研究支援システム Papits における 分類機構を用いた論文収集システムの試作

長谷川 友治† 大囿 忠親†† 新谷 虎松††
Tomoharu HASEGAWA Tadachika OZONO Toramatsu SHINTANI
† 名古屋工業大学大学院 工学研究科 †† 名古屋工業大学 知能情報システム学科

1 はじめに

我々は、研究支援システム Papits の開発を行っている。Papits は、組織内で取り扱うファイル、中でも論文ファイルを集約、共有することで、より良い知識の獲得を目的としている。[Fujimaki et al., 02]

現在、Papits には、研究室のメンバーによって登録された論文だけでなく Web から自動的に論文を集めて登録する論文収集機能が備わっており、大量の論文が登録されている。これらの論文は，主に研究のサーベイなどに利用されている。このとき自分の研究する分野に関係のある論文を探すことになるが、そのためには論文が分野ごとに分類されていることが望ましい。そこで、論文を自動的に定められたカテゴリへ分類を行う機構が必要となる。

そこで、本論文では，Papits における論文収集時の論文分類機構の実装について述べる。そして、Papits において用いた複数のカテゴリにテキスト分類を行うための属性選択手法を提案し、その有効性を示す。

2 Papits の実装

Papits は，WebObjects を用いた Web アプリケーションとして実装され，Web ブラウザを用いて利用する．実際の Papits のインターフェイスは図 1 のようになっている。

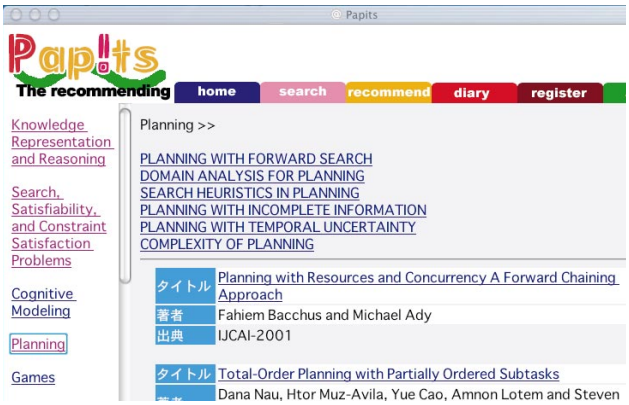

図 1: Papits のインターフェース

論文を閲覧する際には図 1 のようにカテゴリが表示される。このカテゴリは階層構造をとっており、カテゴリ名をクリックすることで、そのカテゴリに所属する論文の一覧とさらに下位のカテゴリを構造を表示する。この下位のカテゴリをたどっていくことで、自分の興味にあった論文を簡単に探し出すことができる。また、閲覧するときには先ほど述べたカテゴリの修正を含めた登録論文情報の修正を行うことができる。

