九州大学学術情報リポジトリ
Kyushu University Institutional Repository

専門用語の関係に着目した医療文献検索手法に関する考察

園田, 亮
九州大学システム情報科学研究科学府 : 修士

伊東, 栄典
九州大学情報基盤センター : 准教授 : 情報学

https://hdl.handle.net/2324/15954

出版情報：情報処理学会火の国情報シンポジウム2009. C-4-3, 2009-03. 情報処理学会
バージョン：
権利関係：

専門用語の関係に着目した医療文献検索手法に関する考察

園田 亮[†]　伊東 栄典[‡]

† 九州大学大学院システム情報科学府情報理学専攻 〒819-0395 福岡市西区元岡 744
‡ 九州大学情報基盤研究開発センター 〒812-8581 福岡県福岡市東区箱崎 6-10-1
E-mail: † ryo.sonoda@i.kyushu-u.ac.jp, ‡ itou@cc.kyushu-u.ac.jp

アブストラクト 近年，医療分野において研究成果を記述した学術論文が蓄積されている．従来これらのデータベースは，その分野の専門家が利用することが多く，登録された情報を研究や実務等に用いてきた．一方では，患者自身が検索を行い，医療文献を見つけたいという要求が生まれた.本論文は情報検索技術を用いて患者の医療文献活用を補助することを目的とする． 具体的には，患者が持っている情報から知りたい情報を含んだ医療文献を導き出す方法と候補文献の順位づけの方法について提案を行った.

A study of information retrieval method to search for medical references, with attention for the relation of technical terms.

Ryo SONODA[†] and Eisuke ITO[‡]

† Dept. of Informatics, Kyushu University 744 Motooka, Nishi-ku, Fukuoka, 819-0395 Japan
‡ RIIT, Kyushu University 6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581 Japan
E-mail: † ryo.sonoda@i.kyushu-u.ac.jp, ‡ itou@cc.kyushu-u.ac.jp

Abstract Recently, a lot medical papers are published and they have been accumulated. The more medical information become available, the more patients are interesting in academic papers. However, academic papers are difficult for non expert because those papers have a lot of medical terms. In this paper, we propose a new method of information retrieve to utilize medical references. We focus on similarity of medical terms to navigate medical references from poor medical knowledge. In a simple examination, we found that our method got good results.

1. 序論

1.1. 研究背景

近年，医療分野において研究成果を記述した学術論文が蓄積されている．米国国立医学図書館では，医学・生命科学分野に特化した文献データベース MEDLINE[7]を 1964 年から作成している．MEDLINE には毎年新たな文献情報が蓄積されており，2006 年には約 62 万件が追加されている．現在 MEDLINE には 1,600 万件を超える巨大な文献情報が保持されている．1996 年には Pubmed と呼ばれる無料で MEDLINE の検索を行うオンラインサービスが開始された．

また，MEDLINE には MeSH (MEDLINE Subject Headings)[9]と呼ばれる論文を特徴づけるためのカテゴリ分類語彙がある．MeSH は木構造をしており効果的な分類が可能である．

ほかの分野でも，特定の分野に特化したデータベースの整備が進んでいる．

一方では，患者自身が検索を行い，医療文献を見つけたいという要求が生まれた．一般的に患者は，医療知識が乏しく，治療を医師に一任したり，調べるとしても図書館に通い本を読んで病気に関する知識を得る程度であった．しかし，医師が治療を行う前に患者に十分に治療法についての説明を行うインフォームドコンセプトや実際に治療を行う前にほかの医師に意見を求めるセカンドオピニオンといった考え方の普及により，患者が納得した治療を受ける権利を求めるようになった.さらに医療技術は日々進歩しており，数年前までは治療法が確立していなかった病気の治療法が開発されたり，従来からの治療法ではなく最新の治療法が確立されることもある．これに加えインターネットの爆発的普及による情報収集の平易化や前述のようなデータベースの充実により患者が自身の健康問題の解決のために医療文献を検索したいという要求がでてきた．

