救援活動支援および防災対策のためのマッチングシステムの提案

山口大学工学部　学生会員　○小俣 尚泰
（株）栗本鐵工所　関根 総一
山口大学工学部　学生会員　江本 久雄
山口大学工学部　正会員　宮本 文穂
山口大学工学部　正会員　中村 秀明

1. まえがき

近年，駿河トラフ，南海トラフを震源とする東海・東南海・南海地域での地震の発生による大規模かつ広範囲の災害の発生が懸念されており，阪神淡路大震災の経験を基にして各機関において対策が急ぎ進められている[1]．そこで，本研究では災害の現場で迅速にニーズに適した人材・物資・機材・情報が入手できるようなマッチングシステムを提案する．本システムでは，シーズデータベースから最適なシーズを探索し，手配する．マッチングシステムにおいて重要な部分となるシーズデータベースは，構築・保守において，多大な労力を要するため，それらのコストを削減することが望まれる．本研究では，この問題に対し World Wide Web (以下 Web)に注目し，Web からの情報を利用してシーズデータベースを構築することを試みた．

2. マッチングシステム

マッチングシステムとは一定の規則に従って効率的かつ迅速に「需要（needs）」と「供給（seeds）」を一致（matching）させる仕組みの総称である．防災の現場で発生するニーズは，内容は普段発生するものと大差がないが，①災害の状況に応じ刻々と変化，②短時間に大量に発生，③ニーズを伝達・把握が困難，といった特徴がある．これらの特徴を踏まえ本研究におけるマッチングシステムの機能要件を図―1のように定義する．

3. Web データの分類

上記要件 1)，3)を満たすためにシーズデータベースは多くの情報を用意していなければならない．この情報を集め入力する作業は多くの労力を要するので，上記要件 7)を満たすために効率化を図る必要がある．そのため，データベース構築時に情報の自動獲得を行い，その情報がデータベースの設計に従って適切に分類され蓄積される必要がある．本研究では，パターン認識技術を応用し，Web データの分類を行うシステムの構築を試みた．

要件 1) 扱うシーズは人材・資機材・各種情報
要件 2) シーズを地域内外から探し出す
要件 3) 見つからないと言う状況を排除
要件 4) 地域の連携を支援
要件 5) 各種条件（手配時間，価格など）も含めてニーズを解決する
要件 6) ニーズの入力作業の簡便化を　図り，ニーズの吸い上げを効率的に行う
要件 7) シーズの入力作業の簡便化を図る
要件 8) シーズの手配に関する処理の効率化

図―1　マッチングシステムの要件

3.1 SVM(Support Vector Machine)[2]

SVM は 2 クラスの分類問題を解くために作られた学習機械である．SVM の学習には局所解の問題はなく，学習結果は一意に定まる．また，汎化能力も従来法と比較して高い．

ここで，訓練サンプルを，x_1，…，x_n と表す．また，それぞれのクラスラベルを y_1，…，y_n と表し，訓練サンプルがクラス A に属していれば，$y = 1$，クラス B なら，$y = -1$ とする．

このときの識別関数を，次式に示す．

$$f(\Phi(x)) = \sum_{i=1}^{n} \alpha_i y_i \Phi(x)^T \Phi(x_i) + b = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \quad (1)$$

また，学習問題は次式に示すようになる．

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

$$\text{subject to} \quad 0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (3)$$

目的関数 z を最小化すれば，識別面 $f(\Phi(x))=0$ が得られる．このとき K はカーネル関数と呼び，高

次元の空間に写像を行い，線形分離性を高めるために導入される．

3.2 特徴分析・特徴抽出

テキストデータの特徴ベクトルは，そのカテゴリに対しての単語の価値として表現する．本研究では単語の出現頻度と分散を用いてその単語の価値を求める tf· idf[3] を用いて，以下の手順で分野・カテゴリに対する単語の重要度を求める．

1）まず，正例集合の単語を洗い出し，単語の集合を作成する．

2）得られた単語の集合を用いて次式により，正例集合に対して t_i^+，負例集合に対して t_i^- を求める．

$$t_i^+ = n_i^+ \log \frac{M^+}{m_i^+ + \varepsilon}, \quad t_i^- = \frac{M^+}{M^-} n_i^- \log \frac{M^-}{m_i^- + \varepsilon} \qquad (4)$$

