



データセット利用研究の動向

大山 敬三（IDR事務局）

国立情報学研究所・教授

データセット共同利用研究開発センター長

本資料についての注意

- 本資料に記載の内容は、IDR提供データセットの利用者から提出された2017年度分までの研究成果のうち、公開となっている論文を分析したものです。
- 使用したカテゴリは本発表者が本分析のために主観に基づき設定したものです。カテゴリへの分類も、本発表者の主観により、基本的に論文タイトルのみを用いて行ったものです。そのため、分析結果の客観性や再現性は保証されませんので、客観性を要する議論の根拠としての利用はしないでください。

論文発表：分析の観点

- 各データセットを使用した論文全体を概観する
- 指向する「出口」を念頭に傾向を見る
- 次の方針でカテゴリを設定
 1. 適用の対象や機能がある程度想定できるもの
⇒ 対象や機能に対応するカテゴリ
 2. 上記以外でデータセットごとの特徴を利用するもの
⇒ 利用する部分と利用の形に対応するカテゴリ
 3. それ以外

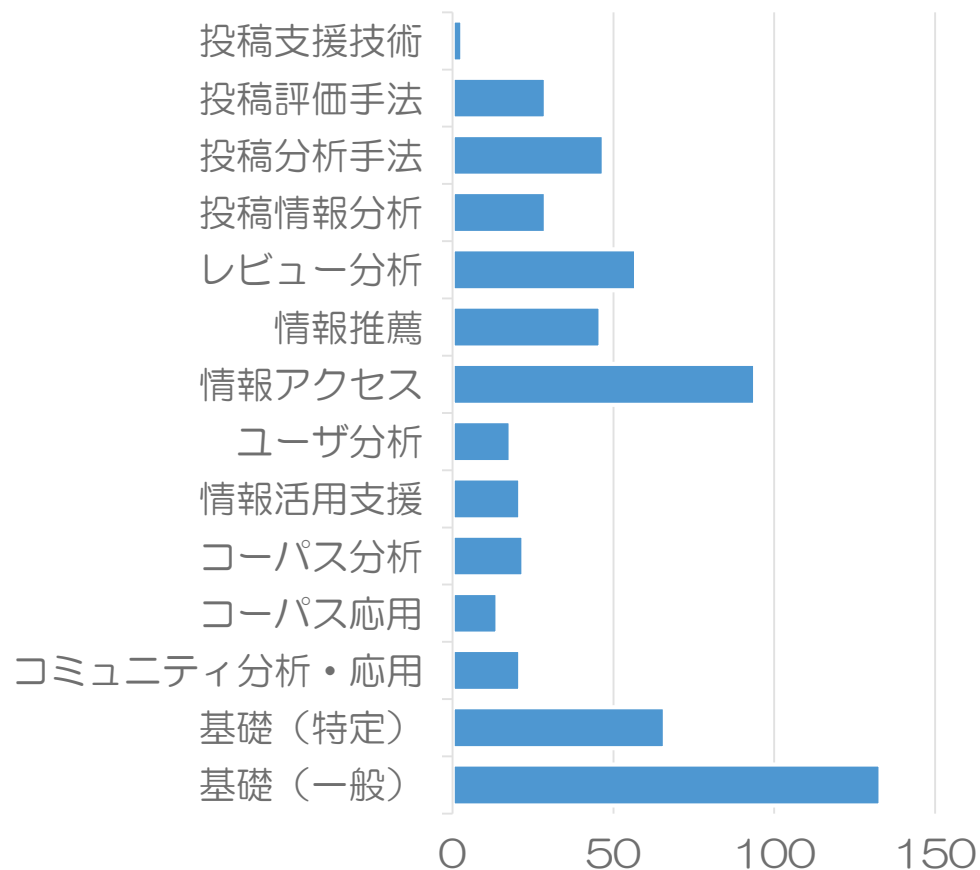
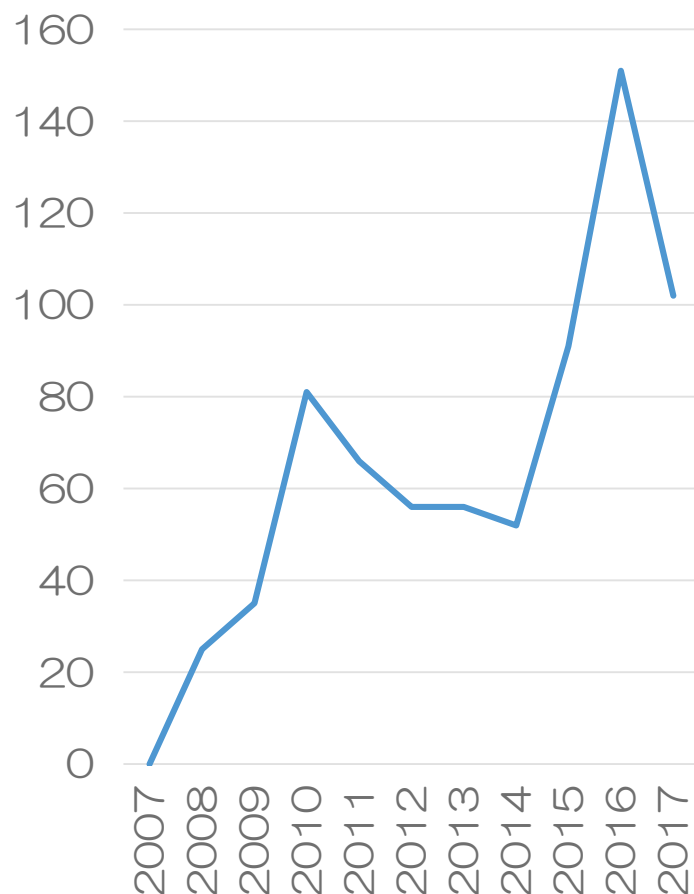
※ 研究分野の動向については今回は対象外
⇒ ポスター発表に関連研究あり

