3B4-06

研究支援システム Papits における 論文分類機構のための特徴選択

Feature Selection for Paper classifier in Research Support System Papits

長谷川 友治[*1] 大囿 忠親[*2] 新谷 虎松[*2]
Tomoharu Hasegawa Tadachika Ozono Toramatsu Shintani

[*1]名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

[*2]名古屋工業大学知能情報システム学科
Department of Electrical and Computer Engineering, Nagoya Institute of Technology

We have developed a research activity support system , Papits. Users of Papits can share various research information, such as PDF files of research papers. Now, Papits is equipped not only with papers registered by a member of a laboratory but with the paper collection function to collect and register papers automatically from the Internet. Therefore, a lot of papers are registered, and these papers are mainly used for the survey of research. In that case, it is desirable to classify papers for every field.

In this paper, we describe implementation of a paper classification mechanism, and propose a feature selection for classifying documents that is represented by bag-of-words.

1. はじめに

近年，組織内での知識共有，知識経営に注目が集まっている．これは組織内のデータを知識として共有し，その知識を必要なメンバーが効率良く再利用することでより知的生産性が向上すると考えられている．我々の研究室では，研究室内の活動を向上させるためのシステムとして研究支援システム 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][Ozono 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インターネット上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．しかし，インターネット上から機械的に集められた大量の論文をひとつひとつ閲覧し，分類することは不可能である．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．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と自動分類の精度が悪くなるといった問題が発生する．

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特徴選択手法を提案する．この手法は，属性の重要度をカテゴリを 2 値変換した上で情報利得を計算する．これは，一つのカテゴリに含まれるデータ数が増加した状態にできるため，データが少ないことで発生する過学習を防ぐことができると考えられる．最後にこの手法による論文分類の実験を行い，有効性を確認する．

連絡先: 長谷川 友治，名古屋工業大学大学院知能情報システム工学科新谷研究室，〒 466-8555 名古屋市昭和区御器所町，電話 052-744-3153，FAX 052-735-5477，tomoha@ics.nitech.ac.jp

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．

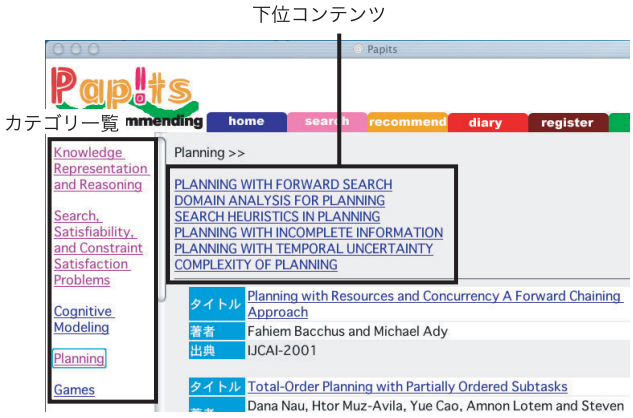

図 1: 閲覧部分のインターフェース

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．

まず論文登録が行われる．Papits にはインターネット上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

