Webデータを活用したマッチングシステムの復興活動支援への適用に関する研究

Application of Web based matching system to rehabilitation activities support in a disaster

小俣　尚泰	関根　聡一	河村　圭	宮本　文穂
Naoyasu Omata	Souichi Sekine	Kei Kawamura	Ayaho Miyamoto
山口大学大学院	(株)栗本鐵工所	山口大学	山口大学
Yamaguchi University	KURIMOTO, LTD.	Yamaguchi University	Yamaguchi University

Absract: Recently, there are anxious about the large-scale and wide range calamity by the outbreak of the earthquake in Tokai, Tonankai, and Nankai areas. Therefore, each organization is taking the countermeasures based on the experience of the great Hanshin Awaji earthquake disaster. The purpose of this study is a matching system that can get quickly the information suitable for needs in disaster area. This system use information on Web (World Wide Web) in order to search from a seeds database and arranges for things such as goods, talented people, and so on. When the database is constructed or maintained, the seeds database demands the great labor of the developer. Thus it is necessary to categorize the Web data using by SVM (Support Vector Machine).

1. はじめに

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，過去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発生するニーズに適した人材・物資・機材・情報が迅速に手配できるようなマッチングシステムの構築を目指している．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にかつ迅速に「需要（needs)」と「供給（seeds)」を一致（matching）させる仕組みの総称である．

本研究では，このマッチングシステムの仕組みを利用した災害復興活動支援を行うシステムの構築を目指している．図 1 に災害地でのワークフローを示す．災害対策本部ではニーズの把握，シーズの把握を行い何が適材適所であるかを判断し，災害地に送り込むシーズの手配を決定する．平常の仕事より多くの仕事が短時間に到来するため，大きな負担が災害対策本部にかかるという問題がある．

マッチングシステム導入すると図 2 のように，災害対策本部のタスクを一手にシステムが引き受けるようになる．ニーズとシーズの把握を，その情報をデータベース化という形で行い，システム内部で災

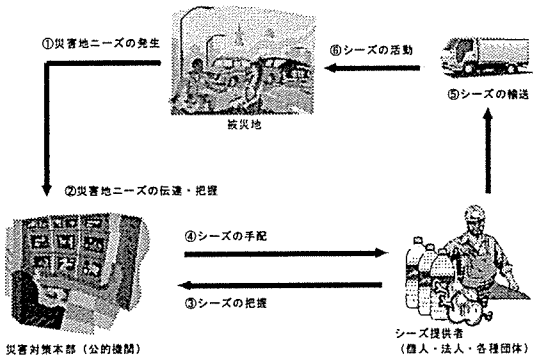

図 1　災害地のワークフロー

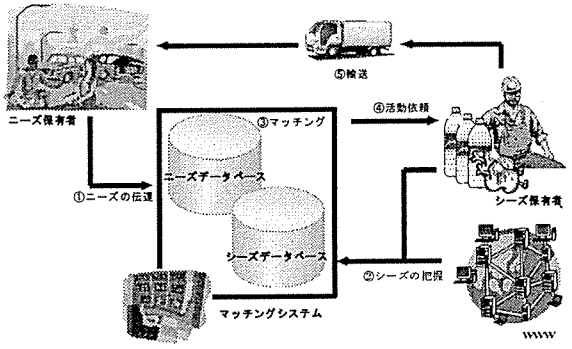

図 2　システム導入後のワークフロー

害の状況などの各種情報と照らし合わせながら解決を行い適材適所にシーズを送り込む.

3. Webデータ分類システム

マッチングを行うための情報が見つからないという状況を極力排除するために，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．そのため，データベース構築時に情報の自動獲得を行い，適切に分類され蓄積されれば，この問題は解決できると考えた．本研究では，マッチングシステムのデータベース構築をサポートすることを目的として，表 1 に示される Web データ収集システムの構築を試みた.

Webデータの分類を実現するにあたってはパターン認識技術を応用した．図 3 に Web データ分類システムの構築フローを示す．このシステムは訓練フェーズを必要とする．ひとつの分類カテゴリの分類器を得るために，左側の訓練フェーズのフローを行う必要がある．以下に各処理について説明する.

① 前処理：ここではHTMLタグを削除し，残った日本語テキストデータに対して形態素解析を行い，単語群を得る．形態素解析において名詞，未知語と判定された形態素を単語群としている.

② 特徴分析：前処理済みの正例データ，負例データを用いてカテゴリの特徴について分析を行う．結果は特徴抽出器として出力される．詳しい特徴分析の方法については後述する.

③ 特徴抽出：特徴分析で得られた特徴抽出器を用いて単語群データを数値データにし，1ページのHTMLページをひとつのベクトルデータに変換する.

④ 分類学習：正例，負例それぞれの特徴ベクトル群を用いて分類境界を算出し，分類器を作成する．本研究では分類器に SVM(Support Vector Machine)を用いている．SVM について詳しくは後述する.

