

A Reproducibility Study of PLAID

Sean MacAvaney, Nicola Tonellotto

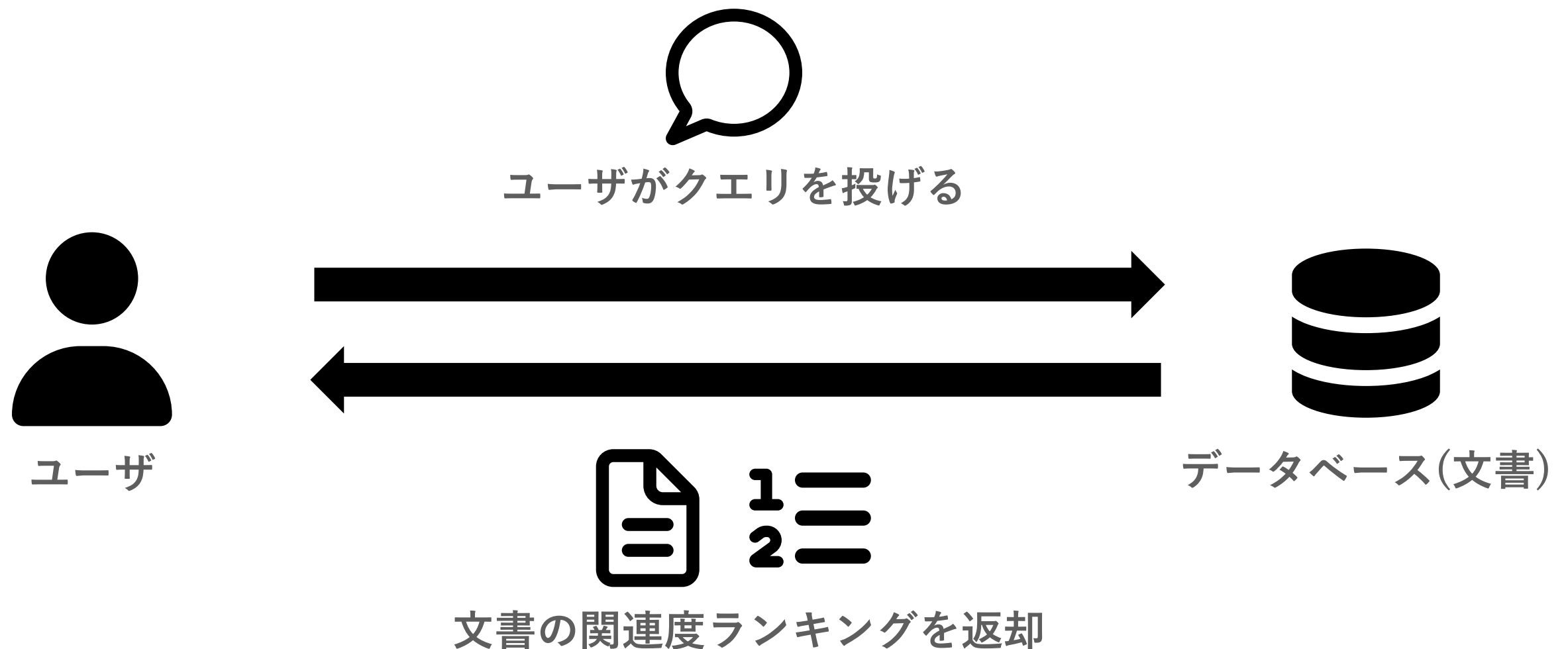
In Proceedings of the 47th International ACM SIGIR
Conference on Research and Development in Information
Retrieval (**SIGIR '24**), July 14–18, 2024, Washington, DC, USA.
ACM, New York, NY, USA, 9 pages.