個人向けに解放されている医療文献検索サイトも存在している．その例として個人向け医療文献検索サイトは医中誌[10]や JDreamPetit[11]がある

1.2. 研究目的

本論文の目的は情報検索技術を用いて患者の医療文献活用を補助することである[17]．現存の医療文献検索システムでは患者が利用する場合に問題がある．１点目は，患者が知りうるような単語や知識では曖昧な検索しかできず，検索結果が多くなり絞り込み検索も困難な点である.２点目は候補論文の表示順が内部ＩＤの順であったり，投稿日時の順などであり検索語に対する文献の重要度を考慮していない点である．

本論文では，以下の２点を解決することを主な目的とする．

1. 患者が持っている情報から知りたい情報を含んだ医療文献を導き出す方法
2. 候補論文の順位づけの方法

2. 関連研究

2.1. 医療文献情報サービス

2.1.1. 従来の医療文献検索システム

医療文献検索システムの最も有名なものとして 米国の国立生物工学情報センターが公開している PubMed があげられる．PubMed を使用することで世界最大の医療文献データベース MEDLINE の全文献を検索することができる．利用者は検索語のほかに著者名，学会名，論文形式，年齢などを制限を加えて検索を行うことができる．検索結果は論文に付与されている ID や著者名，登録日時などで並び替えることができる．

医療関係者限定の会員制のサービスとしてはメディカルオンラインがあげられる.医中誌 Web では日本国内で発行された医学，歯学，薬学，看護学などの関連分野の定期刊行物の論文情報を検索することができる．医中誌は法人だけでなく，個人でも利用することができる．JDreamPetit では MEDLINE のほか日本国内で発行された文献に対しても検索を行うことができる．

これらのサービスは専門家を対象としているため，現状では患者が使用するには支障がある．前述のとおり 患者が知りうるような単語や知識では曖昧な検索しかできず，検索結果が多くなり絞り込み検索も困難であり，候補論文の表示順に検索語に対する文献の重要度が考慮されていない．

2.1.2. 既存の医療情報提供サービス

患者向けの医療情報提供サービスの例として，がんナビ[12]がある．このサイトには，がんについての治療法などの情報や生活に対するアドバイスが掲載されている．しかしながら，あくまでも一般的ながんに対する情報提供を行うものであるため，医療についての専門的な情報は提供されていない．

患者が自身の健康に関する医療の専門情報を調べる場合には，医療文献検索を行う必要があるものの，従来からある医療文献システムでは専門家ではない患者が的確な文献を効率的に探し出すことは困難である．このように既存の医療文献検索サービスや，既存の医療情報提供サービスでは，患者が自身の健康に関する専門的な情報を効率的に調べることは難しい．

医療文献集合は，対象の母集団が大きいため医療文献検索には情報検索技術を適用することが望ましい．次節では，情報検索技術を医療文献に適用した先行研究について述べる．

2.2. 医療文献集合からの関連語発見手法

ここでは医療文献からの関連語発見手法について

述べる．関連項目を発見できれば検索語に近い用語を含む文献の提示が可能になり，専門知識を持たない人の文献検索に応用することができる．

2.2.1. Swanson の手法

Swanson[1][2]らは ABC モデル(図 1)の提案，ABC モデルを用いた新たな項目間の関係の発見方法の提案（2 種類）を行い Arrowsmith と名付けた計算機を用いた文献からの知識発見システムを開発している．

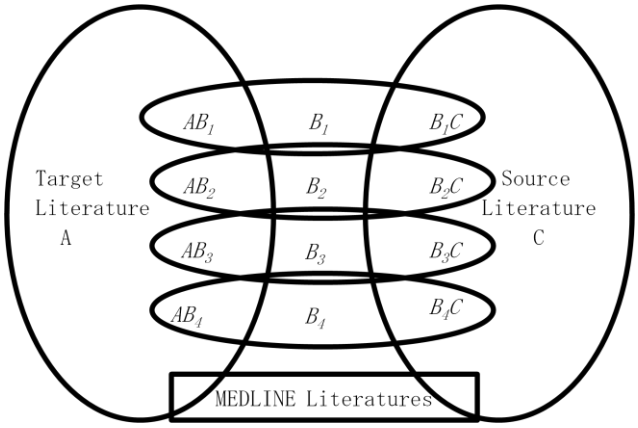

図 1 ABC モデル

このシステムは医療文献データベース MEDLINE を対象としており，利用者が単語 A と単語 C を入力すると，その二つの単語と互いに関係のあると思われる単語 B を共起関係に基づいて算出し，A と B，B と C が共起した論文タイトルを表示する．その結果である B を介して，未発見である A-C の関係を導き出すことを支援する．

