

文書クラスタリングにおける文書属性抽出ツールの汎用化と拡張

納富 一宏¹・山口 俊光²・斎藤 恵一³・藤本 哲男⁴

¹情報工学科

²情報工学科 学部4年

³東亜大学経営学部経営学科

⁴芝浦工業大学工学部機械工学科

Generalizing and Extending Document Attributes Extraction Tool in Document Clustering —Attribute Extraction and Information Visualization on Classifying—

Kazuhiro Notomi¹⁾, Toshimitsu Yamaguchi²⁾, Keiichi Saitou³⁾, Tetsuo Fujimoto⁴⁾

Abstract

In this article, we propose a generalizing and extending method of information extraction tool for clinical case clustering system using Self-Organizing Maps (SOM). This system runs on our WWW based DBMS (Database Management System) for clinical cases. The system was implemented in Java and C++ on the Internet server, and all documents of clinical case in the database are written in Japanese. Our method is available to design widely some SOM based clustering systems for medical information.

Keywords: Self-Organizing Map, Clinical Case, Natural Language Processing, Document Clustering

1. はじめに

近年、インターネットに接続されたサーバ/クライアント間で、医療情報の検索・閲覧を行なうネットワークデータベースシステム(Network Database System)の構築が重要視されている。

我々は、こうしたシステムの一つとして、WWW による臨床症例データベースの構築を行なっている^{[1]-[3]}。

最近では、臨床症例データベースにおいては、新規症例の登録、病態把握のための検索・閲覧などをはじめとして、データベース内に存在する全症例の疾患系の類似度による分類(クラスタリング)を行なうことを目的として、システムの拡張を手がけてきた^{[4], [5]}。

具体的には、データベース化された臨床症例報告文書から、その疾患についての特徴的な語句や表現、さらには各種検査項目に関する情報を自然言語処理的に自動抽出して、類似度を計算するための属性値を求め、これら属性値から自己組織化マップ(SOM: Self-Organizing Map)への入力ベクトルを構成し、学習アルゴリズムの適用を経て、2次元マップを生成することで、情報の視覚化や検索インタフェースへの応用^[6]を考慮した「臨床症例クラスタリングシステム」の構築を行なっている。

本稿では、医療支援、診断支援までを視野に入れ、

医療情報向けの汎用クラスタリングシステムの構築手法について検討する。特に、症例文書からの属性抽出と分類結果の視覚化を行なう上で、より汎用的なシステム設計について述べる。

2. 基本システム構成

2.1 自己組織化マップ(SOM)の概要

SOM は、トポロジカルマッピングを拡張した教師なし競合学習型ニューラルネットであり、入力層とマップ(出力)層の2層構造をなす。また、データ間の特徴類似度による汎用的なクラスタリング能力を持つ。SOM を用いた文書情報検索システムとしては、WEBSOM^{[7], [8]}が知られている。

SOM モデルは、入力層では n 個、マップ層では2次元的に配列された m 個のニューロンからなる。入力層とマップ層の各ニューロンは全結合であり、それらの結合荷重は、 $m \times n$ 行列で表現される。

今、 j 番目の n 次元入力ベクトルを \mathbf{x}_j 、 i 番目の重みベクトルを \mathbf{w}_i とすると、ベクトル間のユークリッド距離 $\|\mathbf{w}_i - \mathbf{x}_j\|$ を最小とする組を k とすると、SOM アルゴリズム

ムによる重みベクトル w_k の更新は次式で示される。

$$w_k^{new} = w_k^{old} + \alpha(x_j - w_k^{old})z_k \quad (1)$$

α は「学習率」と呼ばれ、学習回数 t の単調減少関数である。 Z は競合作用値であり、 k に一致した場合のみ 1、それ以外では 0 を与える。

実際の学習では、 k 番目のニューロンの幾何学的近傍についても式(1)を適用し、重み更新を行なう。

2.2 症例クラスタリング

一般的な症例文書とは、A4 で数ページ(通常、全角 1,000~4,000 文字程度)の文書であり、記載医の所属・氏名、患者の氏名(イニシャル)・年齢・性別・生年月日・主訴、診断、入退院の日付、担当医名(受持、外来)、入院目的、現病歴、既往歴、家族歴、生活歴、入院時現症、入院時検査所見、退院時処方、問題点などが記載されている。