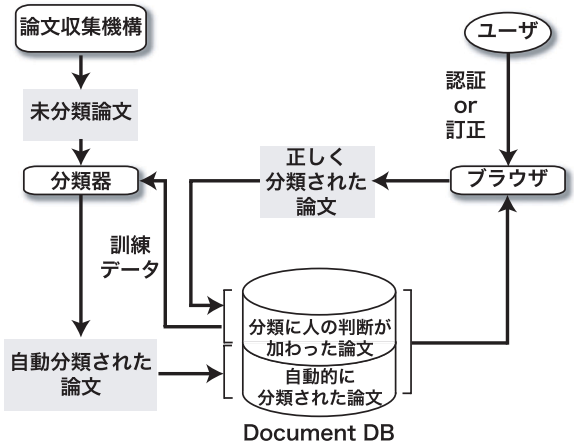

図 2: 分類機構のアーキテクチャ

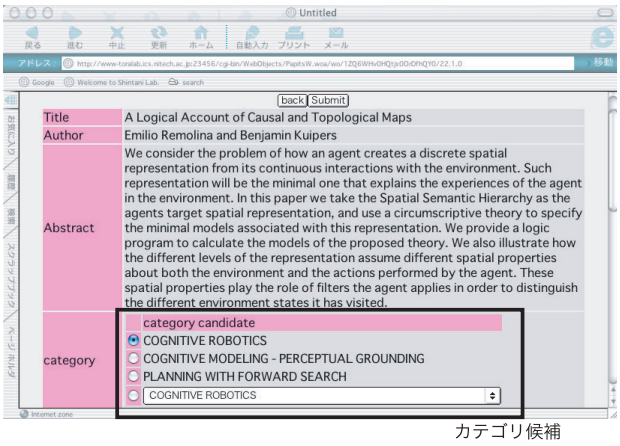

図 3: ユーザによるカテゴリ判定

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．ユーザはその論文のカテゴリを自由に選ぶことができるが，図 3 のように分類器が予測した 3 つのカテゴリの候補を提示を行い，分類器の判断を参考にした決定もできる．これらのユーザの判断によって分類された論文は，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特徴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．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．

この類似度計算には cosine 尺度 $Cos(x_1, x_2)$ を用いた．この尺度によってドキュメント x_1, x_2 の類似度を示す．

$$Cos(x_1, x_2) = \frac{\sum_{j=1}^{n} a_j(x_1) \cdot a_j(x_2)}{\sqrt{\sum_{j=1}^{n} a_j(x_1)^2 \cdot \sum_{j=1}^{n} a_j(x_2)^2}}$$

ここで，ドキュメント x は，$< a_1(x), a_2(x), \cdots, a_n(x) >$ というベクトル形式で表現されているとし，$a_1(x)$ はドキュメント内に単語 a_1 が存在する場合に 1 を，そうでない場合に 0 を返す関数であるとしている．

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ており，中でも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 [Lewis and Ringuette, 94][Yang and Perdersen, 97]．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することが難しくなる．

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カテゴリ分類に影響すると思われる単語をより選択しやすくなると考えられる．また，関連のあるカテゴリがまとめられた状態での属性の評価ができる可能性があり，分類の決め手となる属性を発見できると思われる．

そのアルゴリズムを図 4 に示す．まず始めに，すべてのカテゴリの中で，l 個以下の組合わせのカテゴリの集合 C_A とその C_A 以外のカテゴリの集合 C_B を考え，これらを新しいカテゴリの形式とする．図 5 は，カテゴリ $\{c_1, c_2, \cdots, c_i, \cdots, c_j, \cdots, c_n\}$ が存在するときの $l = 2$ の場合の例である．ここで，l の値分のカテゴリと他のカテゴリという二値化を行うのは，仮に c_i と c_j を含む C_A と C_B を良く分ける語 w_a が存在したとき，c_i と c_k とを含むカテゴリを良く分ける語 w_b という語を発見できれば，w_a と w_b の組合せを評価することで元のカテゴリ c_i, c_j, c_k の分類が可能であると考えたためである．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_A と C_B に対する情報利得を以下のような式で求める．

$$\mathrm{IG}_{C_A,C_B}(w, X) = -\left(\frac{|X_{C_A}|}{|X|}\log_2\frac{|X_{C_A}|}{|X|} + \frac{|X_{C_B}|}{|X|}\log_2\frac{|X_{C_B}|}{|X|}\right)$$

$$+ \sum_{v \in Values(w)} \left(\frac{|X_{C_A,v}|}{|X|}\log_2\frac{|X_{C_A,v}|}{|X_v|} + \frac{|X_{C_B,v}|}{|X|}\log_2\frac{|X_{C_B,v}|}{|X_v|}\right)$$

・ V = 情報利得の値を保持した語の集合
（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
・ l = 2 値化する一方のカテゴリ数
・ $IG_{C_A,C_B}(w,D)$: カテゴリ C_A,C_B に対する情報利得

Feature Selection
- **FOR** すべての語 w
  - **FOR** C の要素 l 個以下のすべての組合わせの集合 C_A
    - $C_B := C - C_A$
    - $IGvalue := IG_{C_A,C_B}(w,D)$
    - **IF** $(max < IGvalue)$ **THEN**
      - $max = IGvalue$
  - 語 w を集合 V に max の値を保持して追加する
- 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 4: 特徴選択アルゴリズム

