

Web上のニュース記事の分類における知識コンバージョンを用いた例外処理

3J-6

小山 誠 桂田 浩一 大原 剛三 馬場口 登 北橋 忠宏

大阪大学 産業科学研究所

1 はじめに

近年、インターネットの普及に伴い個人が入手できる電子化文書の量は増加しており、これらの中から個人にとって必要なものを取り出し、分類・整理するという要求が高まっている。こうした分類作業を自動化する方法の一つとして、個人の分類基準を分類ルールとして形式化し利用する事が考えられる。ルールの表現形式としては一階述語論理が有用であるが、一階述語論理に基づくルールは、常に成り立つルール（完全ルール）として形式化されるため、例外を含み得る個人の主観的な記事の分類基準を完全ルールとして形式化すると、それらのルールから例外に関して誤った分類結果が導かれる。

こうした例外による誤りを防ぐため、筆者らは、例外に関して誤った結論を導く完全ルールを、常に成り立つとは限らないルール（不完全ルール）へと変換する知識コンバージョン（Knowledge Conversion: 以下KC）[1][2]について検討してきた。本稿では、Web上のニュース記事を自動分類するシステムを具体例として取り上げ、上述の問題を解決するために、ユーザによる分類の変更に基づく、KCを用いた例外処理を提案する。

2 Web上のニュース記事分類システム

本稿で対象とするニュース記事分類システムは、入力としてニュース記事を受け取り、それらの見出しから抽出されるキーワードに基づき各記事を1つ以上のカテゴリに分類し、電子スクラップブックを作成する。システムの構成は図1に示すように、分類ルール、背景知識を蓄積する知識ベースと、ルール学習部、ルール修正部、表示部、キーワード抽出部などから構成され、システムでの処理は大きく分けて、初期フェーズ、分類フェーズ、例外処理フェーズからなる。

まず、初期フェーズでは、初期カテゴリが与えられた学習用記事とその見出しから抽出されるキーワード、カテゴリ間の階層構造を表すルール等を背景知識として用いて、ルール学習部において以下のような形式の初期ルールを生成する。ここで、 x は記事を表す変数である。

$$\text{犯罪}(x) \leftarrow \text{has_逮捕}(x), \text{has_人}(x)$$

このルールは、“逮捕”および“人”というキーワードを含む記事は“犯罪”というカテゴリに割り当てるという意味である。尚、学習アルゴリズムとしては帰納論理プログラミング[3]を用いている。

次に、分類フェーズでは、キーワード抽出部において記事の見出しから抽出されるキーワードと、知識ベース

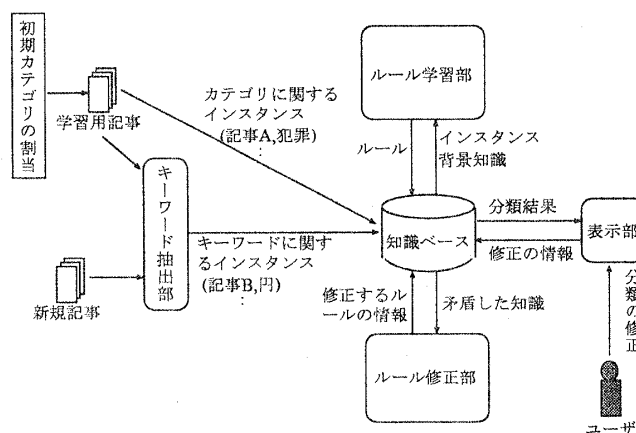


図1: システム構成

中の分類ルールを用いて、新規記事を適当なカテゴリに分類する。分類の結果は、表示部においてユーザに提示され、必要に応じてユーザは分類の修正、新規カテゴリの追加等を行う。

例外処理フェーズでは、ユーザによる分類の変更があった場合に、あるカテゴリへの分類を取り消された記事を、当該カテゴリの分類ルールの例外記事として知識ベースを修正することにより、例外記事の誤分類解消と、分類精度の向上を図る。

3 知識コンバージョンを用いた例外処理

例外処理フェーズでは、具体的にはユーザによる分類の変更に基づき、1) ルール修正部において、誤った分類を与えるルールを例外を許容するルールに変換し、2) ルール学習部において、修正のあった記事を例外として例外に関するルールを学習する。本章では、まず分類を取り消された記事に関する分類の否定情報の取得について述べ、その後、例外処理のため各手続きを説明する。

尚、本研究では知識ベース中におけるルール、事実を表すインスタンスを図2に示されるようなADAG[4]というグラフで表現する。ここで、トークンはルール、インスタンスから導かれる推論結果を表している。

3.1 否定情報の取得

システムは、ユーザによる分類の変更によってあるカテゴリへの例外記事の分類が取り消されたという否定情報を知識ベース中に保持する。具体的には、当該カテゴリを表すノードに対して排反なノードを知識ベース中にもち、そのノードに例外記事をインスタンスとして追加する。例えば、「キーワード“従業員”をもつならばカテゴリ“雇用問題”に分類する」というルールによって記事 A_1 がカテゴリ“雇用問題”に割り当てられたとする。これに対し、ユーザが A_1 のカテゴリを“事件”に修正

