平成 17 年度　卒業論文概要

入学年度	平成 14 年度	学生番号	14117633	氏名	浦澤 真平

論文題目 関係的知識発見手法を用いた英文からの特徴の抽出

和田・犬塚　**研究室**

1. はじめに

データマイニングとは，大量のデータから隠された知識や新しい規則を発見するプロセスである．データマイニングの手法の一つとして，相関ルールの考え方を用いた**MAPIX**(Mining Algorithm by Property Item eXtraction)が開発された．MAPIX は事例の性質に注目する事で可読性の高いルールを導出でき，その際全ての事例を用いなくとも十分なルールの導出が可能であるという特徴を持つ．しかし，現在のところ MAPIX が実データに応用された例はない．そこで本研究では英文からの特徴の抽出に MAPIX を用いる事で，実データに対しても有効な手法であることを示す．

2．MAPIX

図 1．家族関係

簡単な例として図 1 の家族関係について考える．ここで hiroshi に注目すると，例えば次のような事実が成り立つ．

parent(hiroshi，koji) ∧ male(koji)
parent(hiroshi，koji) ∧ parent(koji，yoichi) ∧ male(yoichi)

上記の事実は hiroshi には koichi という息子をもつことと，yoichi という男の孫をもつことを表している．述語の各引数には入力引数なのか，出力引数なのかを表すモードが定められており，それぞれ+，-で表す．この例で，各述語は parent(+，-)，male(+) である．

モードに注目すると述語は二つのクラスに分けられる．第一のクラスは全ての引数が入力引数である**判定述語**という．この述語は事例に関する具体的な事実を表す．第二のクラスは入力引数と出力引数の両方をもつ述語で**経路述語**という．先程の例のように，いくつかの経路リテラルによって，注目する対象から判定述語まで鎖状つながってる述語の組を**性質**という．

また，性質の述語の引数を変数に置き換えたものを**性質アイテム**という．次に上記に示した性質に対する性質アイテムの例を示す．

parent(A，B) ∧ male(B)
parent(A，B) ∧ parent(B，C) ∧ male(C)

MAPIX の概要を次に与える

1. 与えられた事例の集合からいくつかの事例を選択する
2. 選んだ事例から得られる性質アイテムを全て抽出する
3. すべての性質アイテムを用いて，閾値以上の頻度を持つアイテムセットを興味深いものとして枚挙する

3．英文の論理表現

英文にはペンシルバニア大学の Penn Treebank Project によってタグ付けされた構文木を使用．構文木には次の二つのタグが付けられている．第一のタグは，名詞句や動詞句などの英文の構造を表す構文タグ．第二のタグは各単語の品詞を表す品詞タグである．次に構文木の例を示す．

図 2．構文木の例

図 2 に示した構文木の各タグを述語として，英文を論理表現した．構文タグからなる述語は経路述語となり，品詞タグからなる述語は判定述語となる．よって，英文から抽出される性質は，英文の頭から品詞タグからなる述語までが，構文タグからなる述語でつながっている述語の組である．

4．実験

述語表現を行った英文を 300 例用意し以下の手順で実験を行った．

1. 閾値を 1 %，2 %，5 %，7 %に設定
2. 各閾値において 300 例すべてを使用したときに抽出されるアイテムセット数を記録
3. 使用する事例の数を 0 から徐々に増やして生き，抽出されるアイテムセットの数を記録する

図 3．抽出に使用した事例数に対するアイテム数

図 2 より閾値を下げることにより，アイテムセットの数が飽和するのに必要な事例数が増えている事がわかる．しかし，閾値の一番低い 1 %でも，全体の半分の事例も使用すればほぼ全てのアイテムセットが抽出している．このことから MAPIX が英文のようなデータに対しても有効な手法であることが確認された．

5．まとめと今後の課題

MAPIX が英文のようなデータに対しても有効な手法である事が確認された．

現在は頻度しかみておらず，興味深いルールのみが抽出されているとはいいがたい．よって今後の課題として，別のアプローチによる導出方法を考えたい．