

WWW を用いた自然言語による 臨床症例データベース検索システムの構築

榎本 繁*, 納富 一宏*, 斎藤 恵一**, 藤本 哲男***, 吉田 正徳**, 石井 博章*

*神奈川工科大学情報工学科

**早稲田大学 人間総合研究センター

***早稲田大学 理工学総合研究センター

1. はじめに

医療施設において過去の症例を効率よく集積し検索できるシステムは開発されていない。医療施設における症例情報の大多数は学会誌などにも発表されず書類の形態で各施設に保存されている。これらの情報をデータベース化し簡単に検索できるシステムによってより効率的に医療情報が得られると筆者らは考える。

2. システムの構成

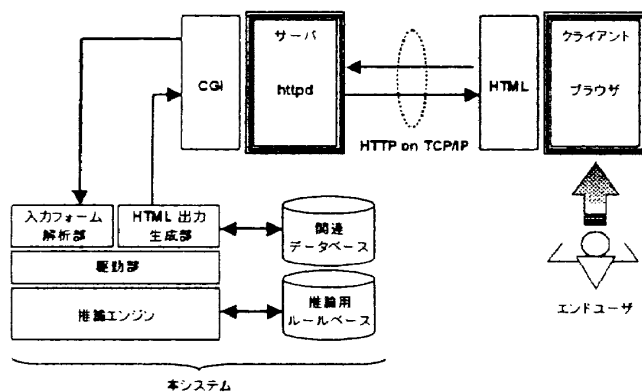


図1 本システムの基本構成

図1に示すように、本システムではサーバ側にプログラム本体を置き、ネットワークを介してクライアントマシンとの通信を行う。クライアント側は、一般的なWWWブラウザを使用する。エンドユーザはブラウザに表示された入力

フォームに症例文章を入力し、内容を送信する。

受信したプログラムでは自然言語で記述された症例情報から自動的にキーワードを抽出し、症例文章を十分反映する検索用インデックスを作成するデータベースを構築する。

自然言語文として入力された臨床症例データを元に、自動インデキシングを通してフルテキストデータベース(full-text database)を構築する。また、検索時に必要なインデックスを作成するため、症例をデータベースに登録する際に臨床データを自然言語解析し、自動でキーワードを抽出および選択を行う。さらに、症例のファジィ測度による系分類を行うことで検索効率の向上を目指す。

3. 自然言語処理

3.1 キーワードの抽出

キーワードの自動抽出処理では、まず入力文章から自立語部分の抽出を行う。抽出する際に形態素辞書を使わずに「字種切り法」を用いて処理の高速化を行う。

日本語文における文字種別の内、平仮名表記の自立語を無視できるものと仮定すれば、自立語となる文字種は、「漢字」、「カタカナ」、「数字」、「英字」のいずれかである。従って、これら自立語を構成し得る文字種の連続からなる部

Development of a new database system on WWW for clinical cases with natural language interface
Shigeru Enomoto*, Kazuhiro Notomi*, Keiichi Saito**, Tetsuo Fujimoto***, Masanori Yoshida**, Hiroaki Ishii*

*Department of Information and Computer Sciences, Kanagawa Institute of Technology

**Advanced Research Center for Human Sciences, Waseda University

***Advanced Research Institute for Science and Engineering, Waseda University

分文字列は自立語、もしくは自立語の一部となる。検索および被検索キーワードは、名詞類からなるものと考えられるため、動詞、形容詞、形容動詞といった自立活用語は名詞化されたもののみに対応できる。活用語尾は平仮名表記されるため、原理的に排除される。

よって、この処理により自立語、すなわちキーワード候補が取得できる。

3.2 用語の選択

次に抽出された自立語集合から検索キーとなる用語を選択する必要がある。すなわち、医学用語などの症例データを特定する語以外の一般的な語は、あらかじめ排除されなければならない。これには、数万から数十万という膨大な医学用語の辞書データと照合するという方法も考えられるが、検索時の処理時間が非常に長くなる恐れがある。よって本システムでは、医学用語に見られる表記文字の代表的なものを文字テーブルとして外部ファイルに登録しておき、先に抽出された自立語集合のうち、これら文字テーブルに登録されたいずれかの文字を含むものを「用語」として選択することにする。外部ファイルにより供給される形式であるため、このファイルを編集することで選択の基準を容易に変更することができる。

3.3 結果の表示

得られた検索結果を入力者の Web ブラウザに表示するために、検索結果の書かれた HTML ファイルを生成する。変換後、入力者のブラウザに生成したファイルを表示する。

4. 結果

次に本システムで症例データを入力し、検索結果した画面を示す。

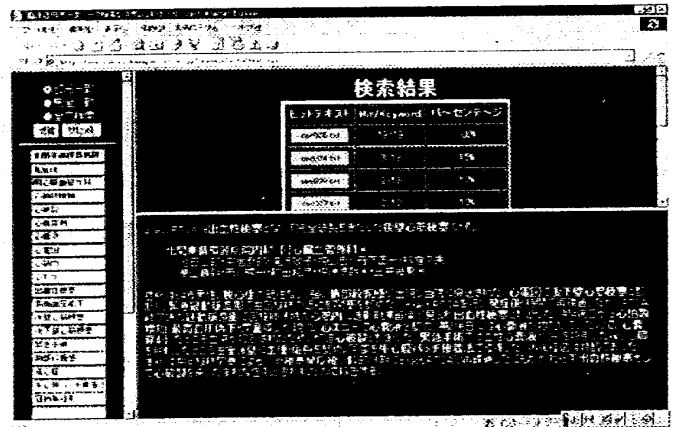


図2 症例データの検索結果

図2の左側に縦に並んでいる文字群が自動抽出されたキーワードになっている。このキーワードで検索した結果が右上に表示されている。今回の入力データでは12個の登録データに合致した。合致した医療データのうちより多くのキーワードを含んでいるものを右下に表示している。

自動抽出されたキーワードの中に、キーワードにならない語がいくつか含まれていた。これは、自然言語処理の意味解析を行わなかったためである。仮に、検索時間への制約が存在しないならば、これは膨大な辞書ファイルとの照合処理を行うことで解決できる問題である。

5. まとめ

利便性を重視するという本研究での開発方針において、今回構築したシステムは必要十分な結果を得ているものとする。

今後の課題として、過去の症例データとその診断結果の因果関係を用いた推論機構の実現があげられる。

参考文献

斎藤,他:ネットワークによる新しい医療情報システムの構築(第2報)ー推論エンジンとハイパーリンクによる診断支援ー,第13回ライフサポート学会予稿集(1997)