論文審査の結果の要旨

氏名 石井 奈都

本論文は序論（第 1 章）を除いて大きく 3 つの部分からなり、第 1 部（第 2 章）は画像の自動分類法、第 2 部（第 3 章）はキータームの自動抽出法、第 3 部（第 4 章）はこれらを実装したデータベースについて述べられている。

学術論文の膨大な情報をデータベース化するにはその論文に書かれた情報を適切にキーワード化して検索・俯瞰できるようにする必要があるが、現状ではこの作業は人手に頼らざるを得ない。高い精度で情報を自動抽出する手法の確立は、大規模データベースの効率的な構築・運用のために不可欠である。本研究はこの目的に向けて、脳科学研究の分野に着目し、論文中の画像に関する情報を自動抽出する鍵となる手法を開発した。

本研究は最初に、「ある画像がどういう実験の結果を示すものであるか」を正しく分類して記述する方法の開発を行った。脳画像の取得には主に MRI, fMRI, CT, MEG, EEG, PET の 6 種類の実験手法が用いられるが、これらで得られる画像は互いに似通っており、画像認識の方法だけでは正しく分類することができない。そのため、関連するテキストを解析してクラス分類することが不可欠である。本研究では、テキスト中の単語をベクトルで表し、サポートベクターマシンを用いて機械学習を行った。従来発表されてきた手法では、この際に個々の単語の集合（bag of words）のみを利用していたが、本研究では 2 単語の組（bigram）を考慮に入れることで特徴を効果的に抽出し、高い精度の多クラス分類を実現した。さらに、従来の研究ではテキストとしてそれぞれの画像に付せられた解説文（figure legend）のみを利用するものが多かったが、本研究では解説文のみを用いた場合に加え、解説文と表題・要旨を用いた場合、さらに解説文・表題・要旨と本文を用いた場合を比較し、後者ほど精度が向上することを見いだした。

個々の単語の集合だけでなく 2 単語の組を考慮した分類手法は本研究が初めてである。図に直接付属した解説文に加えて本文をも考慮に入れる手法は以前にも発表されているが、本研究は解説文のみ、表題・要旨を加えた場合、さらに本文をも加えた場合の分類精度を体系的に比較し、精密な検討を行った初めての研究である。条件を様々に変えた解析はきわめて計画的に行われており、分類精度が向上した場合だけでなく向上が見られなかった場合についても、その原因を議論して、納得できる説明を行っている。この意味で、本研究はオリジナルの研究として高く評価できるものである。

次に本研究は、「ある画像がどういう部位・特徴・作業課題・疾患・遺伝子などに関連した内容を示しているのか」をデータベースにキータームとして記載するための、アノテーション作業を自動化する方法を開発した。そのためには、その画像を掲載した論文を特徴付ける単語を自動抽出する必要がある。そこで、グラフ理論に基づいて単語の相互関係を評価するテキストランク法を利用して、単語の「タームらしさ」を評価する方法を構築した。これによって、ある論文と対照群の文章群データ（コーパス）を比較し

てキーターム抽出する従来提唱されていた手法に比べ、大きな改善が見られた。さらに、一般的な単語や重要な情報を持たない単語など適切でないキータームが抽出されてしまう現象を極力抑えるため、さまざまなフィルタリング手法を検討した。まず、脳科学分野の論文とそれ以外の生物分野の論文のコーパスを比較して前者にのみ多く登場する単語を抽出することにより、生物科学一般に用いられる用語を効率よく排除した。さらに、それぞれの論文で鍵となる用語は出現頻度が非常に高い傾向があるので、頻度分布を利用したフィルタリングによって特例的な用語を排除した。さらに、脳科学分野の多くの論文に共通して現れる単語は特徴性が低いと見なせるため、これら分野一般的な用語を頻度分布に基づいて排除した。これら種々の改善によって、従来よりも大幅に効率的なキーターム抽出を実現した。

この研究はテキストランク法の有効性を実証した点で有意義であると同時に、各種のフィルタリング技法の考案は深い洞察に裏打ちされたもので高く評価できる。特に、ある学術分野の論文とそれ以外の分野の論文のコーパスを比較する手法は効果が大きく、この有用性を実証したことの意義は大きい。本研究で開発されたキーターム抽出手法は、脳科学だけでなくさまざまな分野でも実際に使うことが可能な汎用的なものであり、今後のさらなる発展が期待される。

本研究は最後に、開発した手法を用いて実際に論文画像の自動クラス分けとキーターム抽出を行った情報を搭載したデータベースを作成し、公開した。これは現時点ではまだデータ数が多くなく、網羅的な検索ができる状態には達していないが、本研究で提唱された情報抽出自動技術が実用的に利用できるレベルのものであることを実証した。今後、元データとなる論文データから図版や解説文情報を高精度に認識抽出する技術の改善が進めば、本研究の究極目的である大規模な俯瞰データベースの実現に大きく寄与するであろう。

以上のように本論文は、情報の自動抽出の基本的手法を開発し、実用性を検証した意義の高い研究であると認めることができる。なお、本論文第 1 部は小池麻子・山本泰智・高木利久、第 2 部は小池麻子・高木利久との共同研究であるが、論文提出者が主体となって分析及び検証を行ったもので、論文提出者の寄与が十分であると判断する。

したがって、博士（科学）の学位を授与できると認める。