2.2.2. Surinivasan の手法

Surinivasan[3]らは ABC モデルの考え方を基にして，物質と病名の新たな関係を導き出すためのアルゴリズム(open discovery algorithm)を提唱している．

物質名(A)と MeSH 用語(B)の関連性を共起関係により重みをつける．次に，同様の方式を用いて，重みの大きい MeSH 用語(B)から病名(C)を導出する．

Open discovery algorithm を用いて今まで知られていない新たな関係を導出できたため，Surinivasan らは提案アルゴリズムが未知の関係発見についても有効であろうと述べている．

3. 患者側への医療文献活用方法

本章では患者の医療文献活用方法について述べる

3.1. 患者側が保持している情報

まず，患者側が保持している情報にはどのようなものかを考える．医療文献に含まれる情報は，専門性が非常に高い．そのため，医療関係者でないものは知る機会が少ない．通常患者は断片的で僅かな情報しか持っていない．患者が知りうる情報は，自覚症状から分かるもの，病院からの通知で分かるもの，投与された薬から分かるものに大別できる．自覚症状から分かる情報の項目としては，症状や違和感のある箇所(臓器名や器官名)がある．病院からの通知で分かる情報には病名や検査結果があげられる．投与された薬から分かるものには薬の名前や薬の成分(化学物質名)がある．これらの知識からでは， 自身の手で膨大な医療文献から必要な情報を得るのは困難である．

本研究ではこれらの患者が保持している僅かな情報を基に患者の欲しい情報を導き出す手法を検討する．

3.2. 患者側が必要とする情報

次に健康問題の解決につながる文献とはどの様な内容が記載されている文献であるかについて考える．患者が欲しい情報とは専門家が求める情報とは異なる．専門家は病気の原因や相互関係を調査し，教育，研究，実験に役立て医療技術の発展を目指す．これに対し患者側としては今現在存在する医療技術で最大の効果をあげることを目的とする．具体的には最新あるいは現在受けている治療とは異なる病気の治療法、注意すべき合併症，関係のある病気や臓器名， 渡された薬以外の薬(主に，未だ広範囲に認知されていない新薬)、またはその薬の情報，あるいは副作用情報などが患者が知りたい情報であるといえる．

3.3. 提案手法

患者が保持する情報から，患者が必要とする情報へ導く方法を提案する．まず医療文献集合から通常患者が知ることが困難な項目相互間の関係を導出する．次に医療文献を見つけ出す過程でこれらの関係情報を付与する．最終的にはそれらの情報が載った医療文献の発見および重要度の程度を数値で表す．

3.3.1. 患者保有情報から医療文献を導き出す方法

患者が持っている少ない情報から徐々に追加情報を引き出し，最終的に患者が欲する情報にたどり着くための導き方について述べる．医療文献集合から通常患者が知ることが困難な項目相互間の関係を情報検索技術をもちいて導出する．本研究では具体的な項目相互間の関係計算手法としてコサイン類似度と上位下位関係(intersection rate)を用いることとした.単語間のコサイン類似度を算出することにより似ている単語を結びつけることができる．検索結果に検索語と似ている単語群の情報を付加することで，検索語のみではとらえ切れなかった文献情報を得ることができる．また，単語間の上位下位関係を算出することで単語群を階層化することができる．検索語よりも下位に相当する単語群の情報を付加することで，より詳細で情報豊富な文献情報を得ることができる．

これら関係を用いて患者が通常知ることの無いあるいは無関係であると思い込んでいた単語と患者が入力した検索語を結びつけ，拡張検索を可能とする．

結果として医療文献を見つけ出す過程でこれらの項目間関係情報を付与することになる．このようにして，検索語を追加していくうちに必要とする文献を絞り込むことによって 患者の必要とする情報を既知の情報から必要とする情報が載った医療文献の発見することができる．

3.3.2. 候補論文の順位づけの方法

次に，患者にとってその文献がどれだけの価値があるかについて重要度を定義する必要がある．本節では重要度の算出方法の方針について大まかに述べる．なお，6 章にて具体的な定式化を行う．

