What Do Concurrency Developers Ask About? A Large-scale Study Using Stack Overflow

Syed Ahmed and Mehdi Bagherzadeh

（紹介担当：阿萬@愛媛大）

概要

【目的】

並行プログラミングについて，開発者たちは**どういった話題に興味**を持ったり，**難しさ**を抱えていたりするかを把握

【方法】

Stack Overflowでの質問と回答に対して**トピックモデリング**を行い，それらをカテゴリ分けして傾向を分析

【結果】

質問は大きく**8種類のカテゴリ**に分類された

最もよくある質問はスレッドセーフに関することで，最も（答えを得るのが）**難しい質問は**データベース管理システムに関することであった　等

背景

- **並行プログラミングの難しさ**
  - 並行処理（マルチスレッドや並列処理など）のコーディングでは，どういったトピックが興味深かったり，難しかったりするのか？
- **Stack Overflow** はこれを理解する上で有益な**情報源**になる

Research Questions（RQs）

【RQ1】開発者たちは**どういった話題**について質問しているか？

【RQ2】話題は**どういうカテゴリ**に属しているか？それらの階層はどうなっているか？

【RQ3】開発者の中で**最もポピュラーな話題**は？

【RQ4】答えを見つけるのが**最も難しい話題**は？

【RQ5】**ポピュラーさと難しさは相関**するか？

調査手順

1. Stack Overflow での並行性に関する質問を特定して抽出
2. 抽出した質問（文章）に対してトピックモデリングを行い，質問内容のトピックを手作業で分類
3. 似たトピックを統合していくことでトピックの階層を構築
4. 得られた並列性トピックが先行研究で言われていることと整合しているか調査
5. トピックのポピュラーさと難しさを測定

調査内容(１／３)

- Stack Overflow から 14,995,834 件の質問と 23,489,212 件の回答を収集
  - 各質問には**タグが付けられている**ので，まずは最もよく使われていた 100 種のタグの中から並行性に関するものを抽出
  - 次にそれらのタグが付けられていた質問について，(他にもタグが付いているので)それらを抽出して，タグの候補集合を作成
  - その後，ヒューリスティックに(分類精度に基づく)閾値を決めて**タグ集合を構築**

調査内容(2／3)

- 作成したタグ集合に基づいて，**156,777件の質問**と249,662 件の回答（**accepted answer は 88,764 件**）が得られた
- これらに対して**前処理**を実施
  - ソースコード，HTMLタグ，ストップワード，数字，句読点，非アルファベット，URLを除去
  - ステミング処理（ing 等を削除）

調査内容(３／３)

- LDA(latent Dirichlet allocation)によるトピックモデリングを実行
- トピックのラベルは著者二人によって決められている:決め方はオープンカードソートという方法

ポピュラーさのメトリクス

- そのトピックの質問の**平均閲覧回数**
- そのトピックの質問が**お気に入りとして登録された平均回数**
- そのトピックの質問の**平均スコア**

難しさのメトリクス

- そのトピックの質問のうち，**accept された回答が無いものの割合**
- そのトピックの質問が **accept される回答を得るまでにかかった平均(中央)時間**

結果:RQ1

- 開発者たちの**質問内容は多岐**にわたっていた:
  - スレッドプールから並列計算まで
  - モバイル並行性からWeb並行性まで
  - メモリー貫性から実行速度向上まで
- 最も多く質問されたのは**基本概念に関するもの(８%)**であった

