Mining Context-level Associations in Documents Collections using Passages

永井 健太郎 (350047)

北陸先端科学技術大学院大学 知識科学研究科

2005 年 2 月 10 日

キーワード: Association Rules, Passage, Context, Rough Sets, Text Mining.

本研究の目的は, 文書集合を対象とする相関ルールマイニングにおいて, ルール中の各項に文脈を反映するパッセージを用いることにより, ルールのわかり易さを高めることである.

従来のテキストルールマイニングでは, 単語または複合名詞などを単位として, それらの共起関係を抽出している. しかし, 単語などの表層的な単位を利用することには以下のような欠点があげられる.

- 語の意味が不明瞭になりやすい
- どのような文脈でそれぞれの語が利用されているのかを考慮しない
- 構文情報になりやすい

この欠点は出力されるルールにも影響する. 相関ルールには冗長なルールを生成するという問題が知られている. これに加え, 文書の場合, 意味が不明確な, もしくは文脈が不明確なままの語をルールに用いていることにより, この問題を助長している可能性が考えられる.

そこで本研究では, ルールの各項の意味を正確に把握することを支援するため, 文脈を単位とするテキストルールマイニングを提案する. 文脈を単語の集合として表現する. これにより, 周辺の単語の存在からそれぞれの単語の意味が明確になること, そして文脈というより意味的な単位を用いることにより, 有用なルールが抽出されることが期待できる.

文脈を単位として相関ルールを利用するにあたり, 以下の課題を解消しなければならない.

1. 文脈の抽出 (文書のどこで文脈が変わるのか)
2. 文脈の表現 (文脈は計算機上でどのように格納されるか)
3. 文脈の同定 (どのような時に, 2 つ以上の文脈が同内容であると言えるのか)

文脈の抽出に関しては, 自動抽出の研究が行なわれているが, ここでは問題を単純化し, 段落をパッセージ抽出の単位とした. パッセージ内のキーワードに利用される単語の微妙な表現の差を吸収するため, トレランス・ラフ集合に基づいた語彙の補充を行なった. これは, 通常こうした目的に対しては, 概念辞書を用いることが多いが, これは固有名詞が多いキーワードには適さないと考えられるためである. 最後の課題に関しては, 分布仮説 (distribution hypothesis) 「文脈が類似する場合, そこに出現する単語の意味も類似する」を拡大解釈し, 「二つの文脈が同じような内容を示す場合, それらの文脈は一定以上の単語を共有する」という仮説を立て, 集合操作やベクトル空間モデルを用いて, 閾値以上の類似度を条件とし, 文脈の同定を行なう.

以上をふまえた, 本手法の具体的な処理手順は以下の通りである

1. 文脈の抽出
2. 文脈毎のキーワードの抽出
3. 類似文脈行列の計算
4. 類似文脈行列のトランザクション・データベース形式への変換
5. 相関ルールアルゴリズムによるルール抽出

本手法の特性を明らかにするため, パッセージを単位としたルールの抽出を行ない, ルールのわかりやすさに関する定性的評価, およびルール内の語の意味が不明瞭でないかを定量的に評価した. 抽出されたルールの中には, 従来の複合語に捕らわれない単語の組合せが見られ, 表現の自由度が高まっていることが確認できた. また定量的な評価の手法に関しては, ルールに出現する語の取り得る意味の数が, 同一パッセージ内の語を対象に語意曖昧性解消を行なう前と後で, どれほどの差が出るのかををルールのエントロピーという形で数量化し, 比較を行なった. 小規模の比較ではあるが, 本手法によりエントロピーが半分以下に減少することを確認した. 双方の評価より総じてルールのわかりやすさは向上していると考える.

ルールのより詳細な定性的な評価は今後の課題とするが, 分析した範囲では, 意味を把握できない文脈やルールが多数見られた. キーワードの抽出手法の改善, または意味をより正確に反映した文脈の抽出手法の開発などにより, このような文脈の削減が必要である. また, 計算困難性は文脈の数の２乗となっており, 拡張性, 効率性の面での課題も残している.

本研究では, 文脈を利用することで理解の容易な, そして文書の持つ意味的な単位間の相関を抽出することを提案し, また, 理解の容易さを測る定量的な指標として, ルールの持つ意味的な曖昧性という尺度を導入した. 今後定量的, 定性的な面からの分析・ 評価を行なっていく必要がある.