テキストからの複合指標の計算式抽出

井出智志

本論文では，テキストからの複合指標の計算式抽出に取り組む．複合指標とは，数値を直接計測することはできないが，他の指標の数値に基づいて計算することができる指標と定義され，この複合指標を含む質問に解答することができないという問題を解決する一つのアイデアとして，複合指標の計算式を獲得することを提案する．なんらかの形で指標の計算式を獲得することで，その指標の計算方法を理解できるようになるため，質問に回答することが可能となる．

ここで本研究では複合指標の計算式を獲得するために，テキストから複合指標の計算式を抽出するアプローチをとる．より具体的には，Wikipedia から複合指標を計算する式を抽出する問題に取り組む．例えば，Wikipedia のページから「BMI = 体重 kg÷(身長 m)^2」を抽出する．方法としては，Wikipedia をデータセットに，様々な正規表現を用いて，計算式を抽出し，その後 = や四則演算 $(+ - \times \div)$ など，複合指標となりうる最低上限を満たした計算式のみをルールベースで抽出する．そして機械学習の分類器 SVM を用いて，抽出した計算式のうち，複合指標の計算式のみを獲得する．機械学習を行うための特徴量として，今回は計算式，関連文それぞれから 3 つずつ特徴量を抽出した．これらをもとに数式と関連文から，特徴量を抽出し,．その後分類に用いるデータセットを作成する．実験に用いる計算式と関連文のペアにそれぞれにラベル (複合指標の計算式は 1, それ以外の式には 0) を付与する．そのデータセットを用いて、SVM を用いて機械学習を行い，その計算式が，複合指標の計算式か否かを予測する．また追加実験として，得られた複合指標の計算式から，計算式を構成する変数抽出を行う．例えば，計算式「防御率 = 自責点 ×9÷ 投球回」から，その式を構成する指標「“防御率”，“自責点”，“投球回”」を抽出することを目的とする．

実験では, 数値に関する質問応答を向上させるため, Web 上の多数の知識を獲得するための知識抽出を行なっている. そのため今回は幅広く多数のデータが含まれている Wikipedia を用いる．その後 Wikipedia 全体から特定のセクション (定義セクション，概要セクション計算式セクション，○○の計算式を抽出) のみを取り出す　データセットとして利用する．ここで特定のセクションのみを実験に用いる理由は，ほとんどの計算式がこれらのセクションに集中してたためである．計算式抽出では，定義セクションから数多く計算式を抽出することができ、セクション全体では高い適合率を示した．その後，複合指標の計算式分類，変数抽出をおこなったが, それぞれで高い適合率を示した．

（指導教員　　加藤　誠）