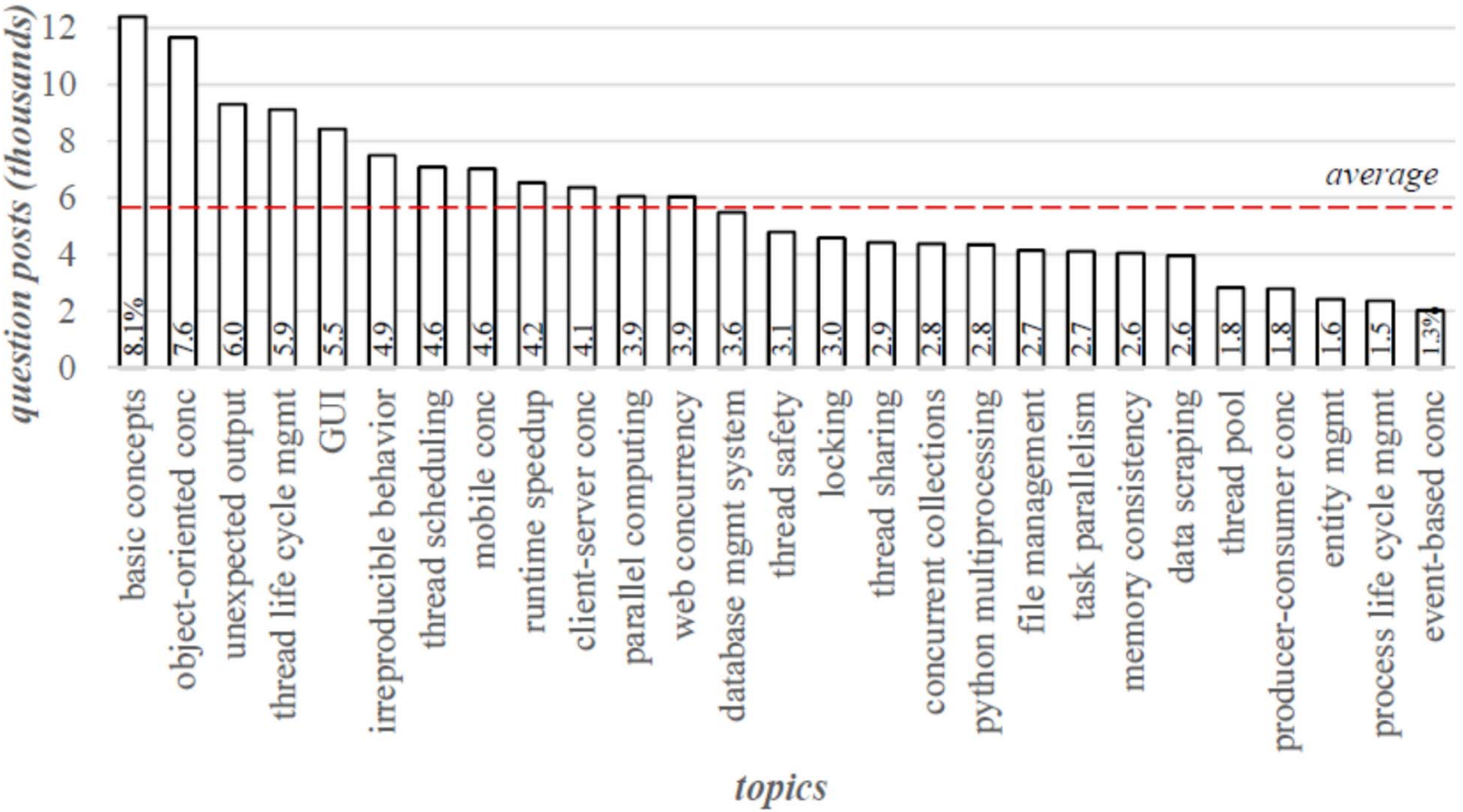

Figure 2: Concurrency topics with individual numbers, average number (dashed line) and percentages of their questions.