カテゴリー一覧

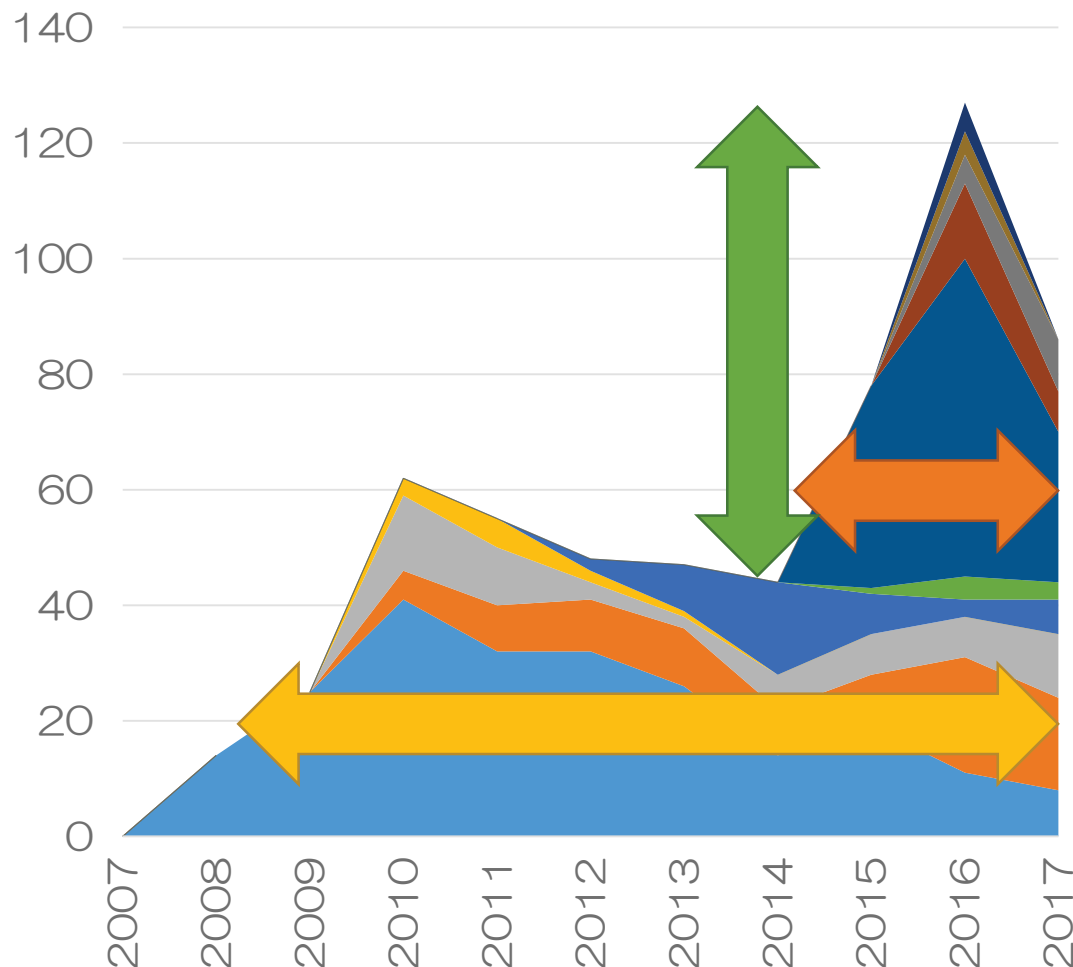
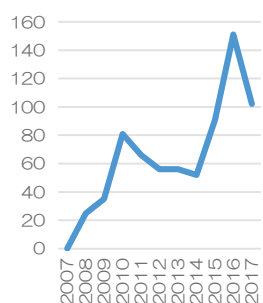
- 出口指向カテゴリ
(情報サイクルに注目)
 - 投稿支援技術
 - 投稿評価手法
 - 投稿分析手法
 - 投稿情報分析
 - レビュー分析手法
 - 情報推薦
 - 情報アクセス
- 特徴利用カテゴリ
 - ユーザ分析
 - 情報活用支援
 - コーパス分析
 - コーパス応用
 - コミュニティ分析・応用
- その他カテゴリ
 - 基礎（特定）
 - 基礎（一般）

