平成24年度　情報工学コース卒業研究報告要旨

長尾 研究室	氏　　名	井　上　慧
卒業研究題目	学術論文からの研究資源情報の自動抽出に関する研究	

研究活動には研究資源の利用が不可欠である。研究で用いられるデータや資料、ツールに加え、既存手法や評価法などの無形のものも広義の研究資源と捉えることができる。これまでに膨大な量の研究資源が構築され、その多くは作成者によって公開・提供されており、多種多様な用途に再利用することができる。このように研究資源を共有し、様々な用途に活用することは、研究者各々の効率的な研究活動を促すとともに、研究分野全体のレベル向上につながる。しかし、現状では既存の研究資源が有効に活用されているとは言い難い。その理由として、研究資源に関する情報が十分に共有されていないため、利用者が自らの目的に合致した研究資源を容易に見つけることができないという点が挙げられる。

そこで、本論文では、研究資源の名称や定義・用途のような、研究資源を検索・選択する際に有用となる情報（以下、**研究資源情報**）を網羅的に取得し提供することを目的に、学術論文からの研究資源情報の自動抽出手法を提案する。学術論文は、一般的なWebテキストと比べて、文章が整っており、記述内容の専門性・信頼性も高いため、有用な情報を精度よく抽出できる可能性がある。

学術論文の本文には、研究資源の名称（以下、**研究資源名**）のみならず、研究資源がどのようなものかを示す表現（以下、**定義表現**）、どのような目的・用途で使用されたのかを示す表現（以下、**用途表現**）などの研究資源情報が含まれている。特に、定義表現および用途表現の中には、利用者の研究活動によって新たに創出された用途も多く存在する。そのため、研究資源名とともに定義表現および用途表現を提供することは、研究資源の広範な利用に大きく貢献すると考えられる。

本研究では、研究資源名と定義表現および用途表現が文内で共起することに着目した。50論文を用いた予備調査によって、定義表現および用途表現の96.8%が研究資源名を含む文中に現れていることを確認した。また、研究資源名と、定義表現あるいは用途表現の間には、研究資源情報抽出の手がかりとなる表現（以下、**手がかり表現**）がよく現れることを確認した。

提案手法では、学術論文の分析によって得られた、(1) 研究資源名の形態素的特徴、(2) 定義表現および用途表現の持つ特徴、(3) 手がかり表現を利用して、学術論文の本文テキストから研究資源名の候補となる形態素列を取得し、SVMを用いて研究資源名を抽出する。研究資源名候補の抽出例を図1に示す。図1の場合は、既知の用途表現の末尾の動詞「構築」と係り受け関係にある手がかり表現「を用いて」を利用して、「を用いて」と連接する「不完備情報ゲーム」を研究資源名候補として抽出する。定義表現および用途表現の抽出には、手がかり表現および研究資源名を利用する。定義表現の抽出例を図2に示す。図2の場合は、手がかり表現「である」を含む文節を根とした部分木から「高次元データの類似性を可視化する手法」を定義表現として抽出する。

また、提案手法を評価するために、研究資源名の抽出実験と定義表現および用途表現の抽出実験を個別に実施した。実験には、2008年度人工知能学会全国大会の講演論文集に収録されている150論文を使用した。研究資源名の抽出実験の結果から、研究資源名の抽出において、既知の定義表現および用途表現の持つ特徴を活用することの有効性が確認できた。また、定義表現および用途表現の抽出実験の結果から、定義表現および用途表現の抽出において、手がかり表現を活用することの有効性が確認できた。

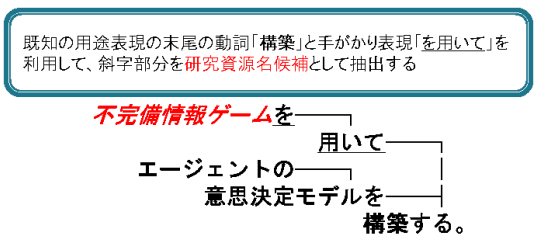

図1：研究資源名候補の抽出例

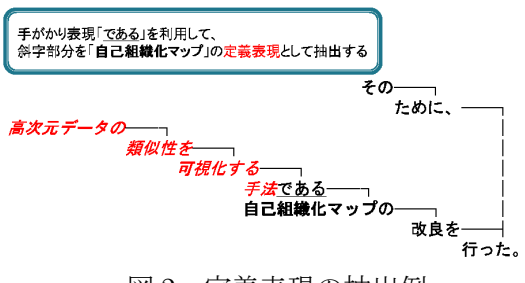

図2：定義表現の抽出例