

P21: 文書の提示順序がブロンズ適合性判定者に与える影響について

酒井 哲也 陶 思捷 (早稲田大学) contact: tetsuya@waseda.jp

情報検索テストコレクションの適合性判定者の分類

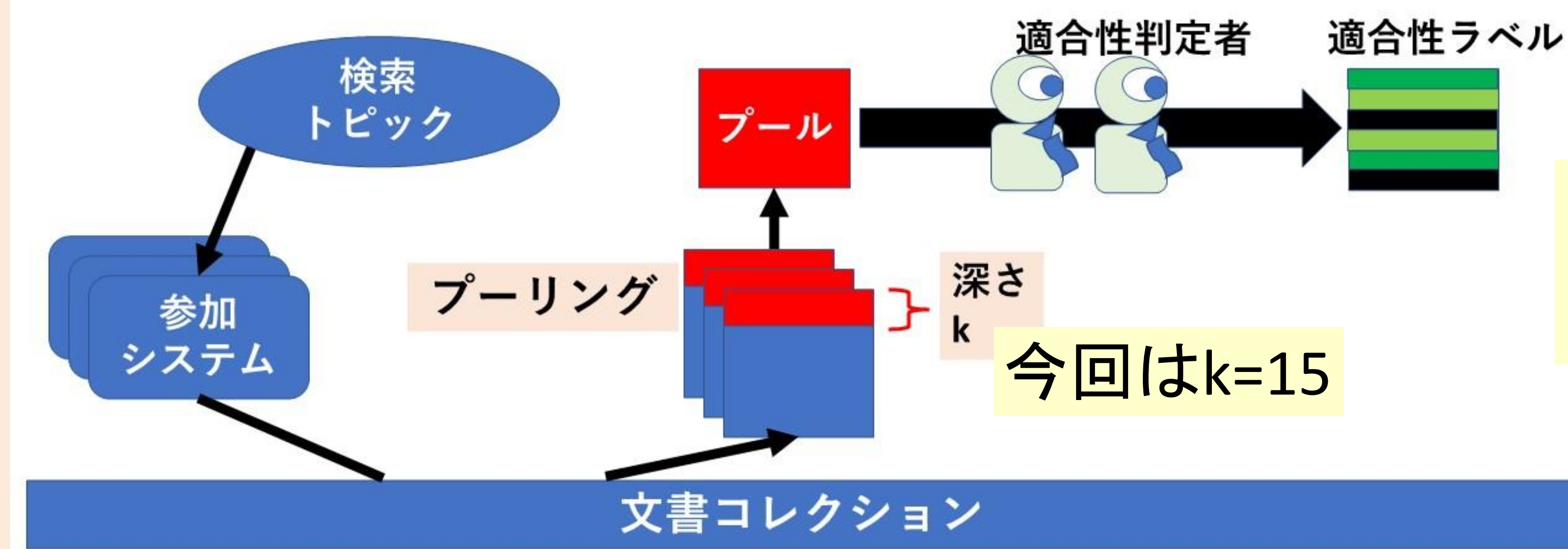
ゴールド: 検索要求をもつ当事者

シルバー: 当事者ではないが当該分野の知識が豊富

ブロンズ: 当事者ではない普通の人 (通常こちら)

深さkの文書のプーリング

検索対象が巨大なため、ごく一部を適合性判定の対象とする



当初の期待

プールされた文書の判定者への提示順序

PRI (prioritise): 多くのシステムが上位に検索した文書を優先

<https://research.nii.ac.jp/ntcir/tools/ntcirpool-ja.html> (NTCIR方式)

RND (randomise): 文書をランダムな順序で提示 (TREC方式)

PRIの利点は過去20+年間明らかにされていなかった!

構築した大規模データセット WWW3E8

- NTCIR-15 WWW-3 (英語ウェブ検索) タスク運営と同時に構築
- 文書コレクション: clueweb12-B13 (約50Mページ)、CMUが提供
- 検索トピック: 160件 (検出力分析によりサンプルサイズ決定)
- 各トピックにつき英語コースの学生から8件の適合性ラベルを収集 (PRI環境で4件、RND環境で4件): 合計259,000件の適合性ラベル

<https://waseda.box.com/WWW3E8corrected>

PRI vs. RND: 主な知見

RQ1: 判定効率が良いのは?

A1: 平均判定時間に実質差なし (PRIのほうが最初の高適合文書を見つけるのが早い傾向はあり)

RQ2: 判定者間一致度が高いのは?