検索結果の文献に対して情報検索技術を用いた重み付けを行うことで重要度を定義する.まず，入力された検索語(クエリ)の各文献に対する重要度を算出する.全ての検索語に対してこの操作を行い，これらの和をとることにより検索語に対する文献 d_k の重要度を算出する．この重要度を D1 とする．

次に，検索語と拡張語との関連性を検索結果に反映をする必要がある.項目間関係情報を付与するために，拡張語の各文献に対する重要度と，検索語と拡張語間の関連度合を表す数値である関連度係数(図２)を用いて文書 d_k 文書の重要度を算出する．これを D2 とする．

図 2　検索語と拡張語間の関連

さらに，医療では古い文献に記載されている既知の情報である場合が多く，必要とする情報は新しい文献ほどよいと考えられる．そのため時系列での評価を取り入れる．

これら３つの計算手法を組み合わせ最終的な文献の重要度の数値を算出する．次章では，分析の対象とした医療文献について述べる．

4. 医療文献

本章では分析の対象とした医療文献について述べる

4.1. 分析対象データ

本研究では分析対象とするデータベースは先行研究と同じ MEDLINE を分析対象とした.MEDLINE は，医学や生物学分野で，最も有名かつ世界最大の文献 DB である．

MEDLINE の全文献を分析対象にするには数が多すぎるため，文献の数を制限する．本研究では，ワールドフュージョン社が提供する知識ベース，LSKB(Life Science Knowledge Bank)[8]において疾患名が「肺がん」(実際の検索語としては「Lung cancer」，「Lung neoplasms」)と記載されている文献で，なおかつ掲載年が 2003 年～2007 年となっているものを実験の第一段階として文献検索により抽出した．本格的な調査のために「高血圧症」(Hypertension)が記載されている最近 25 年の文献(1984 年～2008 年)を抽出した．なお，ここで記載されているとはタイトル，アブストラクト，MeSH のいずれかに記載されていることを指す．

取得を試みた肺がんに関する PubMed ID のリストの数は 23,467 件，高血圧症に関するものは 173,257 件であった．それぞれの MEDLINE 全体に占める割合は 0.15%, 1%であった．第一段階の肺がんを分析対象としたのは，我々の身近な健康問題であり喫煙との関係など興味深くわかりやすい分析結果が得られやすいと考えたからである．また,本格的な調査のために高血圧症を分析の対象としたのは，高血圧症のほうが情報検索技術を用いて処理を行うのに肺がんに比べて適した規模であったためで，高血圧症のほうが副作用や関連疾患などの情報が得たいというニーズがあった為である．そこで項目間類似度の数値化実験には高血圧症に関する文献を用いた.各論文の 2 次情報データのうちアブストラクトは文献の内容を簡潔にまとめた文章であるため本研究ではアブストラクトを主な分析対象とした．

4.2. 医療文献特有の問題

医療文献に特徴的な問題としてシノニム及びホモニムがある．生物学分野には名前を付けるタイミングや，同じものを別のものと思い込んでいた等という歴史上の理由からホモニム，シノニムが存在する．１つのものが複数の名前 をもつ(同物異名)とき，これらはシノニムであるという．例として，アサガオの学名は Ipomoea nil と Pharbitis nil がありこれらはシノニムである[24]. 複数の異なるものに対して同じ名前がつけられている(異物同名)とき，これらはホモニムであるという．分かりやすい例として「はし」には食事に使う「箸」や川にかけてある「橋」があるがこれらはホモニムである．ホモニムとシノニムが複合している場合もある．例えば化合物名「CYCLOSPORINE」は「Sandimmun Neoral」，「Sigmasporin Microoral」，「Cyclosporin A」，「Cyclosporine」それぞれのシノニムとして登録されている．この場合「CYCLOSPORINE」はこれらのホモニムである.文献や学者ごとに使用する名称が異なる場合もある．

5. 分析

本章では取得データの分析方法について述べる．

専門用語間の関係を知るためには専門用語のみを分析の対象とする必要がある.専門用語の分析は共同研究を行っているワールドフュージョン社が保有する専門用語辞書を用いた．分析対象は「高血圧症」を含

む医療文献である．この専用用語辞書は，専門用語の分類だけではなく，その専門用語がどの文献で出現したかについても調べることができる．専門用語の分類は，遺伝子，化学物質名，病名，臓器名の 4 つである．この辞書を用いて「高血圧症」に関する論文について分析を行った..

