原島 純（黒橋教授）

「Studies on Re-ranking and Summarizing Search Results」

（検索結果の並べ替えと要約に関する研究）

平成 25 年 3 月 25 日授与

現在、ウェブから情報を取得するには、検索エンジンの利用が必要不可欠である。検索エンジンを用いるユーザの情報要求は、クエリと呼ばれる単語のリストで表現される。クエリは次の三つタイプに分類できる。

- 誘導型のクエリ … 特定のサイトやウェブページに到達するためのクエリ
- 取引型のクエリ … 商品の購入や予約など、ウェブ上での何らかの取引を目的としたクエリ
- 情報型のクエリ … 広く情報を収集するためのクエリ

クエリの多くが情報型だと言われている。しかし、既存の検索エンジンは、検索した文書のリストを返すのみであり、情報収集に十分有用であるとは言えない。

本研究では、ユーザの情報収集を支援する種々の手法の開発に取り組んだ。まず、検索結果をリランキングする新しい手法を提案した。既存のリランキング手法では、文書の表層に現れる単語の情報のみを用いてリランキングを行う。一方、提案手法では、トピックモデルを用いて、文書に潜在する単語の情報も用いてリランキングを行う。実験の結果、文書の潜在的な情報がリランキングに有効であることが示された。

次に、検索結果からクエリに関する要約を生成する手法を開発した。検索結果から情報を収集するためには、検索された各文書に目を通さなければならない。これは非常に骨の折れる作業である。本研究では、トピックモデルを用いて検索結果からクエリに関する重要文を抽出し、トピック毎に要約を生成する手法を提案した。実験の結果、提案手法によって生成された要約が、クエリに関する情報を収集するのに有効であることが分かった。

最後に、日本語文を圧縮する新しい手法を提案した。要約は、できるだけコンパクトな方が良い。そのような要約を生成するためには、重要文を抽出するだけでなく、抽出した文を圧縮する必要がある。しかし、日本語文に対する既存の圧縮手法は、文節から情報量の小さい単語を除いたり、文法的な圧縮文をつくったりする能力に欠ける。本研究では、ラグランジュ緩和を用いて、文節から情報量の低い単語を除きつつ、文法的な圧縮文を生成する手法を提案した。実験の結果、提案手法が、情報量と文法性を保ちつつ、文を柔軟に圧縮できることが示された。

以上のように、本研究では文書のリランキングと要約、日本語文の圧縮に取り組み、その解法を提示した。今後は、本研究での成果を実社会に還元し、ユーザの情報収集を支援していきたいと考えている。

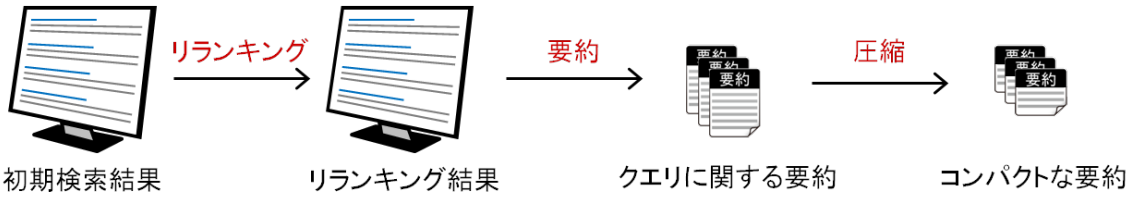

図 1 本研究の取り組み