M^+，M^- はそれぞれ正例集合，負例集合の文書数である．n_i^+，n_i^- はそれぞれ正例集合，負例集合に含まれる文書における該当する単語の出現回数である．m_i^+，m_i^- は，それぞれ正例集合，負例集合に含まれる文書における該当する重要語が含まれる文書数である．

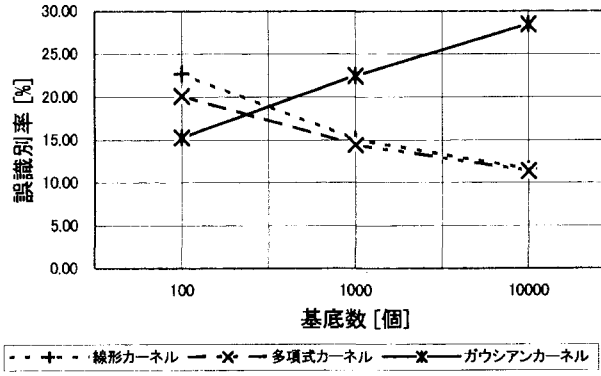
3）得られた t_i^+，t_i^- から次式を基に t_i を求め正例が示す分野の単語の重要度とする．

$$t_i = \left| t_i^+ - t_i^- \right| \qquad (5)$$

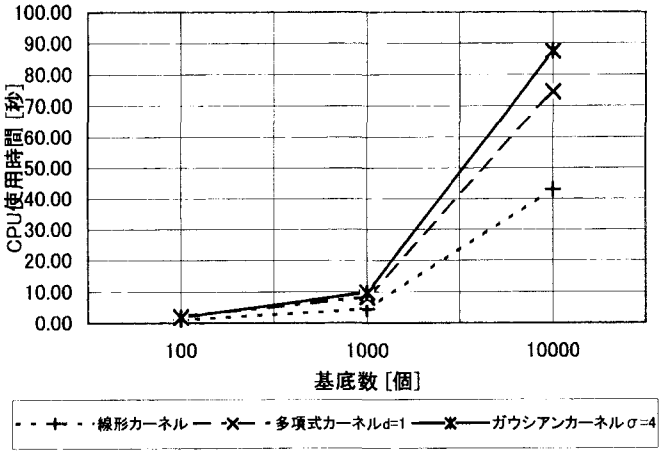
以上より求めた重要度ランキングを用いて特徴ベクトルの基底とする．

4. Web データ分類実験

Web データ分類をマッチングシステムに組み込むことを目指し，Web データの分類が実用できるかどうか評価実験を行った．実験に使用するサンプルデータとして Yahoo!のトップカテゴリ 14 個のそれぞれ深さ 3 で到達できる HTML ページから，ランダムに 1000 個ずつ計 14000 個のファイルを用意し，そのデータを用いて各カテゴリ毎に SVM の学習を行い，識別器を作成した．識別器に入力するベクトルの基底数を 100 個，1000 個，10000 個の 3 パターンと定め，カーネルを線形カーネル，多項式カーネル，ガウシアンカーネルの 3 つを用い，そのパラメータも変えて実験を行った．識別器の検証には Leave-one-out 法を用いた．カテゴリ「エンターテイメント」の識別結果を**図—2**に，計算時間の結果を**図—3**を示す．ガウシアンカーネル以外で基底数の増加が識別性能に良い影響を与えていることがわかる．



図—2 「エンターテイメント」の識別結果



図—3 「エンターテイメント」での計算時間

5. まとめ

① 災害時に発生する多種多様なニーズを解決することのできるマッチングシステムの設計を提案した．

② データベースの半自動構築を目指し，Web データ分類システムを提案した．

③ Web データ分類システムにおいて，汎化能力の高い SVM の導入を試み，検証を行った．その結果，高い識別性能を示し，その有効性が確認された．

参考文献

1) 内閣府防災部門：わが国の災害対策，内閣府政策統括官（防災担当），2002.3.

2) 前田英作：痛快！サポートベクトルマシン，情報処理学会誌，42，pp.676-683，2001.7.

3) Salton, G. and McGill, M.J.: Intoroduction to Modern Information Retrieval, McGraw-Hill, 1983.9.