2.1 論文分類支援機構

論文分類支援は図 2 のような機構で行われていく。

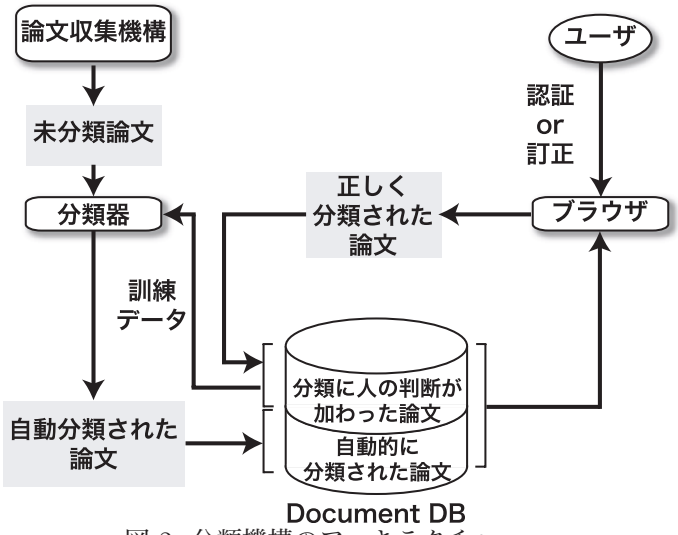

図 2: 分類機構のアーキテクチャ

まず論文登録が行われる。Papits には Web からの論文を自律的に収集して登録する論文収集エージェントが実装されており、データベースには自動的に論文が蓄積されていく。ただし、これらの論文はカテゴリが与えられていないため、分類器によって分類を与える必要がある。ここで、ユーザが分類した論文は正しく分類されたデータであると考えて訓練データとして用いる。その訓練データを使用して分類器が未分類の論文を分類する。機械的に分類器が分類したデータについては本当に正しく分類されていると判断できないため、これらのデータは訓練データとしては使用しない。

ユーザはこのように分類された論文をデータベースから論文を閲覧することができる。その際に、閲覧した論文について、現在属しているカテゴリが正しいことを認証したり、正しいカテゴリに修正したりすることができる。これによって、その論文の属するカテゴリが正しく決定されたと見なし、新しい訓練データとして加えることで分類器の精度が次第に上昇していくことが期待できる。

3 Papits におけるテキスト分類

ここで、Papits において用いたテキスト分類と属性選択の手法を説明する。Papits においては、論文のタイトル、著者名、アブストラクトをその論文を表すデータとして考え、それらの情報から論文を分類する。

3.1 テキスト分類手法

本システムでは、分類手法に KNN(K-Nearest Neighbor:K-最近傍法) を用いて論文のカテゴリ分類を行う。この KNN は分類したいデータと訓練データとの類似度（距離）を計算し、分類したいデータに似ている訓練データの上位 K 個の分類を求める。その K 個のデータの分類を用いた投票を行い、カテゴリを決定する。本システムにおいては、互いのデータにおける出現した単語の有無で類似度を計算する。

Paper Gathering System using Classifier on Research Support System Papits

† 名古屋工業大学大学院工学研究科, Graduate School of Engineering, Nagoya Institute of Technology.
†† 名古屋工業大学知能情報システム学科, Dept. of Intelligence and Computer Science, Nagoya Institute of Technology.

表 1: 分類精度 (%)

	属性数									
	100		200		300		400		500	
	top1	top3	top1	top3	top1	top3	top1	top3	top1	top3
情報利得のみ	51.87	73.79	55.08	77.54	58.82	78.07	52.94	74.33	48.12	75.40
提案手法	55.08	79.67	57.21	80.74	57.21	81.81	59.89	81.81	57.75	82.88

- V = 情報利得の値でソートした語の集合
  (初期状態は空集合)
- D = ドキュメントの集合
- C = カテゴリの集合
- k = 任意の属性数
- $IG_{C_1,C_2}(w,D)$: カテゴリ C_1,C_2 に対する情報利得

Feature Selection
- **FOR** すべての語 w
  - **FOR** C の要素 2 個以下のすべての組合わせの集合 C_1
    - $C_2 := C$ - C_1
    - $IGvalue := IG_{C_1,C_2}(w,D)$
    - **IF** $(max < IGvalue)$ **THEN**
      - $max = IGvalue$
  - 語 w を集合 V に max の値でソートして追加する
- V の中で最も評価が高い k 個の語を属性として選択する

図 3: 特徴選択アルゴリズム

3.2 特徴選択手法

KNN を含めた多くのテキスト分類手法では、分類に関係のない単語を多く評価すると過学習が起りやすく計算時間も増大するという問題がある。そこで特徴選択手法を用いることで分類に影響すると思われる単語を選択する方法が行われる。情報利得を用いた特徴選択手法が一般的には良く行われている [Pascal and Minequ, 01]。しかし、十分な論文数が Papits に登録されていない状態で多くのカテゴリに分類しようとする場合、1 つカテゴリに属するデータが少なくなり、そのカテゴリを特徴付ける単語を選択できない。