⑤ 分類：分類学習によって得られた分類器を用い，入力された特徴ベクトルが，そのカテゴリに属するか否かを判定する.

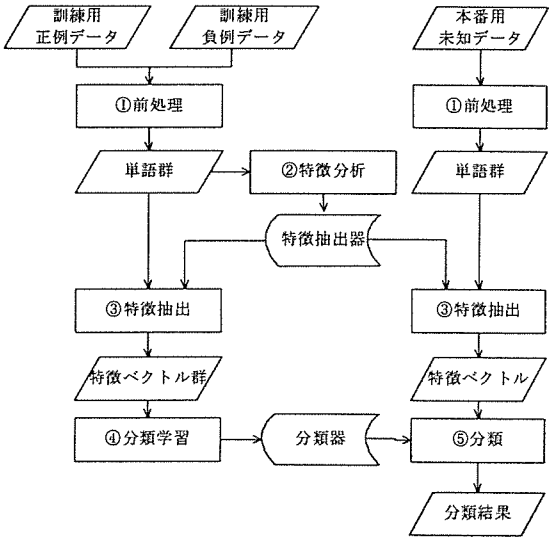

図 3 Web データ分類システム構築フロー

表 1 Web データ収集システムの概要

機能	説明
①収集	WWWからデータを無作為に収集する
②分類	収集されたデータをあらかじめ決められたカテゴリに分類する
③利用	分類されたデータはシーズとして取り扱い，マッチング処理の補助として用いる

3.1. 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 [2] 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.

$$t_i^+ = n_i^+ \log \frac{M^+}{m_i^+ + \varepsilon}, \quad t_i^- = \frac{M^+}{M^-} n_i^- \log \frac{M^-}{m_i^- + \varepsilon} \tag{1}$$

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.

$$t_i = \left| t_i^+ - t_i^- \right| \tag{2}$$

以上より求めた重要度ランキングを用いて任意の数ほど上位から選び，特徴ベクトルの基底とする．

特徴抽出では，特徴分析によって選ばれた基底の単語があれば１，なければ０として，単語群をベクトルデータとして変換出力する．

3.2. Support Vector Machine (SVM)

SVM [3] 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．

$$\begin{aligned} f\left(\Phi(\boldsymbol{x})\right) &= \sum_{i=1}^{n} \alpha_i y_i \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{x}_i) + b \\ &= \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b \end{aligned} \tag{3}$$

また，学習問題は次式に示すようになる．

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{4}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5}$$

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 は式(4)で示される関数である．

$$K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{x}') \tag{6}$$

この関数をカーネル関数と呼び，

$$\Phi : \boldsymbol{R}^d \to \boldsymbol{R}^q \qquad (d < q) \tag{7}$$

となるような高次元の空間に写像を行い，線形分離性を高めている．カーネル関数の例として，多項式カーネルを式(8)，ガウシアンカーネルを式(9)に示す．

$$K(\boldsymbol{x}, \boldsymbol{x}') = \left(\boldsymbol{x}^T \boldsymbol{x}' + 1\right)^p \tag{8}$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = e^{\frac{\|x - x'\|^2}{\sigma^2}} \tag{9}$$

3.3. 多クラス分類への対応

SVM では２値分類しか実現できないため，２値分類器を複数組み合わせて構成する多クラス分類法を適用しなければならない．そこで以下の２種類の多クラス分類法を検討した．

① One versus the rest（1vsR）：N クラスの分類を行いたいときにはN個の識別器を作成する．それぞれの識別器の学習で使われるデータは正例がそのクラスに属するデータ，負例がそのクラスに属さないデータ全てとなる．識別時間は $O(N)$ となる．

② Pairwise：N クラスの分類を行いたいときには ${}_nC_2 = \frac{N(N-1)}{2}$ 個の識別器を作成する．それぞれの識別器で使われるクラスごとのデータ集合の組み合わせとなる．識別時間は原理的に $O(N^2)$ となる．

図 4，図 5 はクラス数が $N = \{3,4,5\}$，各クラスのデータ数は１０個ずつとして，1vsR 法と Pairwise 法の比較実験の結果である．分類器は前述の図 3 の方法により作成される．この実験により以下の結果を得た．

