunyang Chen, Sebastian Baltes, Christoph Treude, and Xiang Chen. 2021. Automated Query Reformulation for Efficient Search based on Query Logs From Stack Overflow. In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). 1273–1285.

[3] Neng Zhang, Qiao Huang, Xin Xia, Ying Zou, David Lo, and Zhenchang Xing. 2020. Chatbot4QR: Interactive Query Refinement for Technical Question Retrieval. IEEE Transactions on Software Engineering (2020).

[4] Bowen Xu, Zhenchang Xing, Xin Xia, and David Lo. 2017. AnswerBot: Automated generation of answer summary to developers' technical questions. In 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 706–716.

目的

検索ログを用いた大規模な分析を行い, SOでの開発者の検索方法を特徴づける

リサーチクエストション

RQ1: SOでの検索活動はどのようなものか?

RQ2: SOでの検索と一般的な検索とでは、
クエリのリフォーメーションに
どのような違いがあるか?

RQ3: SOでのクエリの性質は?

データセット

2017年12月1日から2018年11月31日までのSOのサーバーへのリクエスト747,421,780件を利用

そのうち、1つのイベントに対して2つ以上のイベントに派生する可能性があるものと、botによるリクエスト、ページ更新リクエストを除外

少なくとも1つの検索イベントを持つ4,981,786件のシーケンスを得た

データセット

各リクエストは以下のイベントタイプに割り当てた

- Search : SOの検索ボックスを使った検索
- Post : 特定の質問や回答をリクエストすること
- Home : SOのホームページをリクエストすること
- Begin : 検索セッションの開始
- End : 検索セッションの終了

RQ1:

SOでの検索活動はどのようなものか？

- ・目的

検索ログデータを用いて、検索活動(連続したイベントのペア)を特徴づける

- ・アプローチ

Aprioriアルゴリズム(頻出アイテム集合の検出)を用いて、異なる検索活動間の関連ルールを分析

RQ1:

SOでの検索活動はどのようなものか?

Table 1: Top 9 most common search activity types on Stack Overflow. We present the types of search activities, the corresponding frequencies, the proportion among all search activities, the examples, and the corresponding description.

Activity	#	%	Predecessor	Target	Description
Search→Search	4,143,291	17.20%	/search?q=java	/search?q=java+linux	reformulate the immediately preceding query
Search→Post	3,338,051	13.86%	/search?q=java	/q/215497	visit a post from search result page
Post→Search	2,817,653	11.70%	/q/215497	/search?q=java	search from the posts
Post→End	2,701,072	11.21%	/q/215497		the search session ends up in post
Begin→Post	2,670,252	11.08%		/q/215497	the search session begins with post
Search→End	2,036,898	8.46%	/search?q=java		the search session ends up with search result page
Home→Search	1,285,261	5.34%	/home	/search?q=java	search from the home page
Begin→Home	1,100,320	5.57%		/home	the search session begins from the home page
Begin→Search	642,797	3.67%		/search?q=java	the search session begins from the search page

336種類の検索アクティビティが確認され、
上位9種類の検索アクティビティがSO上の
検索アクティビティの86.07%を占める

RQ1: SOでの検索活動はどのようなものか?

Table 2: Top 10 frequently co-occurring search activity types related to the beginning and the end of the search sessions.

left-hand-side	right-hand-side	support
Search→Post	Post→End	0.536
Begin→Post	Post→Search	0.488
Begin→Post	Search→Post	0.365
Begin→Post,Search→Post	Post→Search	0.348
Begin→Post,Post→Search	Search→Post	0.348
Begin→Post	Post→Search, Search→Post	0.348
Search→Search	Search→End	0.340
Post→Search	Post→End	0.324
Post→Search,Search→Post	Post→End	0.322
Search→Post	Post→End, Post→Search	0.322

検索セッションの62.73%がPostで終了
Postに至らなかった場合, QuestionAskをリクエスト,
つまりSOに質問を投稿する人は0.01%しかいない

RQ2:

SOでの検索と一般的な検索とでは、クエリのリフォーメーションにどのような違いがあるか?