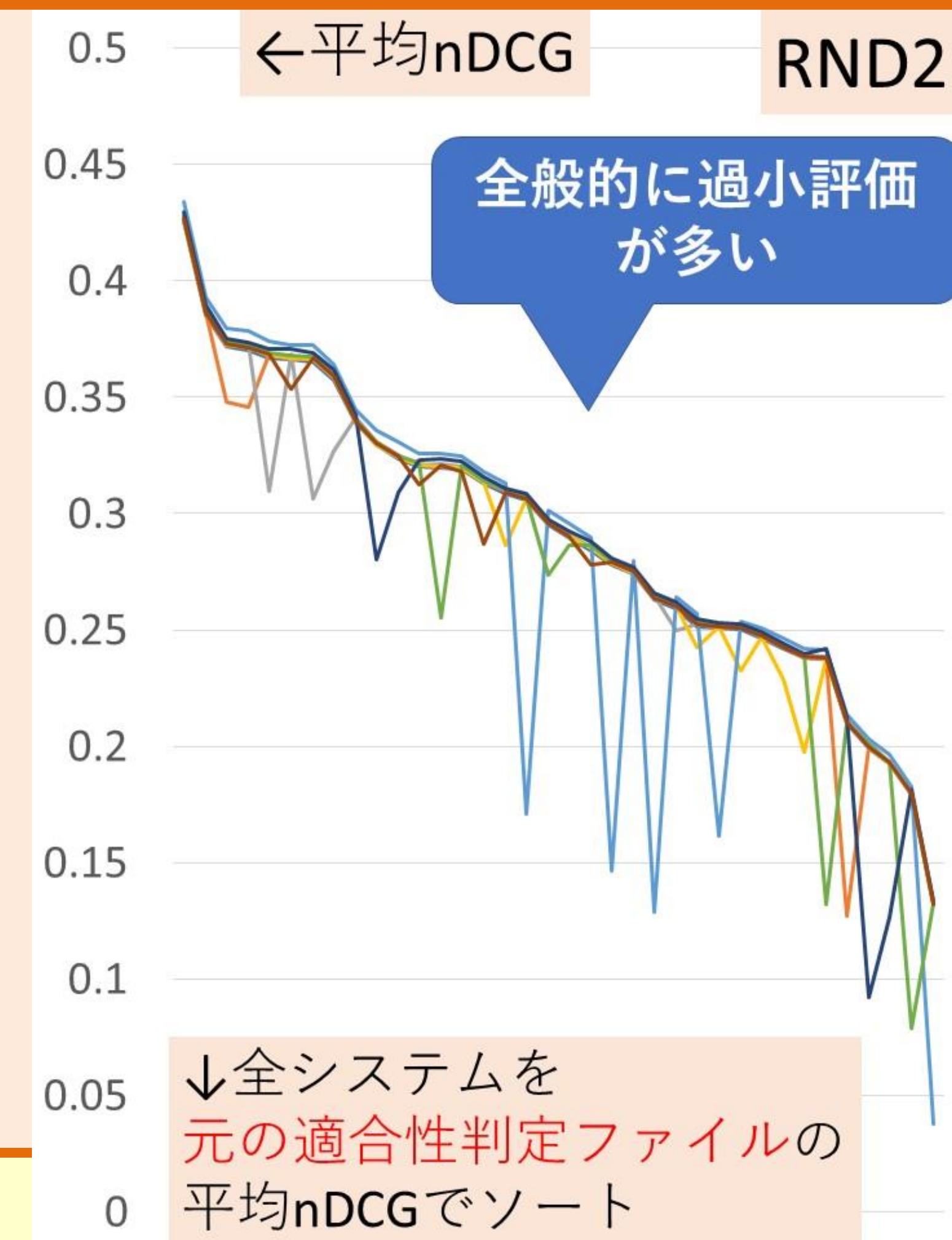
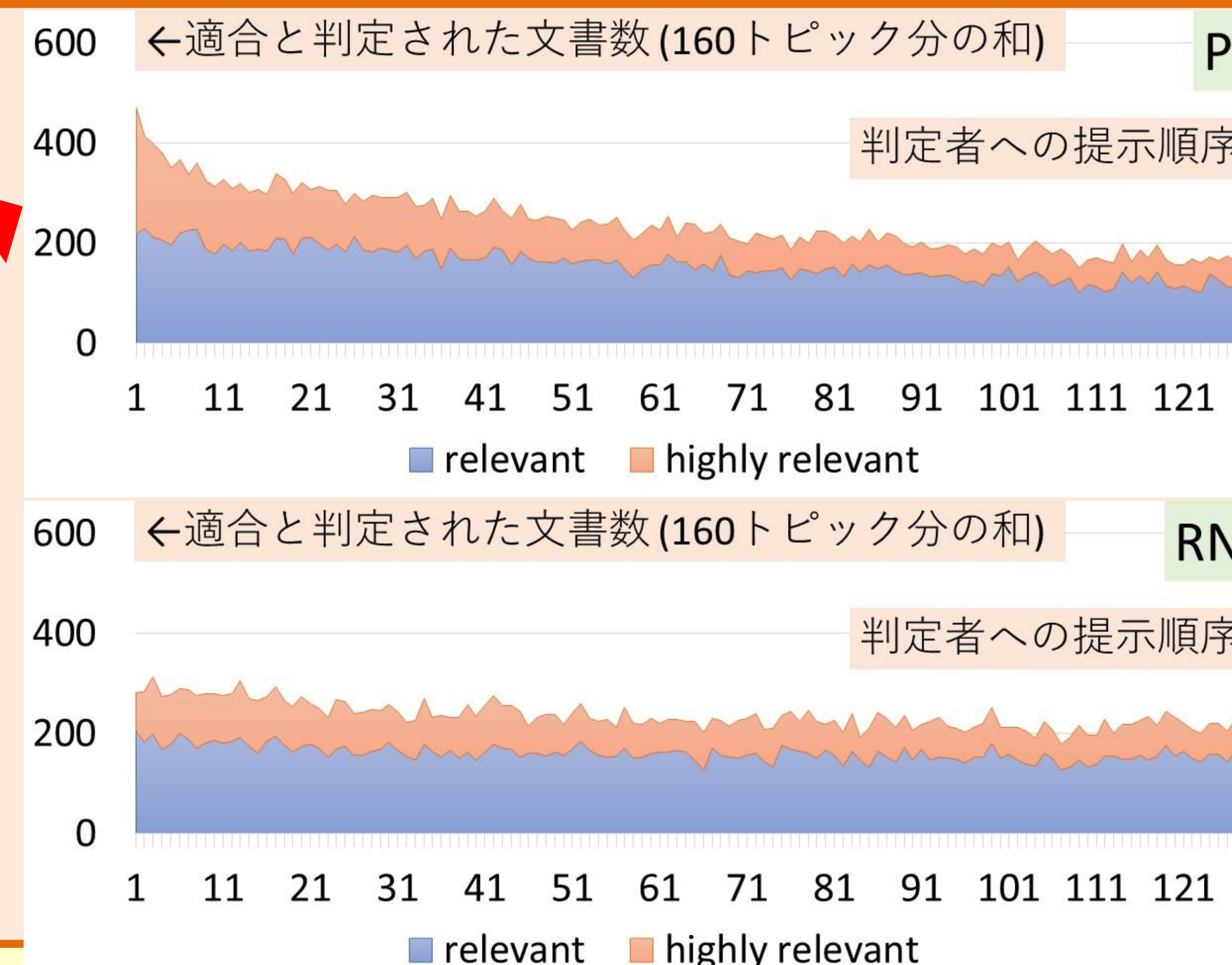
A2: 平均Krippendorff's α で実質差なし (効果量Glass's $\Delta < 0.09$)