症例クラスタリングでは、症例文書から得られた情報を元に、入力ベクトルを生成し、これらに SOM アルゴリズムを適用してマップを生成することで、症例文書そのものを3疾患系(①循環器系、②消化器系、③呼吸器系)毎に分類するものである^{[4],[6]}。

基本システム構成を Fig.1 に示す。

システム構成的にみると、マップ生成のための SOM エンジン、視覚化エンジン、およびユーザインタフェースについては、臨床症例以外の対象についてもそのまま利用することができると考えられる。

そこで、システムの汎用化と拡張を以下の 4 つのフェーズに分けてそれぞれ検討する。

- ① 情報抽出
- ② 入力ベクトルの構成
- ③ SOM アルゴリズムの適用
- ④ 情報視覚化およびクラスタリング結果提示

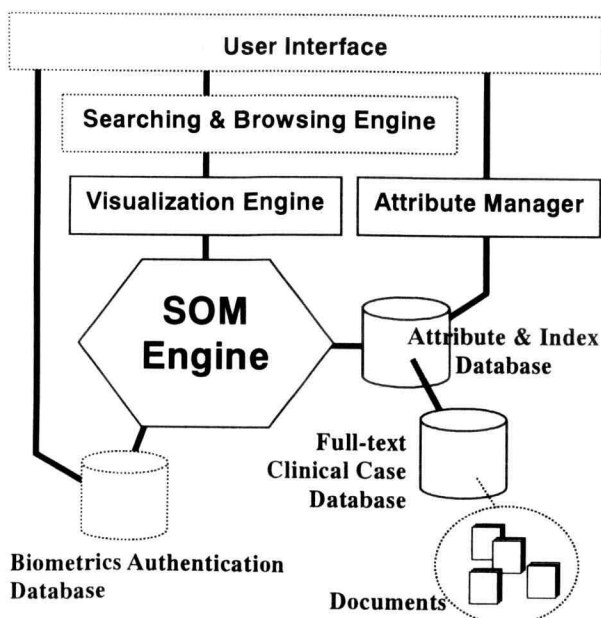


Fig.1 Basic System Construction

3. 汎用化と拡張

3.1 情報抽出

臨床症例を扱う場合は、各症例文書から自然言語処理的なアプローチにより、キーワードと見なせる単語を自動抽出する(Fig.2 参照)。最終的には、これらの出現頻度から出現確率を求め、SOM の入力ベクトルを構成することになる。