5.1. コサイン類似度

専門用語間の類似度合を調査するためコサイン類似度を用いる．文書ベクトルモデルで表現すると，文書間あるいは単語間の類似度をコサイン尺度で評価できる[14]. 文書 1,2...n で各文書が単語 A,B を含む数(TF)をベクトルで表すものとする．

このとき単語 A と B のコサイン類似度を以下の式で計算する．

$$\vec{A} = (a_1, a_2 \ldots a_n)$$

$$\vec{B} = (b_1, b_2 \ldots b_n)$$

このとき単語 A と B のコサイン類似度を以下の式で計算する

$$\cos(\vec{A}, \vec{B}) = \frac{a_1 * b_1 + a_2 * b_2 + \ldots + a_n * b_n}{\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} * \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}}$$

例えば「心臓」(heart)のコサイン類似度が高い語句は，「心臓病」(herat disease)「心筋」(myocardial)などであり，これらは心臓に関係する語句である.入力した専門用語に類似した専門用語を拡張語とすることで拡張検索を行うことができる．

5.2. 上位下位関係

専門用語間のつながりを階層化し，専門用語の上位語，下位語をそれぞれ算出する．

本研究では，用語 s, t の出現回数 DF(s), DF(t)を用いて各専門用語の上位下位関係を以下に示す Intersection Rate で算出する[15][16].

$$|DF(t)| > |DF(s)| \text{かつ} \frac{|DF(t) \cap DF(s)|}{|DF(s)|} > \alpha$$

本研究では α の値を 0.5 と定めた．つまり単語 s の出現する文献のうち半分を超える文献に単語 t が含まれていれば t は s の上位語である．

上位下位関係の具体例として「心臓」(heart)の下位語は「心筋虚血」(myocardial ischemia)や「拡張型心筋症」(dilated cardiomyopathy)であった.入力した専門用語の下位語を拡張語とすることで拡張検索を行うことができる．ここで下位語のみを拡張検索の対象としたのは，上位語をいれると検索対象文献数が大きく増加し探したい文献が見つかりにくくなり，下位語のほうがより詳細な情報を得ることができると考えられるからである．

6. 提案手法の詳細化

6.1. 候補論文の順位づけ

本節では，候補論文の順位づけの方法について述べる． 重要度の決定方法について 3 つの評価基準に分割し，それぞれ定式化する．

6.1.1. 検索語に対する文献の重要度

患者にとってその文献がどれだけの価値があるかについて重要度を定義する必要がある.検索結果の文献に対して情報検索技術を用いた重み付けを行うことで重要度を定義する.入力された検索語(クエリ)の重要度を算出する．

入力された検索語: $Q = \{q_1, q_2 \ldots q_n\}$

ある検索語 q の重み: $= weight(q)$

入力された全ての検索語に対してこの操作を行い，和をとりことにより検索語に対する文献 d_k の重要度を算出する.この重要度を D1 とする．

$$D1(d_k) = \sum_{q \in Q} weight(q).$$

6.1.2. 検索語の拡張語に対する文献の重要度

検索語と拡張語との関連性を検索結果に反映をする必要がある(図 4).そのため検索語に対する文書の重要度に,出現する拡張語に対する文書の重要度に関連度係数を掛けたものを全て加えることで，ある文書 d_k 文書の重要度を算出する．これを D2 とする．

検索語から関連性が定義される拡張語:

$$R = \{r_1, r_2 \ldots r_m\}$$

関連度係数: $\alpha_{i,j}$ ただし $i \in Q, j \in R,$

$$D2(d_k) = D1(d_k) + \sum_{q \in Q} \left(\sum_{p(u,v)} \alpha_{u,v} weight(r_v) \right)$$

P(u,v)とは入力語 q_u と拡張語 r_v について関連度係数 $\alpha_{u,v}$ が定義されることとする．

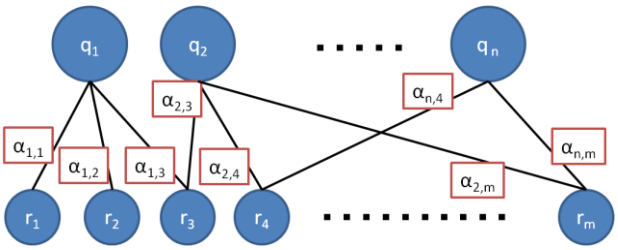

図 3 検索語と拡張語間の関連度係数 α の概念図

関連度係数は追加情報を引き出すための値である.(図 3 参照)