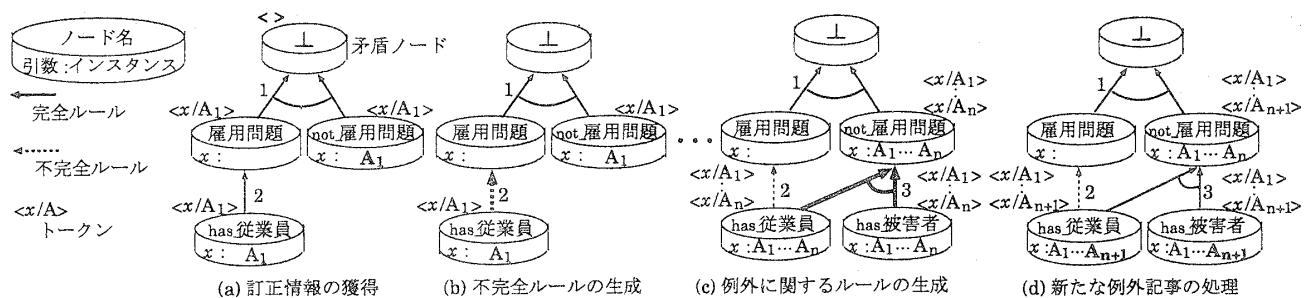


図 2: 例外処理フェーズ

したとすると、システムは、ノード“雇用問題”に対してそれと排反なノード“not_雇用問題”に A_1 のインスタンスを追加する。図 2(a) は、この例において、ノード“not_雇用問題”に A_1 のインスタンスが追加された後の状態を表している。

ここで、ADAG での推論は、インスタンス毎に生成されるトークンをルール（始点）から結論部（終点）に伝播させて行われる [4]。各ルールにおけるトークンの伝播規則は次のようになる。ただし、ADAG の矛盾はノード $\perp$ にトークンが伝播する状態と定義される。

- 完全ルールの伝播規則: トークンがルールの条件部を満たせば、必ず結論部に伝播させる。
- 不完全ルールの伝播規則: トークンがルールの条件部を満たし、かつ結論部に伝播させても ADAG が無矛盾な場合のみ伝播させる

3.2 知識コンバージョンによる不完全ルールの生成

先の例において、ノード“not_雇用問題”にインスタンスが追加されると、このインスタンスから生成されるノード“not_雇用問題”におけるトークンと既存のルール 2 から導かれるノード“雇用問題”におけるトークンにより矛盾が生じる。

このような矛盾を契機とするルール修正部では、誤った分類結果となるトークンが伝播されるのを防ぐため KC を実行する。KC では、矛盾を引き起こしたルール集合の中から例外を含む可能性が最も高いルールを検出し、それを不完全ルールに変換する。このルールの検出のため、矛盾を導いた各トークンの真偽を次のように想定する。

- 対応するインスタンスが知識ベース中に存在するトークンは真である。
- 対応するインスタンスが知識ベース中に存在しないトークンは偽である。

この下で、矛盾を導いたトークンに対して条件部が真、かつ結論部が偽となるルールを不完全化するルールとして選択する。

図 2(a) では、ノード“has_従業員”、“not_雇用問題”におけるトークンが真、ノード“雇用問題”におけるトークンが偽とみなされることから、ルール 2 が変換対象ルールとして選択され、図 2(b) のように不完全ルールに変換される。この結果、記事 A_1 のトークンはノード“雇用問題”へ伝播されず、カテゴリ“雇用問題”への誤分類は解消される。

3.3 例外に関するルールの生成

不完全ルールを生成しただけでは、“not_雇用問題 (A_1)”というインスタンスが与えられた例外記事 A_1 の誤分類は防がれるが、このような否定情報を持たない新たな例外記事の誤分類は防がれない。こうした新たな記事の誤分類を防ぐため、システムは、ノード“not_雇用問題”中に蓄えられた例外記事のインスタンスの数がある値を越えたとき、これらを用いて、ノード“not_雇用問題”を結論部とする例外に関するルールをルール学習部において学習する。

例えば、先の例では、 A_1 と同様にカテゴリ“雇用問題”への分類を修正された記事 $A_2, \dots, A_n$ のインスタンスがノード“not_雇用問題”に追加された時点で、それらの記事を用いて、ノード“not_雇用問題”に関するルールが学習される。図 2(c) は学習結果として、「キーワード“従業員”をもち、かつキーワード“被害者”をもつならばカテゴリ“not_雇用問題”に分類する」というルール 3 が学習された状態を示す。ルール 3 は完全ルールであるので、以後、図 2(d) のように、 $A_1, \dots, A_n$ と同様に“従業員”と共に“被害者”というキーワードをもつ記事 A_{n+1} が追加されても、ノード“not_雇用問題”におけるトークンが優先的に導かれ、ノード“雇用問題”におけるトークンが導かれないためカテゴリ“雇用問題”への誤分類は防ぐことができる。

4 まとめ

本稿では Web 上のニュース記事分類システムを対象に、KC を用いた例外記事の処理について述べた。この例外処理により、既存の分類ルールに従わない例外記事が知識ベースに追加される場合にも、分類精度の低下を軽減することができると考えられる。システムの実装および実験による評価が今後の課題である。

参考文献

- [1] 馬場口 登: 完全/不完全知識を含むデータベースにおける知識獲得, 大須賀節雄ほか編, 知識科学の展開, オーム社, pp.134-142(1996).
- [2] 小山, 桂田, 大原, 馬場口, 北橋: 完全/不完全ルールの属性変換による矛盾解消のための非対話的手法, 1998 年度人工知能学会全国大会論文集, pp.28-31(1998).
- [3] S.Muggleton: Inductive Logic Programming, Academic Press(1992).
- [4] 桂田, 大原, 馬場口, 北橋: 不完全知識の例外に関する観点変更による知識ベース再構成, 人工知能学会誌, Vol.14, No.3, pp.485-494(1999).