統計データ収集のためのフォーカストクローラ

和久井　拓斗

統計データは政策や社会調査，企業などでの新規事業の立案時や，研究機関における基礎データとして有用である．しかし，統計データ自体の重要性は広く認識されている一方で，多くの人々に対して統計データへの高いアクセシビリティを提供することの重要性についてはあまり注目されていない．現状，統計データはさまざまなウェブサイトから公開されているが，それらに統一的にアクセスできるデータポータルサイトが存在しない．

統計データに対してリテラシーの高い専門家に限らず，多くの一般的な人々にとっても容易にアクセスしやすいプラットフォームがないことは問題があると考える．しかし，そのようなプラットフォームを実装する際には次のような技術的な課題にも直面する．それは，日々変容するウェブ空間の中から，どのように統計データを効率良く収集するかという問題である．

本論文では，上述の問題の解決策として，ウェブ上に存在する統計データを効率よく収集するためのフォーカストクローリングアルゴリズムについて提案する．特に，クローラの探索期におけるウェブページの選択方針について議論するものである．提案するフォーカストクローラでは，収集済みのウェブページに基づいて，そのウェブページを含むサイトが統計データを含む確率を推定し，その推定値によって次の探索対象を決定するような適応的な探索方針を用いる．具体的には，各サイトに対する判定誤り確率を推定しながら，その値を部分的フィードバックとするバンディットベースのアプローチによって，次のクロール対象を決定するという方針をとる．その目的は，予め決められた探索クロール (探索期におけるクロール) 回数において，各サイトに対して誤った判定をしてしまった数の最小化をすることである．

提案手法の有効性の確認のため，ランダムにウェブページを選択していく方針をベースライン手法とし，2 種類の実験を行って提案手法と比較する．一つ目はランダムに生成した確率に基づくクローリングの疑似的なシミュレーション，二つ目は実際のウェブページを元にした実データに対する比較実験である．それらについて，精度，適合率，そして再現率を元に比較評価を行った．結果として，少ない探索クロール回数においては，特に再現率において，提案手法がベースライン手法を上回った．このことから，限られた探索クロール回数における提案手法の有効性を確認することができた．

（指導教員　　加藤　誠）