そこで、カテゴリを 2 値変換しそのときの情報利得を計算することで属性の分類に対する重要度を計算する手法を提案する。カテゴリを 2 値にすることで、1 つのカテゴリに属するデータ数が増えるため分類に影響すると思われる単語を選択できると考えられる。

そのアルゴリズムを図 3 に示す。まず始めに、すべてのカテゴリの中で、2 個以下の組合わせのカテゴリの集合 C_1 とその C_1 以外のカテゴリの集合 C_2 を考え、これらを新しいカテゴリの形式とする。そして、出現したそれぞれの語 w についてこのカテゴリ C_1 と C_2 に対する情報利得を求め、語 w について順位付け行う。最後に、上位 k 個の語を分類に用いる属性として選択する。

4 評価実験

4.1 実験方法

ここで本システムの分類精度についての評価を行う。本評価に用いた論文データは、IJCAI'01 の予稿集に収録された 188 論文である。これらの論文のタイトル、著者名、アブストラクトの情報を用いてテキスト分類を行う。ここで、論文は予稿集のいずれか 1 セクションに割当てられているので、このセクションを論文のカテゴリとした。これらのデータに対して、提案手法と情報利得によって属性の重要度を測る方法とを比較した。この情報利得による方法は一般的な属性選択手法であるとされており [Pascal and Minequ, 01]、すべての語の情報利得を求め、最も良い語 k 個を属性として選択する方法である。

精度の基準として 2 通り評価した。1 つ目の基準 (top1) は対象となるデータのカテゴリと KNN が最も適しているだろうと予測した 1 つのカテゴリが一致した場合を正解として精度を求めた。もう一つの基準 (top3) は対象となるデータのカテゴリと KNN が最も適していると予測した上位 3 つのカテゴリの内の一つでも一致した場合を正解として精度を求めた。この精度の基準は、Papits においてユーザにある程度適すると思われるカテゴリが示せることが重要であると考えたため採用した。

それぞれの手法に対して、leave-one-out cross validation を用いて精度を測った。これは、全データ 188 の内の 1 つをテストデータに、残り全部を訓練データとして使用する組合わせすべてを行った場合の精度の平均を求める評価方法である。

4.2 実験結果

実験結果を表 1 に示す。この結果を見るとほとんどの場合で、単純に情報利得を用いるだけの場合よりも分類精度が高いことが分かる。また、top3 の基準においては、提案手法の選択した属性数の方がより少ない場合でも高い精度となっている。これは、評価する属性数を減らしても高い分類精度を維持した状態で、分類にかかる計算コストを下げることが可能であることを表している。

5 おわりに

本論文では，研究支援システム Papits のための論文分類支援機構について述べた．この論文分類支援機構はすでに人手により分類された論文を訓練データとして利用し、KNN 手法を用いた分類を行っている。その際、複数カテゴリへの分類に対応させた属性選択手法を用いることでより精度の高い分類が可能となった。論文分類機構は、自動的に収集されたような未分類の大量の論文を人間が見ることなくある程度の分類を行うことができる。そのため、大量の論文から自分の興味のある論文をカテゴリをたどって探し出すことができるようになる。閲覧した論文に関しては、人間に分類を承認・修正してもらうことにより、その論文を新たな訓練データとして利用ことで自動分類の精度をあげることができると考えられる。

参考文献

[Fujimaki et al., 02] Nobuhiro Fujimaki, Tadachika Ozono, and Toramatsu Shintani : Flexible Query Modifier For Research Support System Papits, Proceedngs of the IASTED International Conference on Artificial and Computational Intelligence, pp.142-147, ACI2002, 2002.

[Pascal and Minequ, 01] Pascal Soucy and Guy W. Minequ : A Simple Feature Selection Method for Text Classification, In the Proceedings Seventeenth of the International Joint Conference on Artificial Intelligence (IJCAI01), pp.897-902, 2001.