※「投稿」は基本となるコンテンツ

論文数の全体的傾向

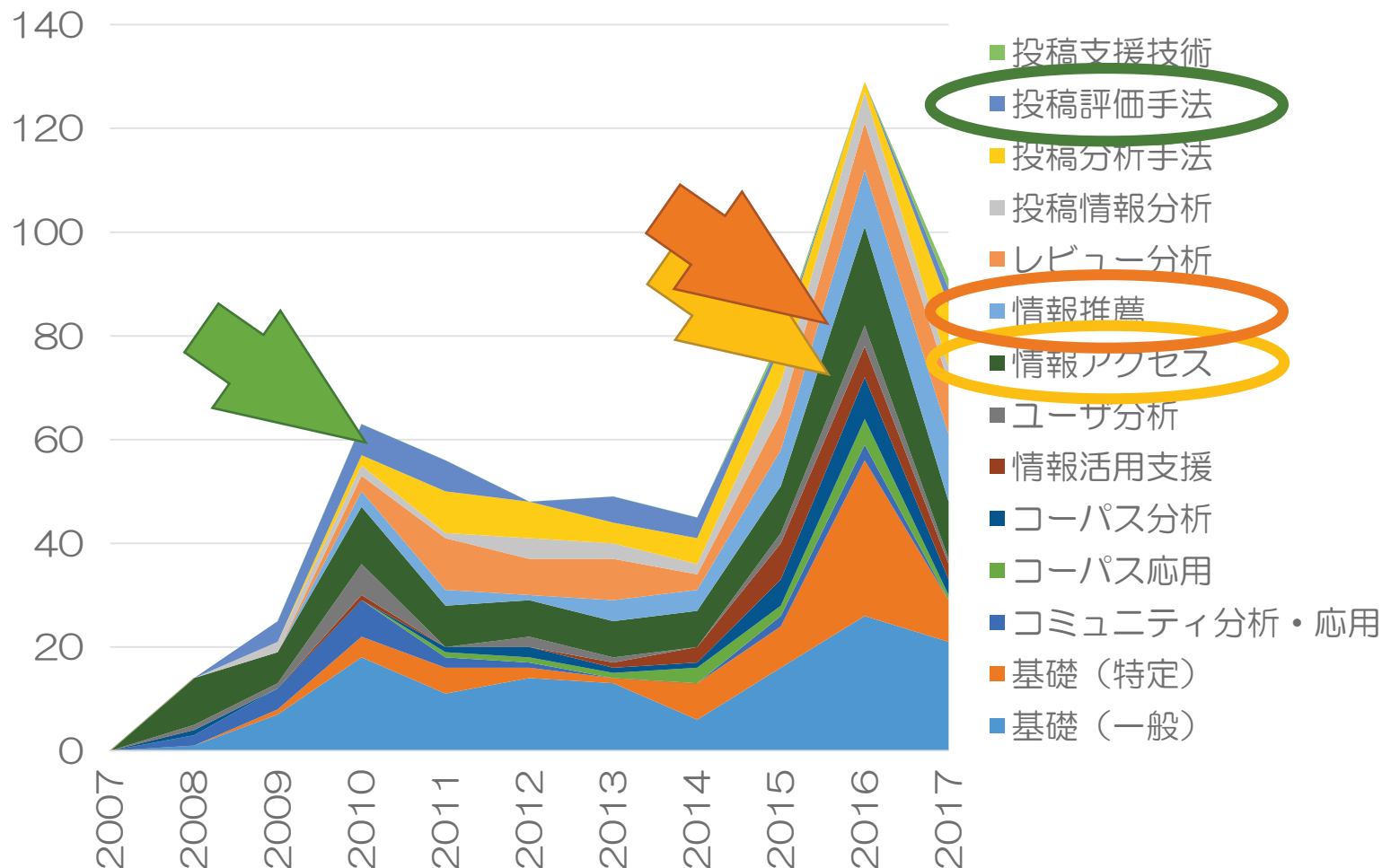
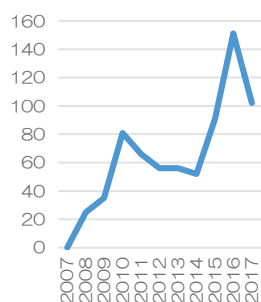