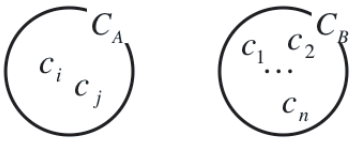

図 5: $l = 2$ の場合の 2 値化

この式はデータをカテゴリ C_A と C_B に分類する場合の語（特徴）w の情報利得となっている．v は w の属性値であり，語 w が出現したとき値 1，出現しなかったとき 0 で表され，$Values(w)$ は w の属性値の集合を表す．$|X_{C_A}|$,$|X_{C_B}|$ は，それぞれカテゴリ C_A と C_B に属するデータの数を表しており，$|X_{C_A,v}|$,$|X_{C_B,v}|$ はそれぞれカテゴリ C_A と C_B に属するデータで w の属性値が v となるデータ数を表している．この値を用いて語 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し，分類に利用する．

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．そのカテゴリは以下のようである．

- *Knowledge Representation and Reasoning*
- *Search, Satisfiability, and Constraint Satisfaction Problems*
- *Cognitive Modeling*
- *Planning*
- *Diagnosis*
- *Logic Programming and Theorem Proving*

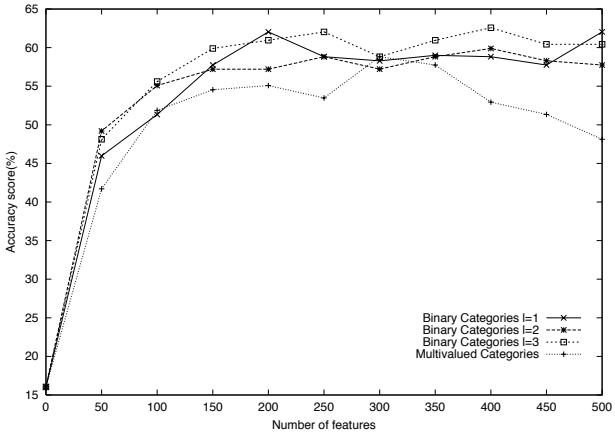

図 6: $N = 1$ の候補による精度

- *Uncertainty and Probabilistic Reasoning*
- *Neural Networks and Genetic Algorithms*
- *Machine Learning and Data Mining*
- *Case-based Reasoning*
- *Multi-Agent System*
- *Natural Language Processing and Information Retrieval*
- *Robotics and Perception*
- *Web Applications*

これらのデータに対して，カテゴリを 2 値化して情報利得を計算する提案手法と通常のカテゴリに対する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．通常のカテゴリに対する情報利得を用いる方法は一般的な特徴選択手法であるとされており [Soucy and Mineau, 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．提案手法では，$l = 1, 2, 3$ の 3 通りについて実験を行い比較を行った．これらを区別するため通常のカテゴリに対する情報利得で属性選択する手法を “Multivalued Categories”，提案手法を “Binary Categories” としておく．

精度の基準として N-best 候補についての評価を行った．1 つ目の基準は，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．つまり，$N = 1$ の N-best 候補と一致した場合を正解としてカウントする．もう一つの基準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．つまり，$N = 3$ の N-best 候補と一致した場合を正解としてカウントする．この精度の基準は，図 3 のような場合に 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

実験結果を図 6, 7 に示す．それぞれ横軸が属性数，縦軸が $N = 1$,$N = 3$ の N-best 候補についての精度を表している．この結果を見るとほとんどの場合で，“Binary Categories” の方が “Multivalued Categories” の場合よりも分類精度が高いことが分かる．また，$N = 3$ 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