① Pairwise，1 vs R の各手法は，学習時間，識別時間のトレードオフな関係にある．

② マッチングシステムでは識別時間を優先し1vsR 法を用いた方が良いと思われる．

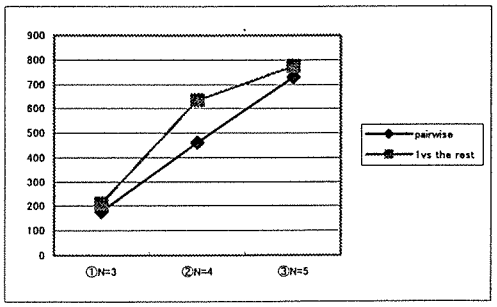

図 4　多クラス分類：学習時間

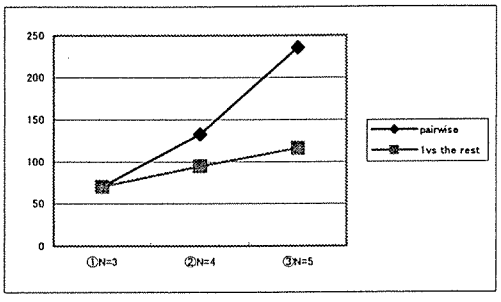

図 5　多クラス分類：識別時間

4. マッチングシステムプロトタイプ

図 2 に示される構想をもとにプロトタイプシステムを作成した．システムの構成は表 2 のように①資機材マッチングシステム，②人材マッチングシステム，③Web シーズ収集システム，④ユーザ管理システムからなる．

図 7, 図 6 は資機材マッチングシステムの画面である．図 7 の画面でシーズ探索条件を入力し，システムに対して問い合わせを行う．図 6 は条件に大分類：「事務機器」，小分類：「パソコン」，その他は未入力で問い合わせをした結果例である．上段の表となっている部分が手作業で入力されたデータであり，下段のハイパーリンクが Web シーズ収集システムにて登録されているデータである．この Web データは前述の Web データ分類システムにより大分類：「事務機器」，小分類：「パソコン」というカテゴリに分類されたデータである．このように Web 上のデータを結果提示の補助として使うことによって，情報が見つからないという状況を排除することが実現できているといえる．

5. 終わりに

本研究では災害復興を目的としたマッチングシステムの活用の研究を進めている．一般的に情報が不足しがちなデータベースが抱える問題に対して，Web データを活用で解決することを試みている．Web データの活用法にあたっては SVM を用いた分類器を作成することにより，Web データにカテゴリを割り付け，従来法によるデータとの連携を実現した．

今後の課題としては多クラス問題に対する学習時間及び識別時間の双方の向上，実際の災害を想定したシミュレーション形式のシステム検証を行う必要などがある．

表 2 プロトタイプシステム構成

①	資機材マッチングシステム	①災害地で必要とされている物資・機材の情報（資機材ニーズ）の収集
		②災害地で利用するための物資・機材の情報（資機材シーズ）の収集・管理
		③資機材ニーズと資機材シーズの管理
②	人材マッチングシステム	①災害地で必要とされている人材の情報（人材ニーズ）の収集・管理
		②災害地で利用するための人材の情報（人材シーズ）の収集・管理
		③人材ニーズと人材シーズの管理
③	Webシーズ収集システム	WWW(World Wide Web)から資機材・人材の情報を収集し，マッチング結果に反映する．
④	ユーザ管理システム	システム上でニーズの保有者，シーズの提供者としてユーザを管理する．

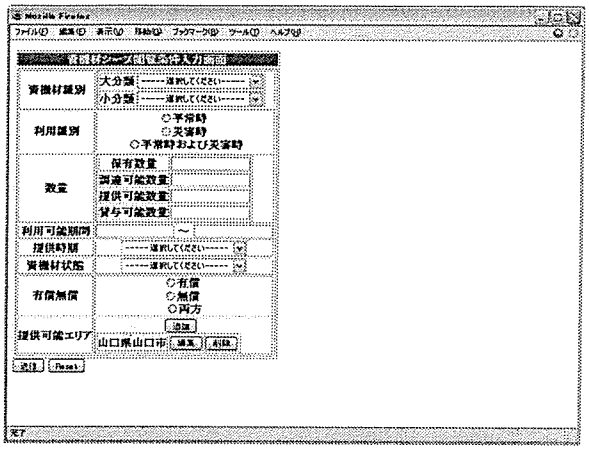

図 7 資機材シーズ閲覧条件入力画面

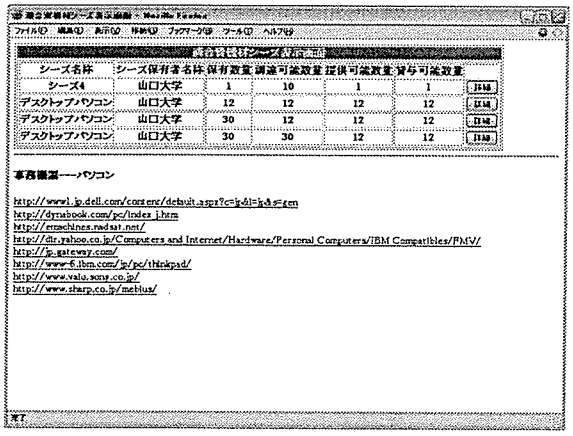

図 6 マッチング結果例

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

[3] 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.

連絡先：

小俣 尚泰

山口大学大学院理工学研究科知能情報システム工学専攻

〒755-8611 山口県宇部市常盤台 2-16-1

Phone: 0836-85-9530

Fax: 0836-85-9530

E-mail: omata@design.csse.yamaguchi-u.ac.jp