データセット別推移（件数）

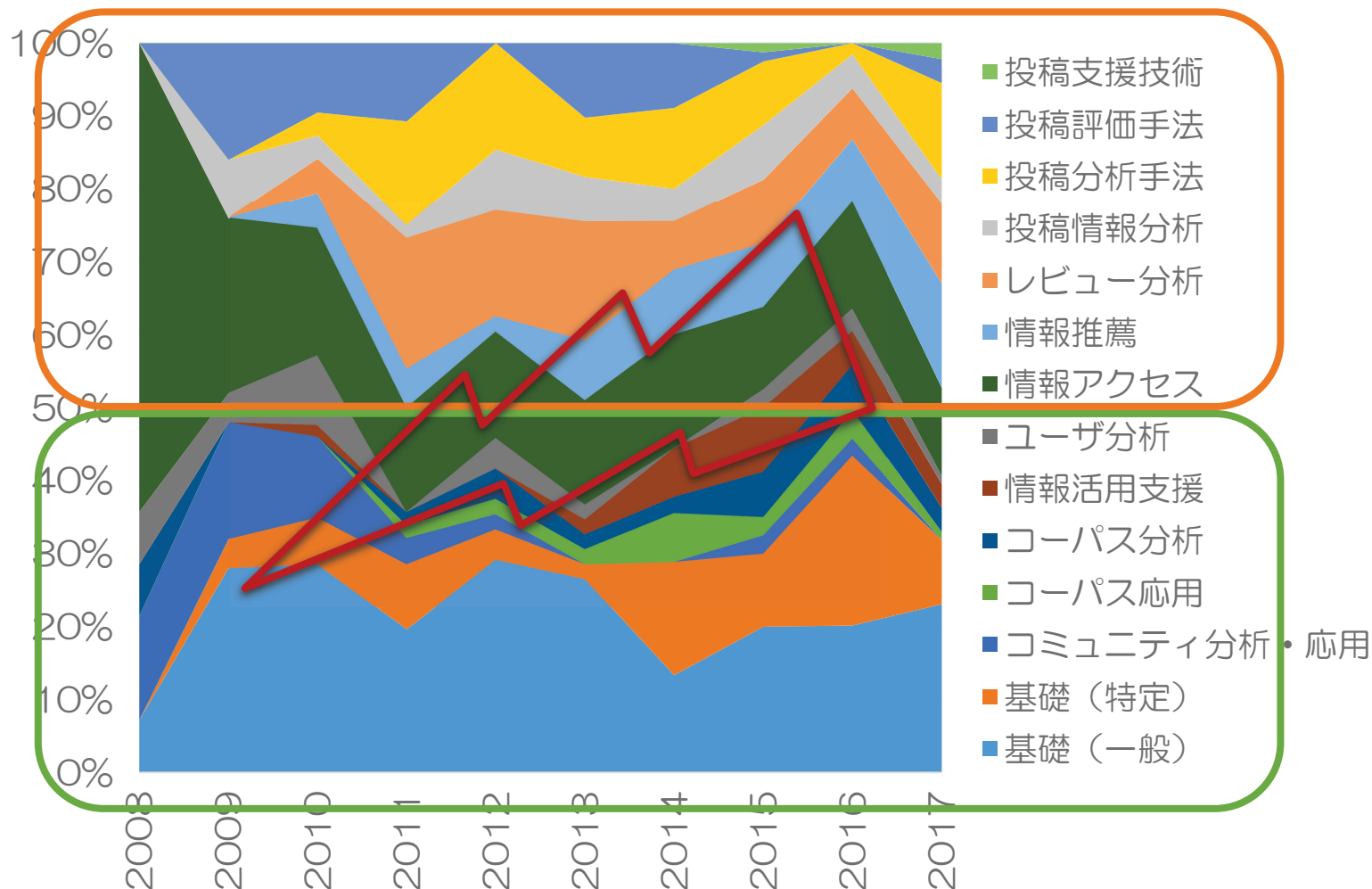
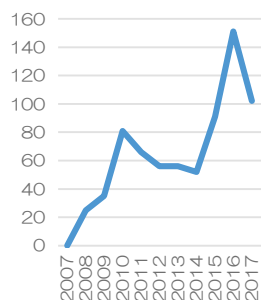