S2310157 谷口

情報検索

大量の文書からユーザの欲しい文書を取得

- ・ 関連文書を、素早く取り逃さず高順位に置くことが理想



検索手法

BM25

- ・ 単語の一致で関連度を算出 (**高速** 低精度)

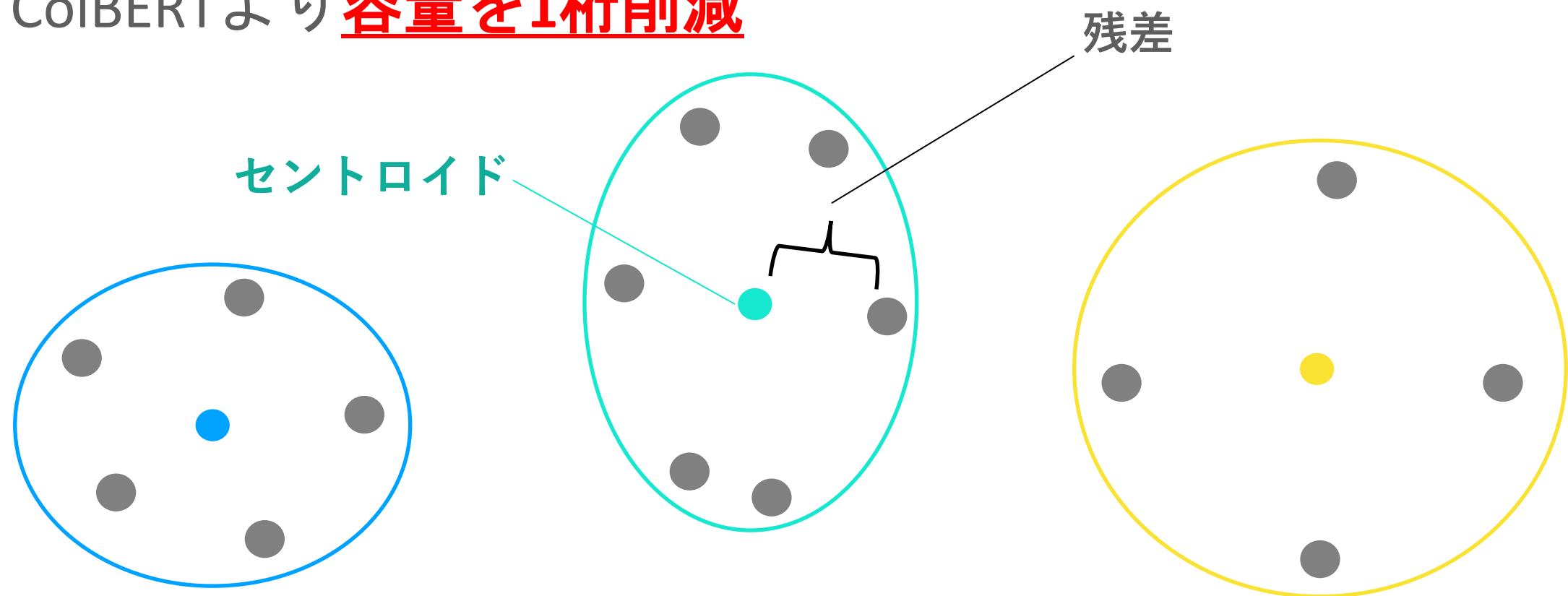
ニューラルネットワークモデル

- ・ 文書や単語をベクトル化+ベクトル間の類似度を算出
 - Dense Retrieval : 1文書1ベクトル(中精度)
 - Late interaction : 1単語1ベクトル(低速 **高精度**)
 - ColBERTはLate interaction

CoLBERTv2

CoLBERTの容量圧迫

- CoLBERT : 文書の全単語をベクトル化し保存 ⇒ 容量圧迫
- CoLBERTv2 : ベクトルを「セントロイド+残差」で表現
⇒ CoLBERTより 容量を1桁削減