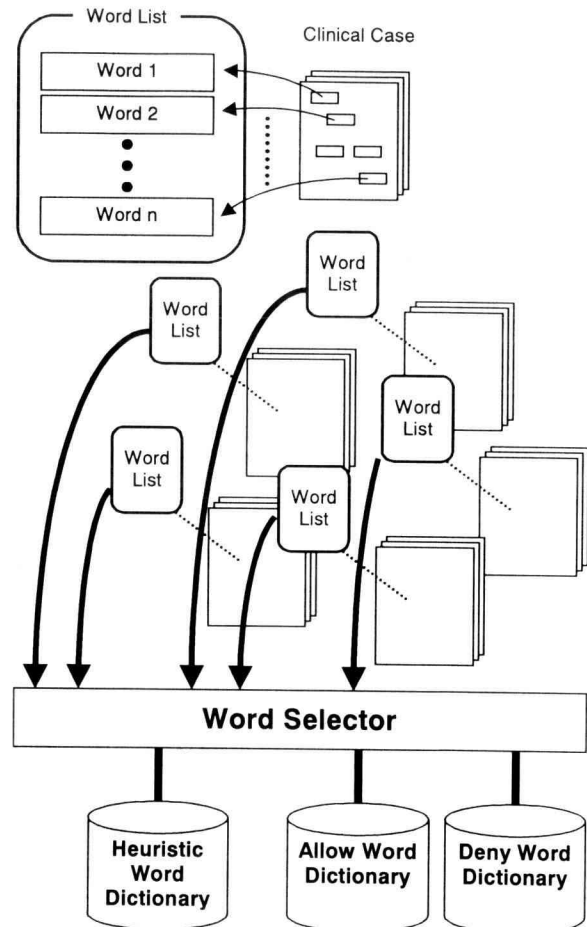


Fig.2 Keywords extraction from clinical cases

単語抽出におけるフィルタリング処理としては、特定の用語を積極的に抽出するための辞書照合、逆に特定の用語を排除するための辞書照合を考えることができる。

通常、前者は医学関連用語辞書、後者は一般用語辞書を用いることで対応することができるが、形態素解析(字句解析)や辞書照合の負荷を軽減する場合は、文字種別によるヒューリスティクスを用いる方法を我々は既に提案している^{[1],[3]}。

頻度情報の文脈的な抽出(文字、あるいは語などの接続共起情報の抽出)を行なう場合は、形態素解析における n -gram を用いることで対応できる。一般に、 n は 3 程度までであり、 $n=1$ の場合、uni-gram と呼び、単語出現確率に一致する。 $n=2$ の場合を bi-gram、 $n=3$ の場合を tri-gram と呼び、 n 個接続した単語(形態素)の組をエントリと考えて、頻度(出現確率)を求める方法である。こ

の n -gram 方式を用いる場合は、対象となる文書量が比較的大きいものである必要がある (Fig.3 参照)。

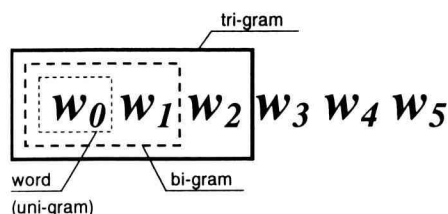


Fig.3 N-gram for keywords

3.2 入力ベクトルの構成

基本的に、入力層に与えるデータ群は、数値化された任意の属性値からなるベクトル表現であれば、SOM によるクラスタリングマップを生成することが可能である (Fig.4 参照)。

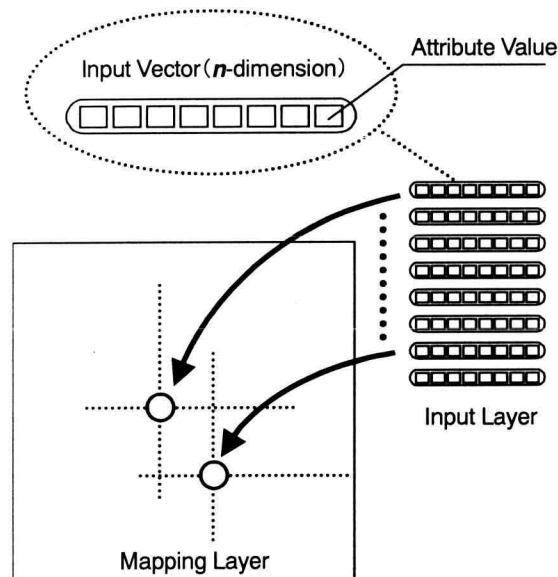


Fig.4 Data mapping of input vectors (n - dimension)

実際に我々が試みたクラスタリングマップ生成のための属性値 (医療関連データ) の例としては以下のものがある。

- ① 症例文書中の医学関連用語の出現頻度
- ② 肝機能血液検査結果の数値データ
- ③ 胃内視鏡画像診断の医師所見に基づく数量化データ
- ④ 腹部超音波診断の医師所見に基づく数量化データ