結果:RQ2

- 質問は大きく**八つのカテゴリ**に分かれた
  - **並行モデル**，**プログラミングパラダイム**，**正確さ**，**デバッグ**，**基本概念**，**永続性**，**パフォーマンス**，**GUI**
- カテゴリ別では**並行モデル**が最も多く28%，GUIが最も少なく5%
- **パフォーマンスよりも正確性**の方が質問されることが多かった（7% vs. 12%）

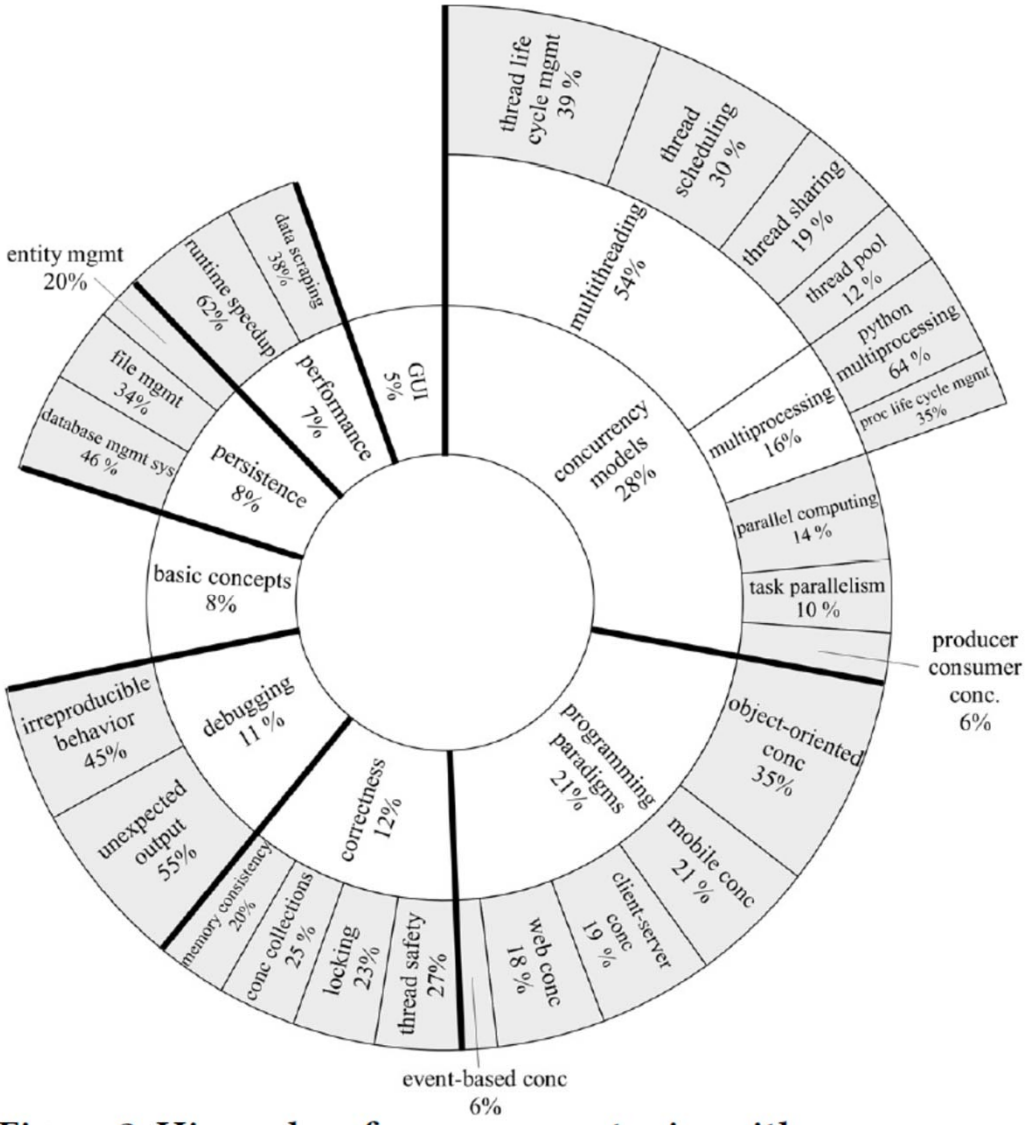

Figure 3: Hierarchy of concurrency topics with concurrency topics in gray and their categories in white.