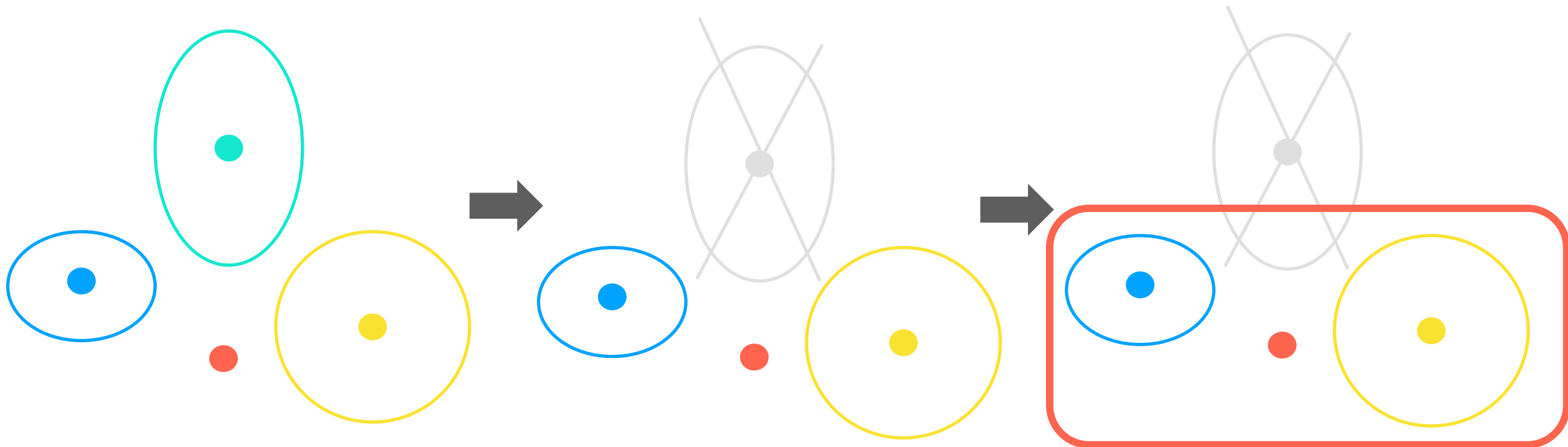
似たベクトル同士でグループ分け
グループの代表点(セントロイド)とセントロイドからの距離(残差)

PLAID

セントロイドを検索の高速化に活用する手法

▪ nprobes, t_cs, ndocsは適宜変更するパラメータ

1. クエリの1単語トークンからnprobes個のセントロイドを取得
2. 類似度がt_cs以下のセントロイドを除外
3. 上位ndocs件の文書に対してColBERTv2による検索



PLAIDの検討不足

PLAIDを提案した元論文

- ・ PLAIDはColBERTv2を用いた標準的手法より優れた実験結果
⇒ いくつか問題点がある

元論文の問題点

- ・ PLAIDのパラメータ: `n_probe`, `t_cs`, `n_docs`の説明不足
- ・ 「BM25で候補を絞る⇒ColBERTv2で再検索」という
シンプルな手法との比較が為されていない(ベースライン不足)