- ニコ動コメント等データ
- ニコニコ大百科データ
- 不満調査データ
- HOMESデータ
- クックパッドデータ
- HPビューティーデータ
- 楽天レシピデータ
- 楽天GORAデータ
- 楽天トラベルデータ
- 楽天市場データ
- Yahoo!知恵袋データ

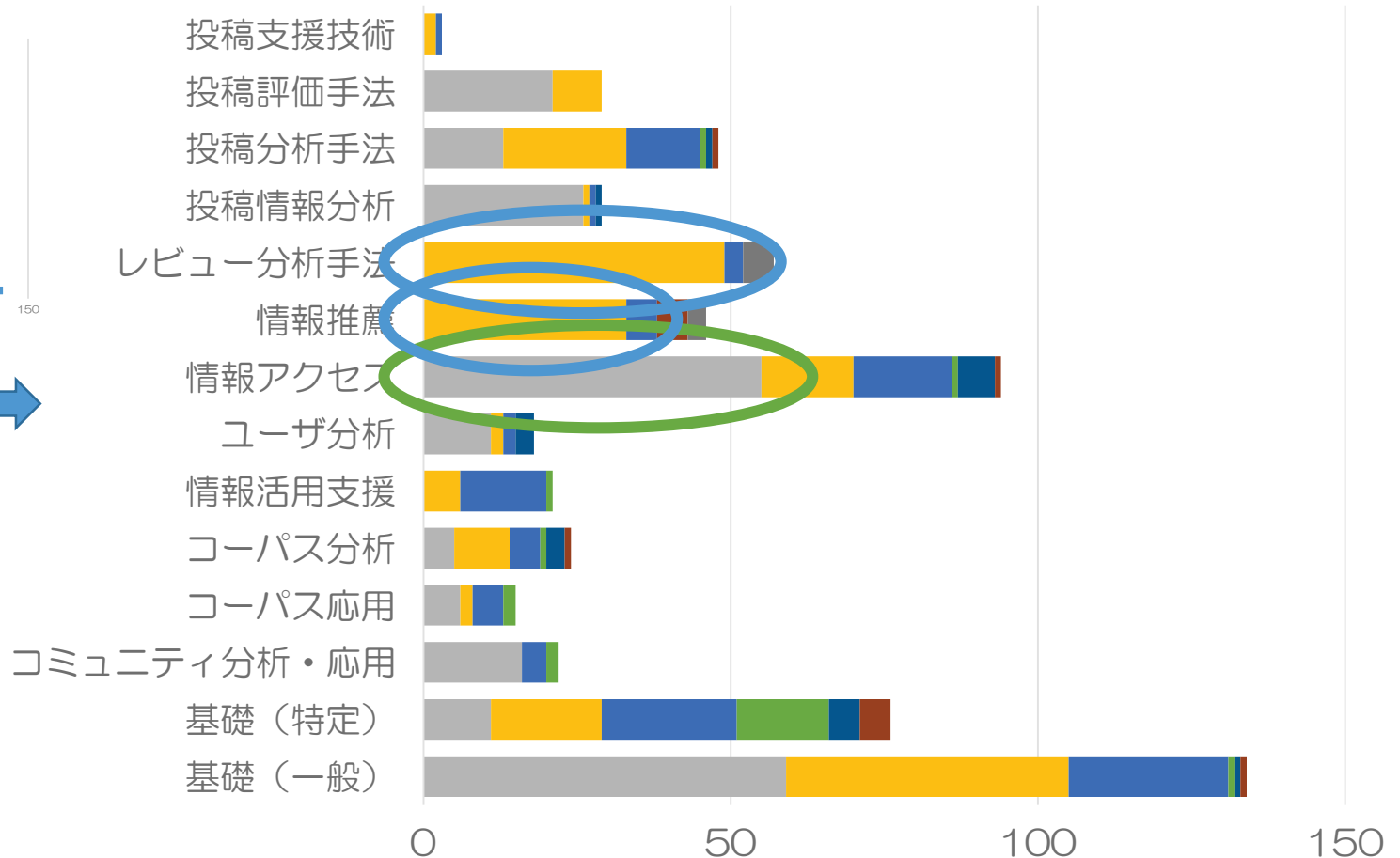
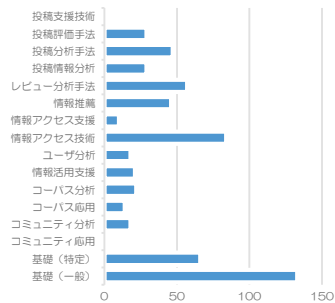
カテゴリ別推移（件数）



