情報検索における失敗の自動分類

中島　百花

本論文では，機械学習手法を用いて，情報検索における失敗を自動で分類する問題に取り組む．NTCIR-4 Web テストコレクションの適合性判定 8,781 件 (文書: 7,636 個，検索課題: 80 個) を分析対象として用いた．

情報検索の失敗の分類方法として，Buckley が提案した 9 種類，難波と酒井が Buckley の分類に追加した 2 種類，Savoy が提案した 6 種類の 3 つの分類方法がある．しかし，これらの失敗の分類方法は，人手による分類であるためコストがかかることや，分類のカテゴリの内容に重複があること，分類の粒度が揃っていないなどの問題点がある．そこで，分類を自動化するにあたり，失敗分類の再構成を行った．既存の 3 つの分類方法の共通点をまとめることによって，新たな分類方法として，以下の 8 種類のカテゴリを提案する：(I) 観点を外している，(Ⅱ) クエリの構造理解が必要，(Ⅲ) 入力されたクエリに誤りがある，(Ⅳ) 事前処理に問題がある，(Ⅴ) オントロジーなどの外部知識が必要，(Ⅵ) 単語の一致の有無を超えた推論が必要，(Ⅶ) クエリ・文書中の単語が曖昧／不明瞭，(Ⅷ) 文書中の単語の近接関係が必要．

NTCIR-4 Web テストコレクションの失敗事例 100 件に対して，提案する 8 種類のカテゴリの分類方法で失敗事例を分類し，失敗分析を行った．失敗分析の結果，多くの失敗事例は「観点を外している」というカテゴリに分類され，強い偏りが見られることが明らかとなった．よって，提案する 8 種類のカテゴリの分類方法では，失敗原因が 1 つのカテゴリに偏るため，失敗分析には有用ではないと推測された．

そこで，ボトムアップ型のアプローチとして，失敗事例に対して階層的クラスタリングを行い，各クラスタの特徴を分析した．クラスタ間の距離関数には Ward 法，データ間の距離の計算方法にはユークリッド距離を採用し，クエリの長さや BM25 などの 12 個の特徴量を用いた．クラスタ分析に基づいて，新たに情報検索の失敗の分類方法として，以下の 6 種類のカテゴリを提案する：(I) クエリ語が文書に出現する割合が高いが，単語の近接性が低い，(Ⅱ) クエリ語が文書に出現する割合が高く，単語の近接性も高いが，検索者の情報要求を満たす内容が文書に含まれていない，(Ⅲ) クエリの中で重要度の高い単語の出現回数が少なく，近接性も高くはないため，観点を外している，(Ⅳ) クエリ語が文書に出現する割合が低いため，観点を外している，(Ⅴ) クエリの長さが長く，また，クエリに珍しい単語が使われている可能性があるため，オントロジーなどの外部知識が必要である，(Ⅵ) クエリの最初の単語，すなわち，検索上，最も重要な単語のみを強調し，その他の観点を外している．

最後に，各クラスタに分類された事例を用いて，クラスタ分析に基づいて新たに提案した分類方法で自動分類を行った結果，高い分類精度が達成できることが明らかとなった．

（指導教員　　加藤　誠）