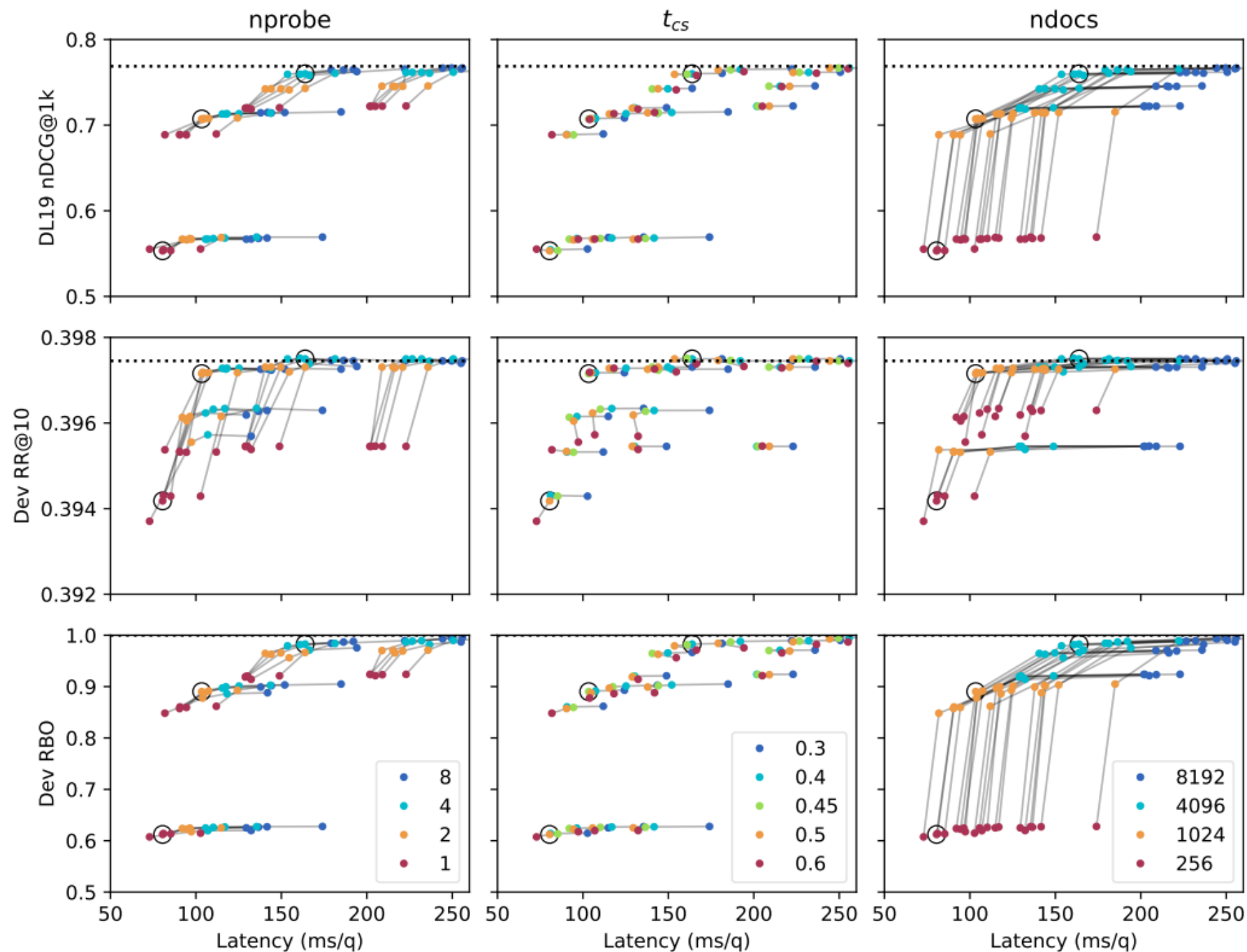
PLAIDのパラメータ

元論文での推奨値

System	nprobes	t_cs	ndocs
PLAID(a)	1	0.50	256
PLAID(b)	2	0.45	1024
PLAID(c)	4	0.40	4096