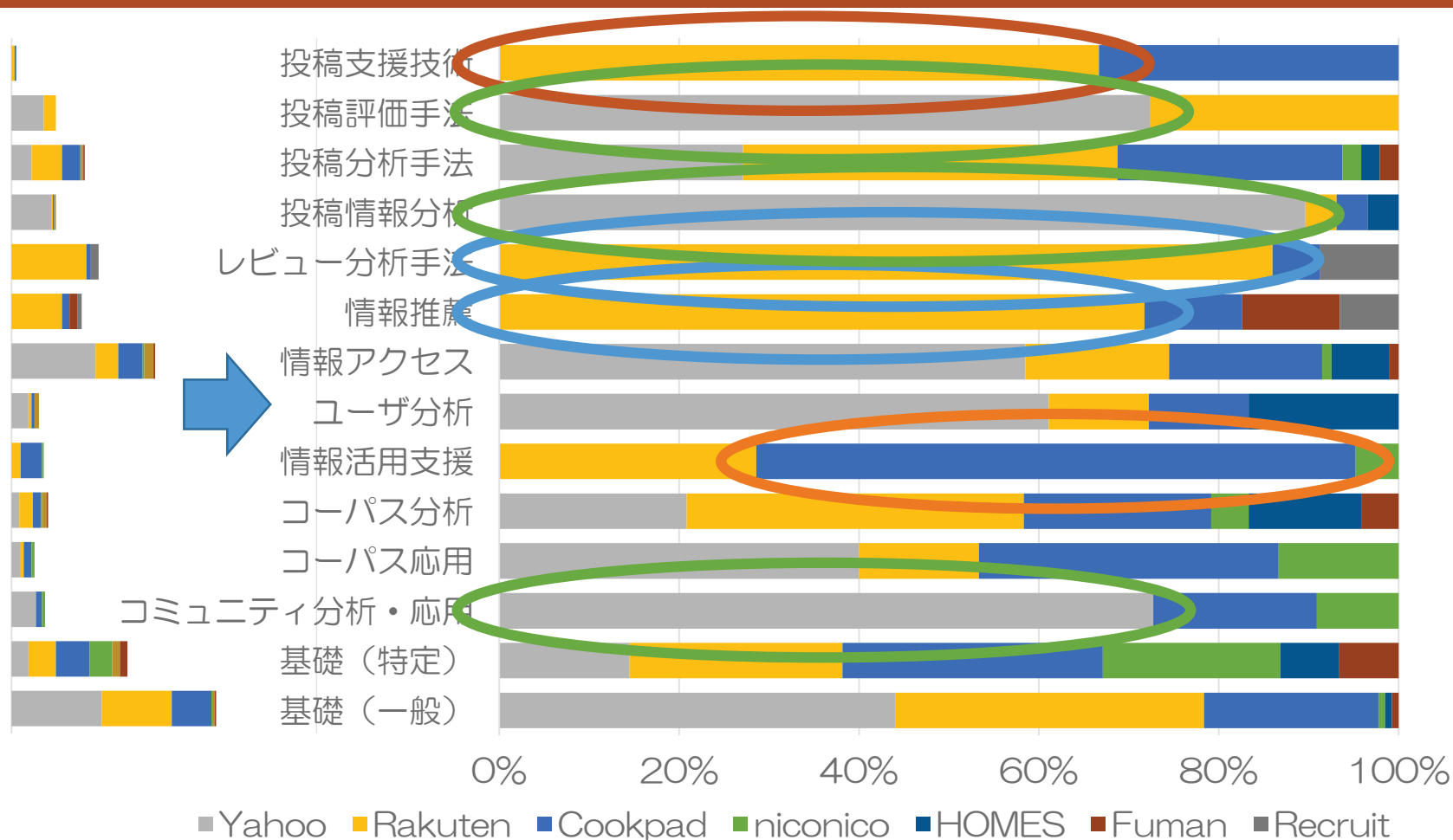
カテゴリ別推移（割合）



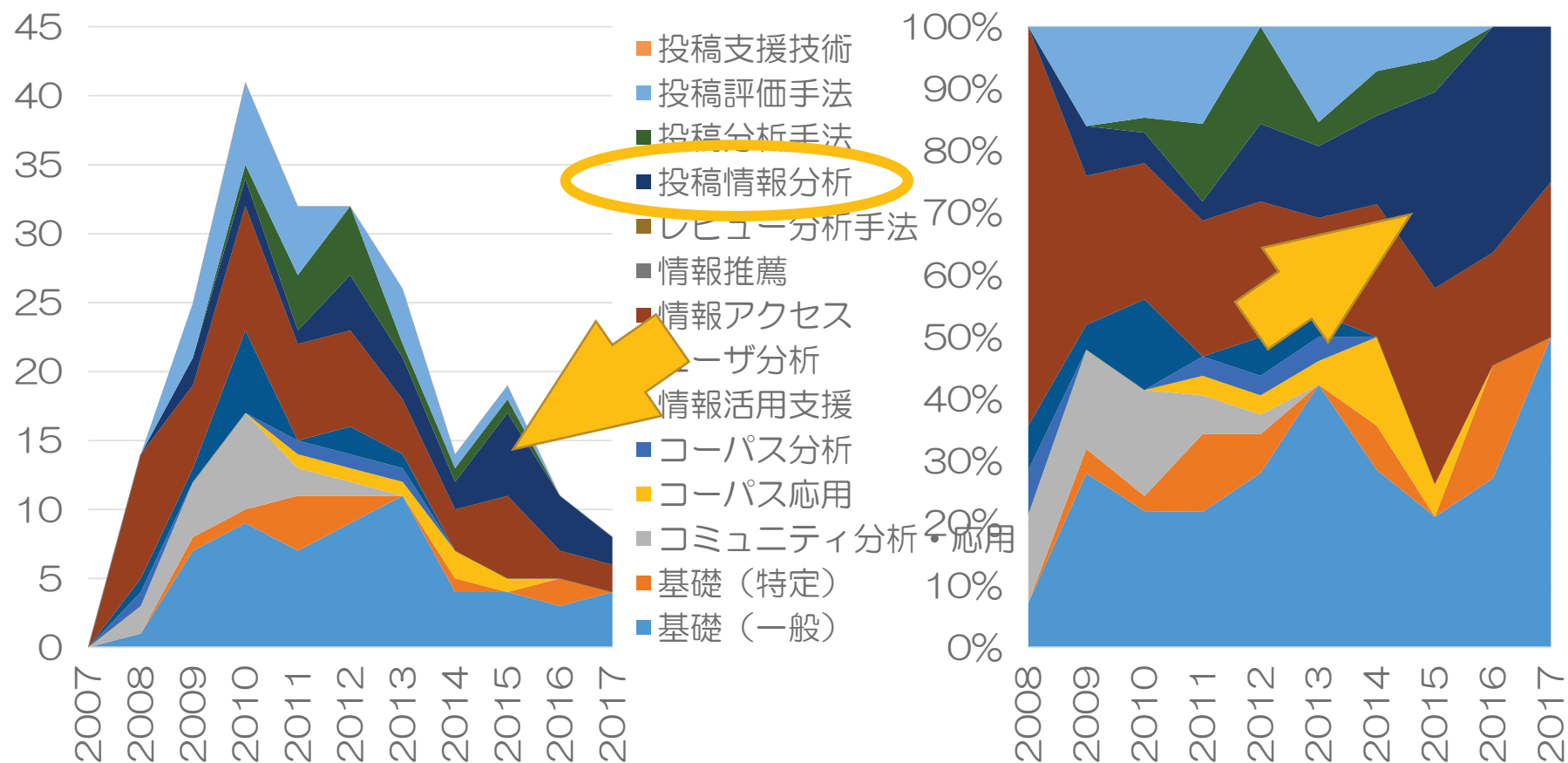
カテゴリ対データセット（件数）



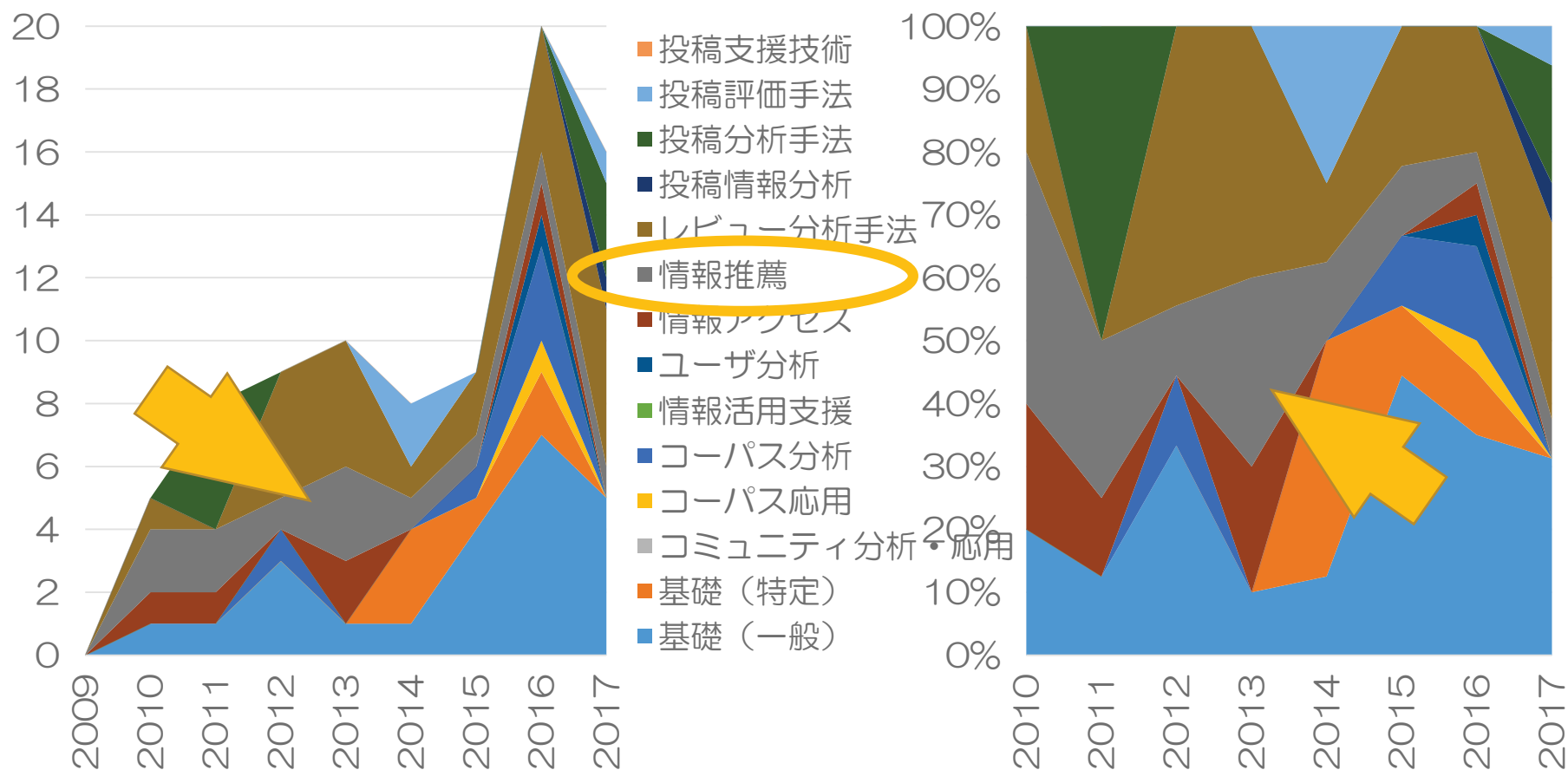
データセット対カテゴリ（比率）



Yahoo!知恵袋データ カテゴリ推移



楽天市場データ カテゴリ推移



傾向は？

- データセットにより傾向は大きく異なる
 - Yahoo!知恵袋データ
 - 情報アクセス : 件数が大