- ・アプローチ

HuangとEfthimiadisの研究に従う

→一般的な検索エンジンの検索ログから学習し、
再定式化戦略の分類法を作成

RQ2:

SOでの検索と一般的な検索とでは、クエリのリフォーメーションにどのような違いがあるか?

SO上の4,048,518件のリフォーメーションのうち、2,905,998件(71.8%)がHuangとEfthimiadisの研究によるリフォーメーション戦略に当てはまった

一般的な検索で特定された戦略に当てはまらない再定式化については、その26.7%がドメイン固有の特性に関連している

RQ3: SOでのクエリの性質は?

- ・目的
クエリの内容, 検索意図, それらの関係を明らかにする
- ・アプローチ
3人の著者により, 検索セッションを意味的タグと検索意図のカテゴリにそれぞれ分類

RQ3:

SOでのクエリの性質は?

・意味的タグ

23.44%の検索セッションでは、追加条件を持つクエリが1つ以上あり、追加条件を表す形容詞を持っている

Table 6: Categories of semantic tags in queries at the concept level. We also present the description, example, and the proportion of the search sessions that have corresponding semantic tags among the sampled 1086 search sessions.

Category	Description	Example	%
Programming language and technology	The name of the coding technology, e.g., programming language, package name, library name, OS, schema.	python, angular, pyqtgraph	61.98%
Task	The action verb refer to the task to be implemented	jQuery upload file	42.71%
Data type	The data types that can be assigned with values, entities, etc.	different colors in the same plot matplotlib	27.34%
Object	The non-programming terminologies nouns that are predicated by task verbs.	count vowels assembly languages	24.74%
Condition	The adjective refer to conditions to the tasks and the objects.	c# maximim integer value	23.44%
Method	The callable method of a programming technology in language, library, package, etc.	heapq.nlargest,	10.67%
Error information	The bug trace, error code, or description that indicate the code cannot run as expectation.	[c#] vertical input error	9.64%
Development tool	The tool that are used to help with coding process, e.g., testing suite, CI/CD tools.	git commit -a	7.81%
Question word	The question words in English.	why, how, what, where	6.25%
Code	A code snippet.	atoi(argv[1])	1.30%

RQ3: SOでのクエリの性質は?

・検索意図

18.30%は、関数の呼び出しに関するもの
→公式ドキュメントに掲載されている例題が
開発の助けにならないことが考えられる

Table 7: Categories of search intention. We present the description, the top 3 most common combinations of semantic tags, and corresponding examples, as well as the proportion of the sampled 1086 search sessions that are associated with the intention.

Category	Description	Typical combinations of semantic tags	Example	%
Technology manipulation	Search for how to manipulate programming technologies, e.g., assign values, find the max value.	data type	mappolygon	19.33%
		technology name, data type	ILIST ROW	
		technology name, data type, task	delete row dataframe	
Function call	Search for the use example of certain callable method, API, etc.	technology name	Libgdx	18.30%
		method	heapq.nlargest	
		technology name, method	groupby r	
Concrete development task	Search for how to implement a certain development task in certain programming technology. Developers usually identify such task when coding in editor.	task	order by	18.30%
		technology name, task, object	jQuery upload file	
		task, technology name	jquery capitalize	

SOへの実用的な提案

- ユーザーの0.01%しか質問しない理由を調査すること
- WordNetのようなデータベースを構築すること（プログラミング用語間の技術的な関係を含む）
- ユーザーに検索意図をどのようにクエリに表現するかなどのガイドラインを提案すること

結論

- SOのウェブサーバへのHTTPリクエストに基づいて、直前のクエリを再構築することが最も頻繁な活動であることを観察
- 検索セッションの23.44%は、追加条件を表す形容詞を含むクエリを持っている
- 今後、技術的な面で関連のある用語を別の用語に置き換えるなどの最適化方法を提案する予定

所感

- ・自分自身, SOでの検索があまり上手でないので興味深かった
- ・SOで質問を投稿するハードルは高いが, 質問を投稿しやすくなると, 質問の質が低下する可能性もあると感じた