- ・ この推奨値はどのようにして見つけたのか？
- ・ 各パラメータを変更すると検索結果にどう影響するのか？

PLAIDのパラメータ



- $ndocs$ が最も影響が大きく1024以上が望ましい
- その次に $nprobes$ の影響が大きく、小さくし過ぎると精度に影響

PLAIDのベースライン不足

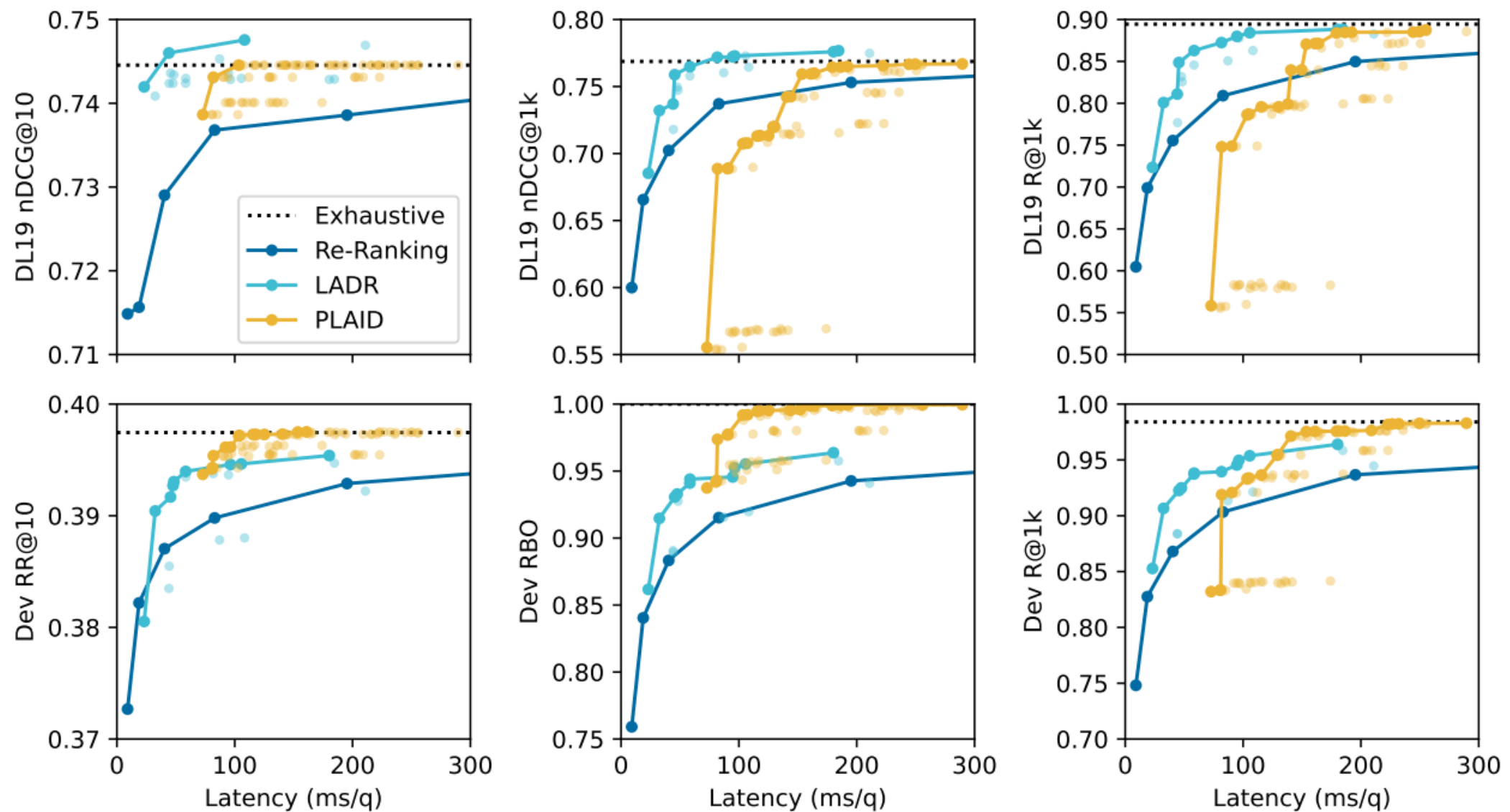
元論文のベースライン

- ColBERTv2
- SPLADEv2(Sparse Retrieval)
- BM25

足すべきベースライン

- 再ランキング手法(BM25 $\Rightarrow$ ColBERTv2)
- LADR(再ランキング手法の発展)