 - 投稿情報分析, 投稿評価手法, コミュニティ分析・応用 : 比率が大
 - 楽天データ
 - レビュー分析手法, 情報推薦 : 件数・比率とも大
 - 投稿支援技術 : 比率が大
 - クックパッドデータ
 - 情報活用支援 : 比率が大

傾向は？

- 全般的に，出口を意識した研究と基礎的な研究が拮抗
- 年次推移の大勢に明確な傾向は見られない
 - ヤフーと楽天以外はまだ年数が浅い
 - 年による変動が大きい（特にクックパッドの影響）
- カテゴリごとに見ると
 - 情報アクセスはほぼ一定の件数で安定
 - 情報推薦は安定的に緩やかに件数が増加
 - 投稿評価手法はNTCIRのCQAタスクが影響
 - 少数だが分野の広がりを感じるものもあった
 - Yahoo!知恵袋データの投稿情報分析

まとめ

- 「出口」を念頭において，発表論文を通じて研究動向の分析を行った。
- データセットごとの特徴に応じた多様な研究に利用されている。
- 主観的には一部だが研究の裾野の広がりが感じられた。

今後の課題

- 研究分野や適用技術の観点を入れた分析
⇒ データセットと研究テーマの間の適合性マップによる一覧性の向上
- データ提供企業の課題や期待との整合性に関する分析
⇒ データ提供企業と研究者が課題を共有できる仕組みに展開