3.2.1 属性値の正規化・標準化

上記①のように、頻度データを扱う場合は、実際には、出現確率や共起確率を求める必要がある。それ以外の数値データや数量化データは種類の異なるものを混在

させる場合、特定の属性値の影響を抑えるために、正規化・標準化を施す必要がある。

単純な正規化の例としては、同一属性内の相対的な割合としてパーセント値を用いる方法がある。あるいは、特定のレンジ内に振り分ける一次変換を考えることもできる。更に、属性の2値化を行なうことにより、論理型として扱うことも可能である。

多くの場合、各属性のレンジオーダーを揃えた方が良いが、ある属性値が他の属性値に対して独立でない場合、予想通りのクラスタリング結果とならない場合があると考えられる。

3.2.2 属性の選択と入力層への投入順序

汎用データを扱う場合、クラスタリングを行なうべきデータと属性値との対応付けが重要である (Fig.5 参照)。

属性数が多いと、SOM 学習に要する時間が長くなるため、全データフィールドから属性を取捨選択できた方が良い。属性の順列に関しては、学習への影響はないと考えられるが、入力ベクトルの把握を行なうのは、ユーザであるため、インタフェース設計としては、並びが可変である方が良い。

また、SOM 学習に関わる属性値の投入順序は、一般にはランダムに行われるが、これをあらかじめグループ化されたデータ毎に投入する方法や、グループを循環させてシーケンシャルに投入する方法など、SOM 入力層へのベクトル投入に関するいくつかのバリエーションを考えることができる。

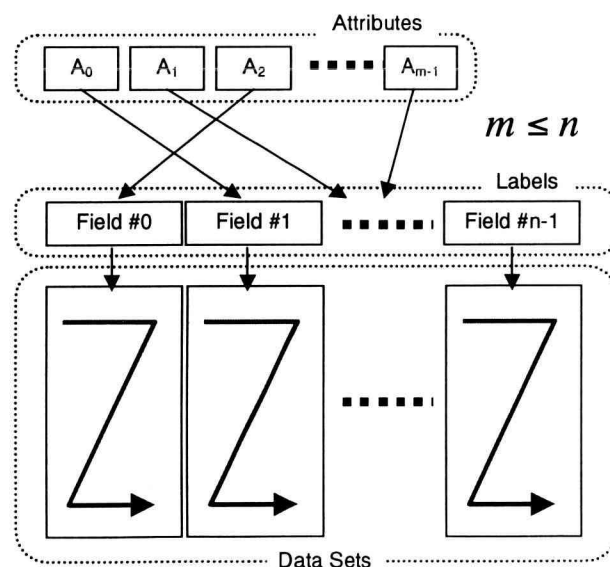


Fig.5 Attribute selection of data sets

3.3 SOM アルゴリズムの適用

SOM アルゴリズムの適用部分では、学習回数、マップサイズというパラメータの他に、学習率曲線、近傍判定曲線の選択が可能である方が良い。これは、生成されたマップの良さを比べる上で必要な機能となる。