PLAIDのベースライン不足

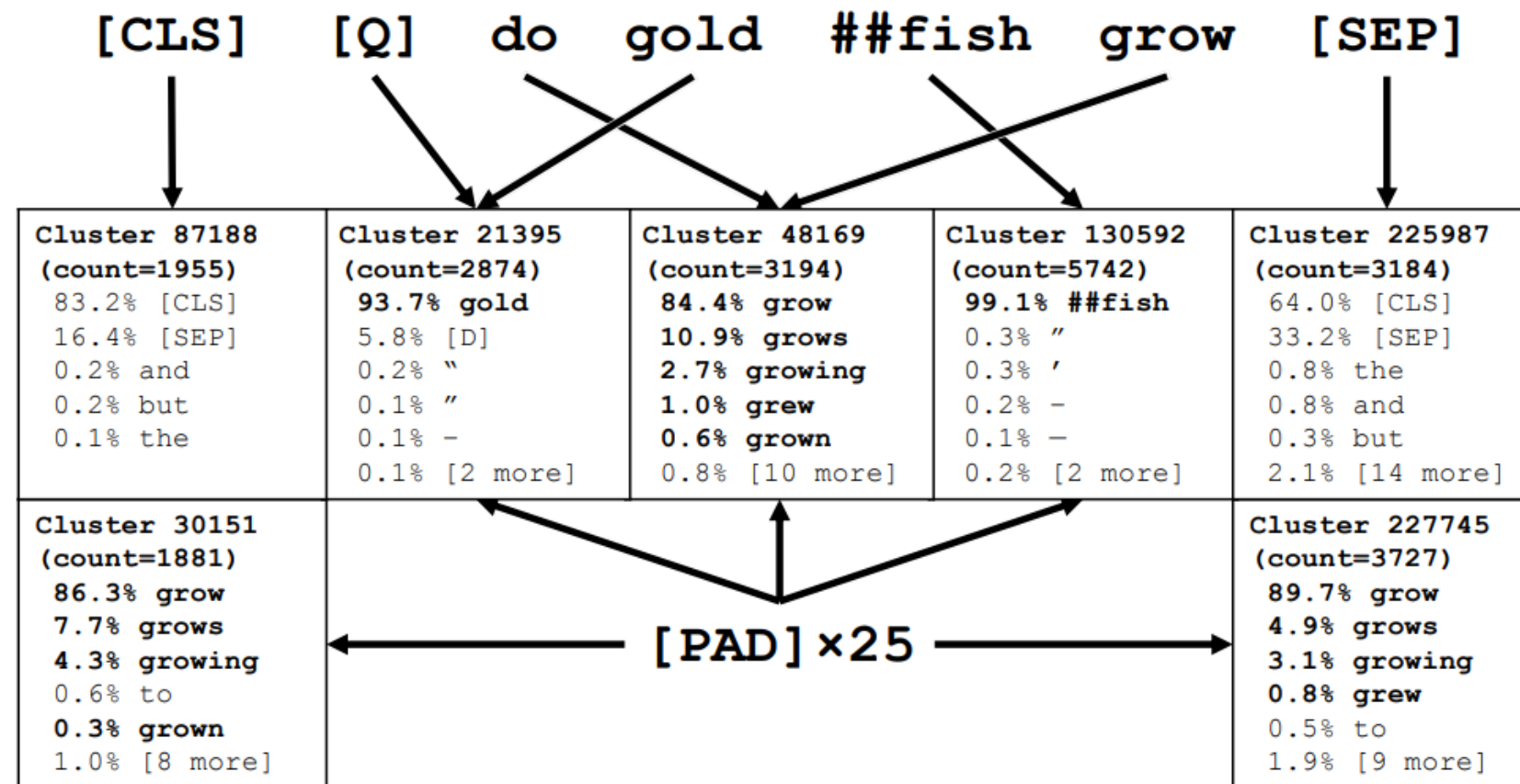


- 速度は再ランキング、LADRが優れている
- LADRは精度速度の両方で上回るケースも

PLAIDのクラスタ分析

LADRがPLAIDに拮抗した要因

- PLAIDはベクトルをグループ分けしてクエリに近いグループの中から文書を探す手法



- 同単語を変形したものが同じグループに入っているケース
- BM25のような単語の一致を見ているのと同ほとんど同じ

まとめ

情報検索手法

- 単語の一致を見るBM25
- ベクトルに変換するColBERT ⇒ グループ分けで容量削減したv2
- ColBERTv2の特性を生かしたPLAID

PLAIDの元論文の問題

- パラメータの説明不足
- BM25を用いた再ランキング手法との比較の欠如
- 再ランキング手法の一部はPLAIDに拮抗する結果
- PLAIDのグループ分けが本質的にBM25に類似している可能性

所感

- 自分の研究のベースライン探しを兼ねようとして失敗
 - PLAIDという手法を提案する論文なのかと思ったら
PLAIDを提案した論文について再検討する論文だった
- LLMではなく小規模なニューラルネットワークモデルを用いる手法で、自分の研究にはあまり関連しなさそう
- パラメータからベースラインまで論文内のことは
全部論理的に、必然的に導かれたものであると十分に説明
できることが求められるのだなと思った