6.1.3. 時系列評価

医療では古い文献に記載されている既知の情報で

ある場合が多く，必要とする情報は新しい文献ほどよいと考えられる．そのため時系列での評価を取り入れる.評価法は，対象期間を１とし経過年月の割合により０～１までの評価を与える．ただしこの場合時系列の評価値により，文献の重要度が大きく変化する可能性がある.そのため，全体に１を加えることで１～２までの評価を与えることとした．

ある文献 d_k の時系列評価は以下のような式で定義する．

$$T(d_k) = 1 + \frac{\text{経過年月}}{\text{対象期間}}$$

これら３つの計算手法を組み合わせ最終的な文献の重要度の数値を算出する．それぞれ D3,D4 とする．

$$D3(d_k) = D2_add_T(d_k) = T(d_k) + D2(d_k)$$

$$D4(d_k) = D2_multiply_T(d_k) = T(d_k) * D2(d_k)$$

これらの数値を計算し，７章で述べる試作システムを作成した.さらに専門家に依頼し計算手法ごとの比較，評価を行う．

6.2. 単語の重みづけ

本節では単語の重みづけについて述べる．単語の重みは検索結果の文献の重要度を定義するために必要となる．本研究では入力されたある検索語 q の重要度を求めるために局所的重みと大局的重みを使用する．局所的重み付けとして TF を用い大局的重みづけには文書頻度の逆数(inverse document frequency:IDF)を用いる.これらの積を検索語に対する文書の重要度と定義する．

ある単語 q が文献に出現する場合以下のように単語の重み weight(q)を定める．

$$weight(q) = TF(q) * idf(q) = TF(q) * \log\left(\frac{\text{全文献数}}{DF(q)}\right)$$

6.3. 関連度係数

単語間の関係を正しく定めることができるコサイン類似度及び上位下位関係の値を関連度係数として使用する．入力語 q_i と拡張語 r_j との間の関連度係数 $\alpha_{i,j}$ は以下のように定める．

コサイン類似度を拡張語導出の指標とする場合，

$$\text{関連度係数}\ \alpha_{i,j} = \cos(\vec{q}_i, \vec{r}_j),$$

上位下位関係を拡張語導出の指標とする場合

$$\text{関連度係数}\ \alpha_{i,j} = \frac{|DF(q_i) \cap DF(r_j)|}{DF(r_j)}$$

ただし $DF(q_i) > DF(r_j)$

これらの値は実験段階では専門用語ごとに分けて計算を行っていたが，次章で述べるシステムを実装する段階では病名，化学物質名，臓器名３つの専門用語群を混ぜた 28,753 語を分析の対象とした．遺伝子に関する専門用語ついては実験段階で大量の本来関係のない遺伝子どうしに関係が定められてしまったため除外している．

7. 試作システム

本章では，提案手法の計算手法を実装した試作システムについて述べる．システムは WebCGI を用いて作成し，利用者は Web ブラウザを通じてシステムを利用することができる.拡張語導出にコサイン類似度もしくは上位下位関係を用いることができる．どちらを用いるかは利用者が自由に決めることができる．

画面上部に検索したい語句を入力する入力欄がある．ここには病名，化学物質名，臓器名３つの専門用語群に出現する 28,753 語を入力することができる．検索結果は画面下部のフレームに表示される．まず入力された語句に対する拡張語が全て表示され，その下に検索結果の文献が各計算方法に対して重要度の順で上位 20 位まで表示される．

計算手法は６章で述べた D1,D2,D3,D4 を用いる.なお D1 の手法は従来の情報検索手法そのものである．

7.1. 拡張検索機能

高血圧症を対象として患者が保持する情報から患者の知らない情報を拡張語として導き出し，その拡張語の重みを検索対象文献に付加し患者が必要とする情報を見つけ出す試作システムを作成した．

