論文の内容の要旨

Studies on Document Clustering Considering Users' Interests
(利用者の視点を反映する医学生物学文献自動分類に関する研究)

山本泰智

生物医学分野における研究技術及びそれに伴う研究の進展により、発表文献数の急激な増加と領域の細分化が生じている。このため、興味のある遺伝子の機能についてなど、これまでに得られている知見を網羅的に知ることが困難になっている。問題は、関連文献の数が研究者一人では到底読みきれないほどに非常に多いことと、関連文献を探し出すことが容易ではないことである。これらの問題に対処するため、計算機を利用し、研究者の関連文献調査にかかる手間を軽減し、ひいては大量の文献に書かれた知識に基づく仮説形成の一助になるシステムを開発することは有益であると考えられる。

現在、利用者が入力する語(キーワードまたは検索語)に関連する文献を検索する文献検索システムは存在するが、幾つかの問題がある。通常一つの研究課題 (例えば、ある遺伝子に関する研究) は複数の観点に立つ課題 (当該遺伝子により発現するタンパク質の機能や、当該遺伝子が関わる疾患に関する研究など) に分かれるため、優れた文献検索システムをもってしても、利用者がそれらの観点に関する知識をある程度持ち合わせていない限り、適切な文献を取得することが難しい。課題を示す検索語 (例えば、遺伝子名) だけでは検索される文献数が多くなりすぎる一方、複数の検索語の組み合わせが不適切であれば、得られる文献は不十分なものとなるだろう。更に、検索結果を閲覧して初めて存在を知る観点もあることが考えられ、別の視点からもう一度検索結果を見直すこともありうる。例えば、ある遺伝子がある疾患に関係していることを知った後で、検索結果を遺伝学に関する視点から見るといった場合があげられる。

この問題に対する解決法の一つして、既存の文献検索システムから得られる、ある研究課題に関する文献情報を、利用者の与える視点を反映して階層的あるいは非階層的にクラスタリングするシステムを開発した。個々のクラスタは、当該研究課題のそれぞれの観点に対応する。ここでいう文献情報とは、文献の題目、要旨、著者、関連語などである。また、研究課題に含まれるものは、当該検索システムが検索語として受入れられるもの全てとする。例えば、ある遺伝子名を検索語とした場合には、本システムにより得られる結果は、当該遺伝子の機能に関する文献情報のクラスタ、当該遺伝子と疾患に関する文献情報のクラスタなどである。

本システムの特徴は、利用者の視点に応じてクラスタリングを動的に変化させることが出来る点である。この結果、利用者はさまざまな視点から検索結果を俯瞰することが可能となる。利用者は視点を、システムが提示する検索結果に含まれる文献に関連した統制語のリストから選択することでシステムに与える。また、取得した複数クラスタの中から任意の数だけ選択し、そこに含まれている文献情報のみを対象とした階層的もしくは非階層的クラスタリングを行うことができ、従ってクラスタリングの繰り返しによる、クラスタリング対象領域の絞込みも可能である。

本システムは文献の題目及び要旨を処理するために、自然言語処理手法を採用し、また、幾つかの領域固有知識資源を利用している。このうち領域固有語辞書を出現語のステミングに、そして領域固有統制語彙をクラスタのラベル付け、および利用者がシステムに視点を伝えるために用いている。題目及び要旨は、文献情報毎に、出現する語の重要度を数値化したものを要素とするベクトルとして表現され(ベクトル空間モデル)、文献情報間の類似度を取得するために用いられる。利用者の視点を反映させるために、情報検索手法の分野で研究されてきた関連性フィードバックを基にしたベクトル変換を行う。これは領域固有統制語毎に、予め用意した一定数の語と当該統制語との関連性の強さをそれぞれ数値化したものを要素とするベクトルを用意しておき、文献検索システムで得られた文献情報の個々のベクトルと線形和を取ることにより実現している。ある語と統制語との関連性は、統制語毎に当該統制語が付されている全ての文献情報群の題目と要旨に出現する語と、全文献情報群のそれらに出現する語の頻度情報を基にして取得している。ベクトル空間モデルを用いた自然言語処理を行う際に問題となるベクトルの高次元性と多義語・類義語の存在に対し、ステミングのほか、忌避語、頻出語、及び希少出現語の削除、特異値分解を行っている。クラスタリングアルゴリズムは、階層的クラスタリングについては群間平均法を、非階層的クラスタリングについては buckshot アルゴリズムを基にして並列化等の改良を施した k-means を用いている。

クラスタリング結果は、階層的クラスタリングについてはデンドログラム、非階層的クラスタリングについては 2 次元平面上への各文献情報を点とする描画と、より詳細な情報が得られる表により提示される。2 次元平面上への描画を行う際に必要となる各点の座標は主成分分析により取得している。階層的・非階層的クラスタリングいずれの結果に対しても、クラスタ毎にクラスタラベル、複合語を含む関連語が表示され、非階層クラスタリング結果については、クラスタ毎に特徴的な文を当該クラスタに含まれる文献情報から抽出して表示する。

本システムの有効性の評価を、クラスタリングの有効性、及び統制語を与えることによるクラスタリング結果への影響の点から行った。さらに既存類似手法との比較、及び複数のケーススタディについて議論した。