RQ3: 異なる適合性判定ファイルにより複数のシステムランキングを得た場合、ランキング間の類似度が高いのは? (適合性ラベルの違いに対するシステムランキングの頑健性)

A3: PRI-PRI間の類似度のほうはRND-RND間よりも若干高いが実質差なし。ただしPRI-RND間類似度はこれらより有意に低い⇒ PRI、RNDではどの文書が適合とされるかがかなり異なる

RQ4: プーリングに不参加のシステムをより正當に評価できるのは? (未知のシステムに対する頑健性)

A4: PRIのほうが頑健性が高い傾向あり。理由: PRIは「ポピュラーな」(=多くのシステムにより上位に検索される)適合文書を発見する傾向があり、ポピュラーな適合文書は未知のシステムにも検索されやすいため、未知のシステムの正當な評価に役立つ⇒再利用可能性を求めるならPRIのほうがお奨め



今後

- ゴールド判定者だとどうなるか? (Done: 別途発表予定。ブロンズの場合と根本的な齟齬なし)
- ウェブ検索以外のタスクだとどうなるか?
- 一般に、正解ラベルづけのためのデータ提示順序は常にランダムでよいのだろうか?

詳細: Sakai et al.: Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents? (CORRECTED VERSION), 2022. <https://arxiv.org/abs/2211.00981>