近傍モデルに関しては、マップ上のセル形状により、主に正方形や正六角形が用いられる (Fig.6 参照)。

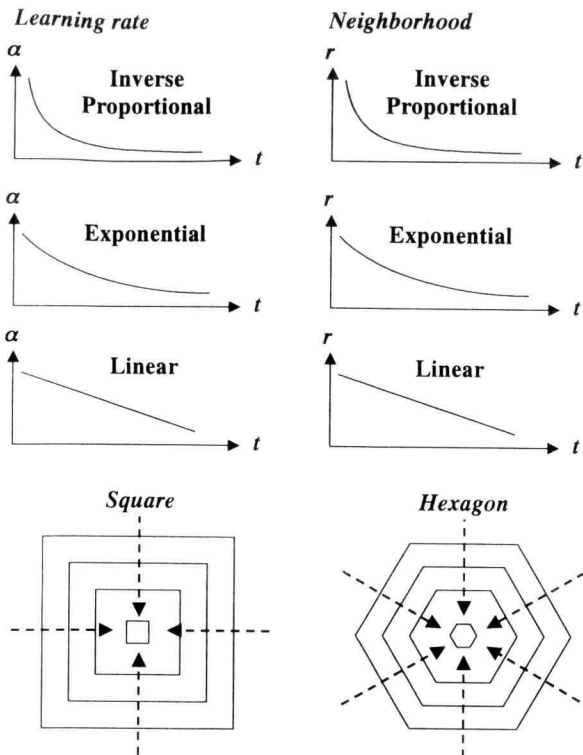


Fig.6 Variation of equation for SOM learning

3.4 情報視覚化およびクラスタリング結果提示

マップ部分を階層構造化することでさまざまな情報提示が可能となる。ノードへのラベル付け、クラスタリング領域のカラー化、属性強度のカラー化、同一セルへマッピングされたノードの個数などが独立のレイヤーで操作できる方がよい(Fig.7 参照)。

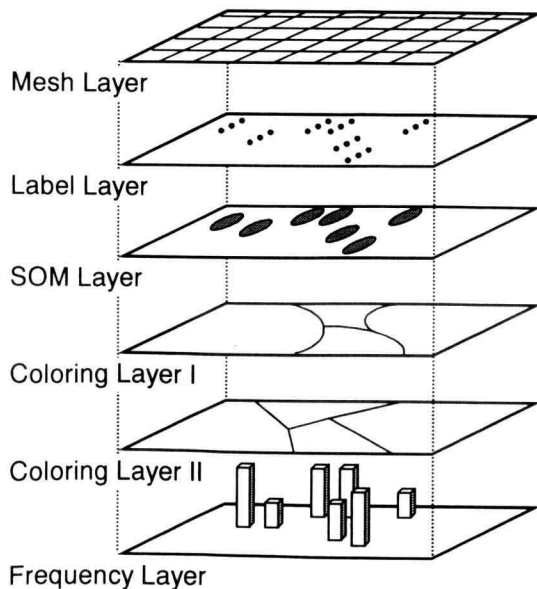


Fig.7 Display layers for SOM viewing

この他、情報視覚化については、3次元的なマップ表示や画面操作インターフェースが望ましいと考えられるが、現行のシステム設計ではそこまで対応していない。

4. 実装システム

C++言語により、情報抽出ツールを UNIX サーバ上に CGI アプリケーションとして実装した。これによりクライアント側は一般的なブラウザのみで実行が可能となる。プログラムは、①自然言語処理部、②CGI インタフェース部、および③入力ベクトル生成部からなり、ソース規模は、3,496 行(①=816 行、②=1,581 行、③=1,099 行)である。