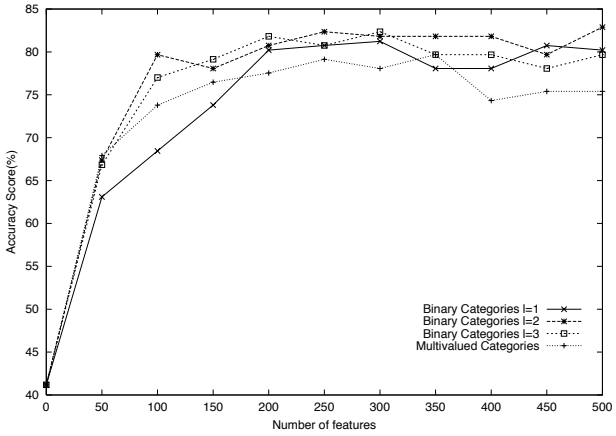

図 7: $N = 3$ の候補による精度

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．

“Binary Categories” の中でも l の値を変えた結果は，$N = 1$ では $l = 3$ の場合が，$N = 3$ では $l = 2$, および $l = 3$ の場合が，$l = 1$ と比べて属性数が少ない段階でも高い精度を誇っていた．これは，l 個のカテゴリで構成される 2 値化されたカテゴリのドキュメント数が増加するため，よりノイズの影響を小さくすることができるため過学習が防がれているためではないかと考えられる．しかし，l が大きくなると計算量が増えてしまい，実際の計算時間は $l = 1$ に比べて $l = 2$ の場合は 4 倍，$l = 3$ の場合は 16 倍になってしまう．ただし，特徴選択は分類システムを稼働させる前にあらかじめ行っておくバッチ処理であり，分類にかかる時間には影響しないためこの計算時間は問題にはならないと考えられる．

また，“Binary Categories” と “Multivalued Categories” で選択された単語の上位１００語を見てみると，“Multivalued Categories” では，”when” や”to” といった一般的な語がいくらか含まれてしまっているのに対して，提案手法で選択された語では，一般的だと思える語は少なかった．この結果からも論文の分野をより表していると思われる語を選択できていると思われる．

5. 関連研究

正確なテキストの自動分類は，テキストの意味理解が必要であると考えられており，計算機にとってはかなり難しい仕事とされている．しかし，機械学習の進展や計算機の発達により表面的な情報を用いた機械学習や数理統計，データマイニングの手法を用いたテキストの自動分類が盛んに行われており，成果をあげている．

テキスト分類のための手法としては，naive Bayes，決定木，KNN などが提案されており，弱学習器を組合わせて高度な分類器を生成するブースティングのような手法も提案されている [Schapire et al., 00][平, 春野, 02]．また，最近では学習手法としてサポートベクターマシンが注目されており，テキスト分類においても良い結果が示されている [Yang and Liu, 99]．

属性選択の方法としては，属性の情報利得などの値を閾値によって評価する方法が多くある．例えば，[Soucy and Mineau, 01] は情報利得と語の共起確率を合わせた評価を使用して属性選択を行っている．

6. 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．

今後の課題として，KNN 以外の分類手法を用いた場合の分類精度などを評価する必要があると考えられる．例えば，現在注目されているサポートベクターマシンのような手法を Papits における論文分類への適用を行い，その評価を行うなどすることが考えられる．

参考文献

[Lewis and Ringuette, 94] D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization, *Third Annual Symposium on Document Analysis and Information Retrieval*, pp 81-93, 1994.

[Fujimaki et al., 02] N. Fujimaki, T. Ozono, and T. 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.

[平, 春野, 02] 平 博順, 春野 雅彦 : トランスダクティブ・ブースティング法によるテキスト分類, 情報処理学会誌,Vol.43, No.6, pp.1843-1851,2002.

[Ozono et al., 02] T. Ozono, S. Goto, N. Fujimaki, and T. Shintani : P2P based Knowledge Source Discovery on Research Support System Papits, *The First International Joint Conference on Autonomous Agents & Multiagent Systems*(AAMAS 2002), 2002.

[Schapire et al., 00] Schapire, R. E. and Singer, Y : BoosTexter : A Boosting-Based System for Text Categorization, Machine Learning, Vol.39, pp.135-168, 2000.

[Soucy and Mineau, 01] P. Soucy and G. W. Mineau, A Simple Feature Selection Method for Text Classification, *Proceedings of International joint Conference on Artificial Intelligence*(IJCAI'01), pp. 897-902, 2001.

[Yang and Liu, 99] Y. Yang and X. Liu, A re-examination of text categorization methods, *22nd Annual International SIGIR*, pp.42-49, 1999.

[Yang and Perdersen, 97] Y. Yang and J. O. Perdersen, A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the Fourteenth International Conference on Machine Learning*(ICML'97), 1997.