コサイン類似度の値は 0.1 を閾値とし，上位下位関係は α >0.5 を条件とした．

システムの実行図を図 5 に示す．現状のシステムでは検索結果が出るのに多少時間がかかる．特に検索語に対する拡張語が多くなるとこの傾向が強くなる．

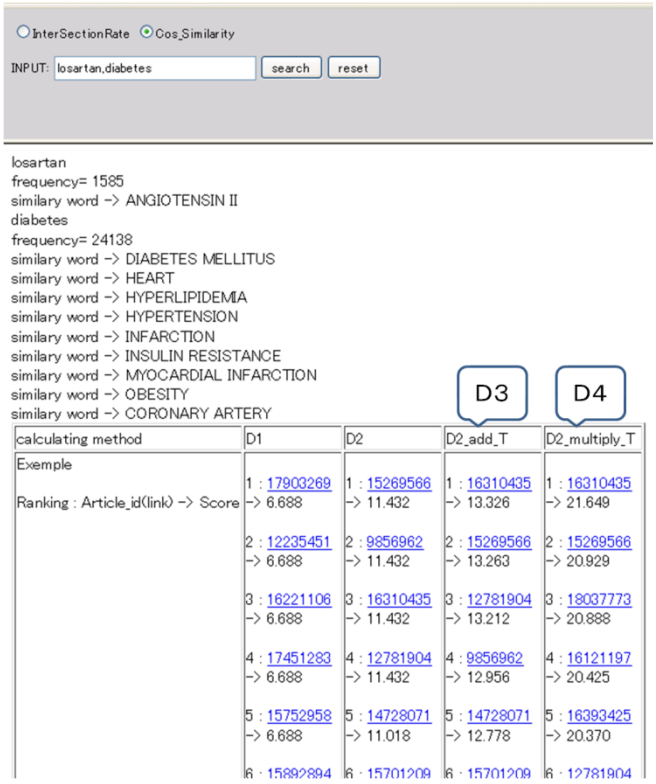

図 4 試作システム

7.2. 評価及び考察

３人の専門家に人出で評価していただいた．試作システムに高血圧症に関係が深いと考えられている単語を数パターン入力しその結果を評価の対象とする．

主な評価項目は，以下のとおりである．

1. 拡張検索を行う際の，拡張語が入力語に対して適切か否か
2. 導出された論文の内容が入力語に対して適切か否か
3. 従来の検索法との比較

検索語に対する拡張語の中にあまり関係のないものも含まれていることがあり，拡張語の制限が甘い傾向があった．拡張語が正確な場合は関連語の出現により文献が上位にカウントされるため，情報量の豊富な文献が優先される傾向があった．この結果から入力語に対する拡張語が余分に選定され，結果として文献の重要度に余計な拡張語分の補正が余計に加算されていることがわかる.導出された論文の内容が入力語に対して適切か否かについては，拡張語が正確に算出されているかどうかで左右されている.また，従来の検索手法では検索されない文献が検索結果に表れていた．

このためコサイン類似度の閾値を 0.2，上位下位関係の閾値をα>0.6 に引き上げて再実験を行った．コサイン類似度に関しては losartan と diabetes の組み合わせについて完璧に近いほど関連語が抽出されていた．これは閾値の引き上げにより関連語の絞り込みができたといえる．文献の内容も全体的に無難な文献が上位にリストアップされていた．しかしながら上位下位関係では，病名の関連語にアルファベット 2 文字のみの（例えば DM）ものがあり，これが文献を選びこむ際のノイズの原因になっている可能性があった．また全体的に上位下位関係では関連性の薄い文献が取れてきがちであった．

8. 結論

本論文は情報検索技術を用いて患者の医療文献活用を補助することを目的とした.具体的には，患者が持っている情報から知りたい情報を含んだ医療文献を導き出す方法と候補文献の順位づけの方法について提案を行った．

まず, MEDLINE から分析対象とする肺がん，高血圧症に関する文献を取得した.次に，検索語と拡張語間の関連性を数値化した．このための計算手法にコサイン類似度と Intersection Rate に基づく上位下位関係を用いた.これを用いて，検索された文献に対し重要度を定義する式を提案した．さらに，上記の計算を行い，結果を表示する簡単な試作システムを作成した.また，特定トピックの例として副作用に焦点を当て，特定トピックに関する論文の選集方法について述べた．