結果:RQ3

- **スレッドセーフに関する質問が最もポピュラー**であった
- 逆に**クライアント・サーバ並行性**に関する質問が最も少なかった

結果:RQ4

- **データベース管理システム**に関する質問が，最も（答えを得るのが）**難しい**ものとなっていた
- 逆に，**メモリー貫性**に関するものが最も答えを得やすかった

結果:RQ5

- ポピュラーさと難しさの間には**負の相関**

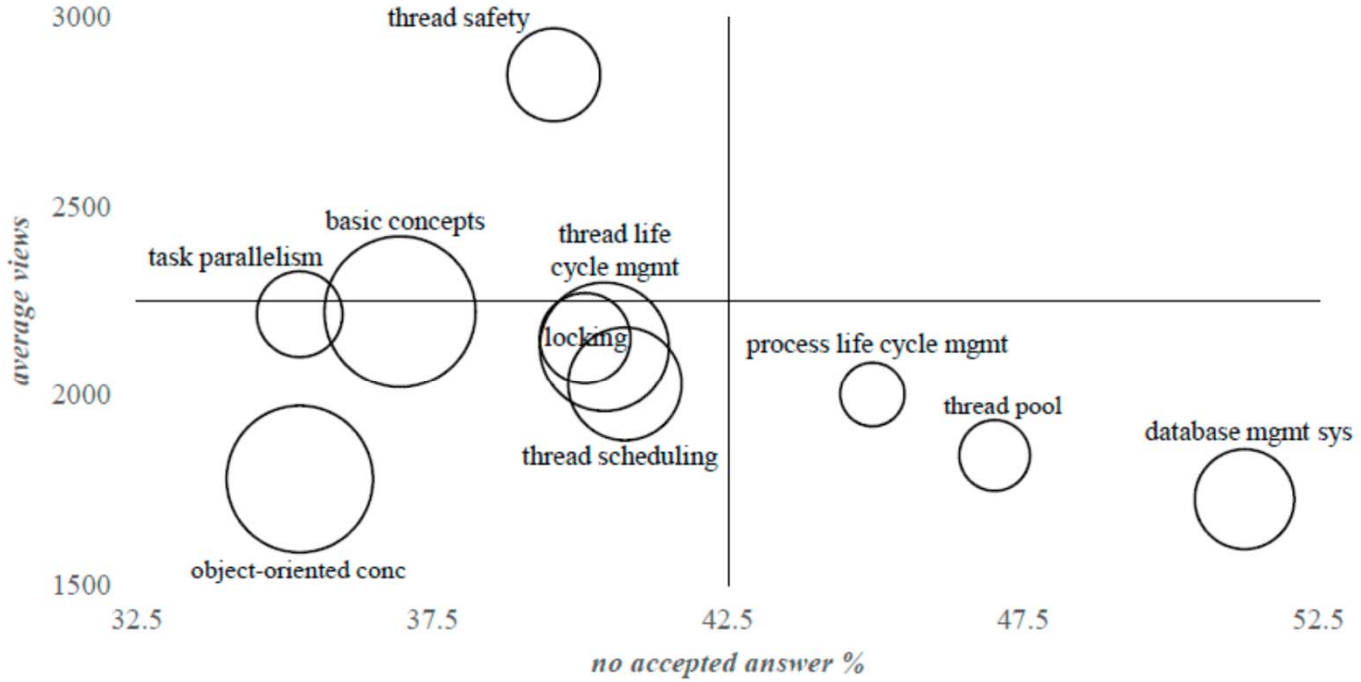

Figure 4: Trading off concurrency topics.

まとめ

データセットは https://goo.gl/uYCQPU

- 並行プログラミングに関し，**Stack Overflow での質問を調査**して開発者たちの持つ興味と困難さを分析
  - **トピックモデリング**を活用
- どういったトピックがポピュラー/難しいのかを明らかにした
  - 並行プログラマたちの（教育も含めた）助けに
  - 今後はコミットログとバグ報告の分析にも応用