本ツールの動作画面例として入力インタフェース画面、抽出結果表示画面、入力ベクトル生成画面をそれぞれ、Fig.8, Fig.9,および Fig.10 に示す。

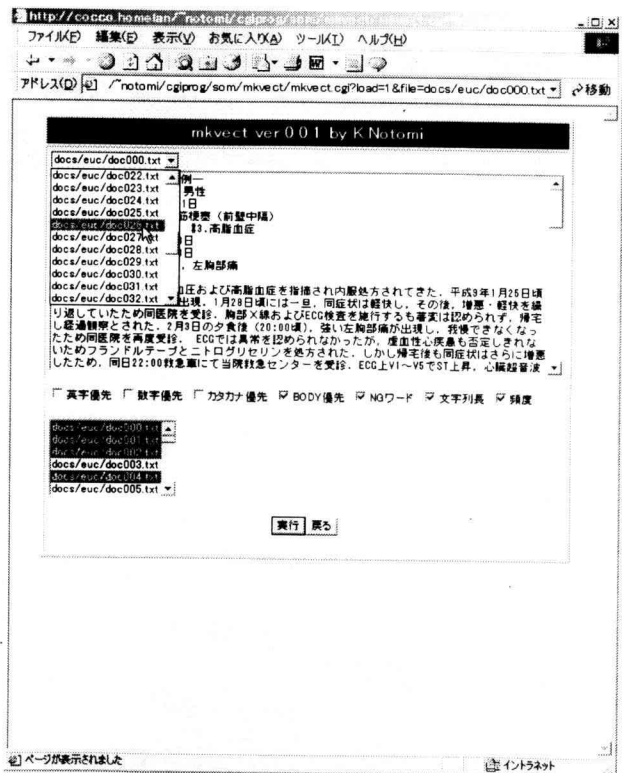


Fig.8 Document browsing & Configuration of keyword extraction

Fig.8 に示した画面は、文書ファイルの内容確認を行なうためのブラウジングダイアログとして機能する一方、キーワード抽出条件の指定ダイアログとしても機能する。抽出条件には、臨床症例文書用の医学関連用語抽出フラグ(画面「BODY 優先」チェックボックス)のほか、文書中の自立語成分として「英字優先」、「数字優先」、「カタカナ優先」のフラグ指定が可能である。これらは、症例文書内の検査項目および検査結果に関する表記に対して効果的である。