コサイン類似度，上位下位関係を用いることで，関連性の高い専門用語どうしが関連づけられていた．これにより医療文献に対しこれらの手法が有効だと考えられた．試作システムについての評価については専門家からみての意見ではあるが検索結果の文献も関連度の高いものが取れていた．

今後の課題について述べる.今回は時間がなく，患者から見た意見が取れなかった.このため患者から見てこの医療文献検索手法が本当に的を得た結果であるかどうかが不明である．

今後は共同研究をしているワールドフュージョン社を通して，日経が運営する患者向けのナビゲーションサイトであるがんナビ[12]での試験的な運用を視野に入れている．ここで患者から見た意見を得られると考えている．

また，提案手法についての定量的な評価を行う必要がある．人手での評価では良い結果が得られたものの，主観的な評価であるため，客観的に評価するためには定量的な評価が必要である．今後，いくつかの検索語に対しての検索結果を定量的に評価し本論文の提案手法の良さを示していきたい．

今回は肺がんと高血圧症に限定して実験や試作システムを作成した.コサイン類似度および上位下位関係について一定の閾値を設定し実験を行った．特に上位下位関係については試作システムでの運用時の評価が芳しくなかった.このため閾値の調整を行うことで

よりよい結果が期待できる.さらに，このアルゴリズムは肺がんや高血圧症だけではなく他の対象についても同様に適用可能である．医療文献内の他の項目や，医療文献以外のデータに適用することで，このアルゴリズムの汎用性を検証したいと考えている．

謝　辞

本研究は，株式会社ワールドフュージョン社との共同研究によるものである．

文　献

[1] Don R. Swanson, Neil R. Smalheiser ”An interactive system for finding complementary literatures: a stimulus to scientific discovery,” Artificial Intelligence, Vol. 9, No. 1, pp.183-203, 1997.

[2] Don R. Swanson, and Neil R. Smalheiser and Vetle I. Torvik, “Ranking Indirect Connections in Literature-Based Discovery: The Role of Medical Subject Headings,” J. Am. Soc. Inf. Sci. Technol., Vol. 57, No. 11, pp.1427–1439, 2006.

[3] Padmini Srinivasan, and Bisharah Libbus, “Mining MEDLINE for implicit links between dietary substances and diseases,” Bioinformatics (Oxford, England), Vol. 20 Suppl. 1, pp. 290–296, Aug., 2004.

[4] Yasunori Yamamoto, Toshihisa Takagi,”Biomedical knowledge navigation by literature clustering” Journal of Biomedical Informatics, 40 pp.114–130, 2007.

[5] Smalheiser,N.R. and Swanson,D.R. (1998) Calcium-independentphospholipase A2 and Schizophrenia. *Arch. Gen. Psychiatry*, **55**,752–753.

[6] Srinivasan,P. (2004) Text mining: generating hypotheses from MEDLINE. J. Am. Soc. Inf. Sci. Technol., **55**, 396–413.

[7] NLM, MEDLINE,1964

[8] 株式会社ワールドフュージョン，LSKB(Life Science Knowledge Bank), http://www.w-fusion.com/J/productlist/lskb.html

[9] NLM. MeSH. (Medical subject headings), 1960. National Library of Medicine.

[10] 医中誌 web:http://www.jamas.or.jp/

[11] JDreamPetit:http://pr.jst.go.jp/jdreampetit/

[12] がんナビ : http://cancernavi.nikkeibp.co.jp/

[13] Wikipedia:http://ja.wikipedia.org/wiki/

[14] 北研二，津田和彦，　獅子堀正幹，情報検索アルゴリズム，共立出版，2002.(ISBN:4-320-12036-1)

[15] Rui Li, Shenghua Bao, Ben Fei, Zhong Su, Yong Yu, Towards Effective Browsing of Large Scale Social Annotations, International World Wide Web Conference archive Proceedings of the 16th international conference on World Wide Web, 2007.

[16] Yoshihiro Shimoji, Sachio Hirokawa, Dynamic Thesaurus Construction from English-Japanese Dictionary, International Workshop on Ontology Alignment and Visualization, 2008.

[17] 園田 亮，伊東 栄典，高崎 哲也，川原 弘三,医療文献からの類似項目発見についての考察,電子情報通信学会第 19 回データ工学ワークショップ (DEWS 2008),2008.03