また、特定文字、もしくは特定語句を NG ワードとして外部ファイルに登録しておき、これらを排除するための「NG ワード」フラグの指定ができる。

さらに、ノイズ除去のための「文字列長」フィルタ、および「頻度」フィルタを備えている。前者は、1文字の自立語成分をカットする。後者は出現頻度1のものをカットする。

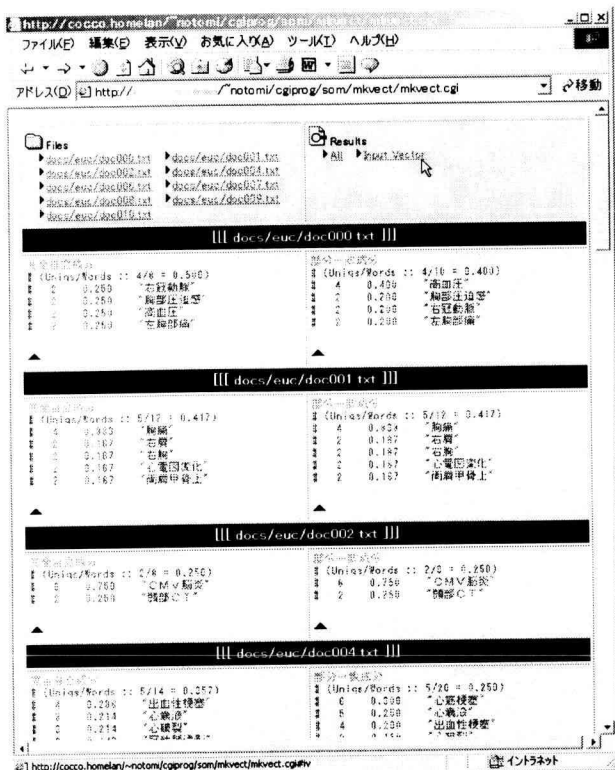


Fig.9 Extracted keyword lists with hyper-links

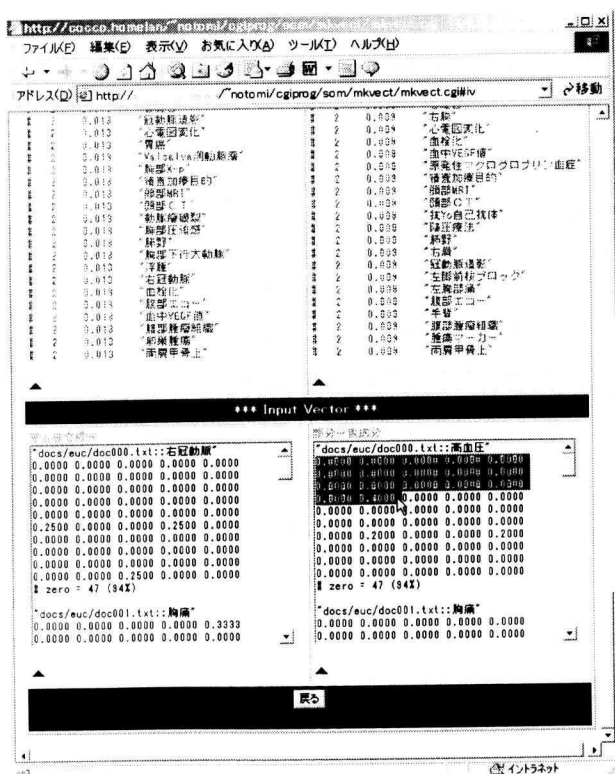


Fig.10 Created Input Vectors

Fig.9, および Fig.10 に示した画面は、結果表示例である。Fig.8 の実行ボタンをクリックすることで、自動抽出されたキーワードリストを提示するとともに、最終的には各キーワードリストから出現確率を算出することで、SOM 学習用の入力ベクトルの生成結果を表示する。

Fig.9 の先頭に表示されている部分が HTML リンクによるショートカットメニューになっている。ここをクリックすることで、ユーザは特定の結果画面を迅速に表示することができる。

入力ベクトルは、Fig.10 に示したように、HTML フォーム(<TEXTAREA>~</TEXTAREA>)を利用して内容表示を行なっているため、ユーザは入力ベクトルの再編集が簡単にできる。また、ユーザはコピー & ペーストを利用することができるため、エディタなど他のアプリケーションとの連携も容易である。

5. まとめ

自己組織化マップ(SOM)による臨床症例クラスタリングシステムの構築手法について述べた。

特に、症例文書からの属性抽出と分類結果の視覚化を行なう上で、より汎用的なシステム設計について検討した。また、この設計に従って、SOM 学習用の入力ベクトル生成に適した属性抽出ツールの実装を試みた。

文書クラスタリングでは、属性抽出量が均一であることが望ましく、文書量が均一化されていない場合に、良い SOM マップが得られないという問題点がある。現在、属性値の選択フェーズにおいて、これを補正する手段を模索中である。

今後の課題としては、医療情報全般を扱えるクラスタリングシステムの汎用化と拡張について更に検討を進めると共に、バイオメトリクス認証によるセキュリティ強化を図ることで、インターネットにおける医療情報アクセスの実現可能性について評価データを収集する予定である。

参考文献

- [1] 納富, 斎藤, 藤本: WWW による臨床症例データベース検索システムの構築, バイオメディカル・ファジィ・システム学会誌, Vol.1, No.1, pp.35-45(1999).
- [2] 斎藤, 納富, 藤本: WWW による臨床症例データベース検索システムの構築(第 2 報)ーファジィ測度論による症例分類ー, 情報処理学会第 58 回全国大会講演論文集, 5G-03 (1999).
- [3] 納富, 斎藤, 藤本: 自然言語処理とアソシエーションを用いた疾患系分類, 情報処理学会第 58 回全国大会講演論文集, 5G-03 (1999).
- [4] 納富, 岡本, 山口, 他: WWW による臨床症例検索システムの開発ー自然言語処理と自己組織化マップを用いた疾患系分類ー, 第 61 回情処全大, 4R-06, (2000.10).
- [5] 山口, 納富, 他: WWW による臨床症例検索システムの開発ー自己組織化マップを用いた打鍵タイミングによる個人認証?ー, 第 61 回情処全大, 4R-03, (2000.10).

- [6] 納富, 山口, 他: 臨床症例データベース管理システムの構築 —自己組織化マップによる情報の視覚化と検索インターフェース—, 第 62 回情処全大, 6Q-01, (2001.03).
- [7] S.Kaski, K.Lagus, T.Honkela, T.Kohonen : Statical Aspects of the WEBSOM System in Organizing Document Collections, Computing Science and Statics, 29, pp.281-290, (1998).
- [8] K.Lagus and S.Kaski: Keyword selection method for characterizing text documents maps , ICANN '99, (1999).