社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－GN－49 (26)
2003／10／24

医療従事者に対する自動論文検索システムの構築および関連論文検索支援と研究戦略支援の提案

折田 憲始[1] 森川 富昭[2] 能瀬 高明[1]
矢野 米雄[3] 西野 瑞穂[2] 森口 博基[4]

臨床・研究・教育を職務として行っている大学病院の医師(研究者)は，独立行政法人化のため，経営にも参画しなければならなくなる．そのため，研究に割ける時間は減少すると考えられる．そこで，我々は医師の研究を支援するために，自動で論文を検索しメール配信する自動論文検索システムを構築した．評価実験により，本システムが有益であることが示唆された．また，更なる医師への研究支援を目的とし，以下の二つの支援を提案する．

第一に関連論文検索支援を行う．今までは見つからなかったが，研究を発展させる上で必要となるであろう論文を医師へ提示することを目的とする．第二に，研究戦略支援を行う．医師の研究レベル向上のため，検索キーワードとAbstractから取得した索引語をユーザプロファイルと見立て，医師の研究の戦略(方向性)を支援する．

The construction of PubMed Check System for medical workers

ORITA Kenji [1], MORIKAWA Tomiaki [2], NOSE Takaaki [1],
YANO Yoneo [3], NISHINO Mizuho [2], MORIGUCHI Hiroki [4]

The Medical doctors in the university hospital handle clinical, the research, and the education now. The university becomes making to the independent administrative agency in 2004. Then, the doctor should do not only clinical, the research, and the education but also the hospital management in the future. Therefore, the time that the doctor researches shortens than before. Now, all doctors cannot effectively use the computer. The knowledge sharing among doctors using the computer is imperfect. Then, to support the research of the doctor, we constructed PubMed Check System that was able to do the paper retrieval automatically.

1 はじめに

現在，大学病院の医師は臨床・研究・教育を職務として行っている．しかし，今後，大学は独立行政法人化になるため，医師は臨床・研究・教育のみならず，経営にも参画しなければ

[1]徳島大学 工学部 知能情報工学専攻
[2]徳島大学 歯学部 歯学科 小児歯科学講座
[3]徳島大学 工学部 知能情報工学科 知能工学講座
[4]徳島大学 医学部 附属病院 医療情報部

ならない．そのため，医師が研究に割ける時間は減少すると思われる．

現在，医師は論文検索にPubMedを利用することが多い．PubMedとは，1997年からサービスを開始したNLM(National Library of Medicine，米国国立医学図書館) によるプロジェクトで，医療系英語論文(医学，歯学，薬学，生物学，看護学分野)を無料で検索できる有名なWebページである[1], [2]．PubMed全体には，2003年 9月 現在，約 1270万件という膨大な量の論文が登録されている．また，現在も一日に約 3000件のペースで日々論文が登録されている．

しかし，PubMedは利用者のキーワードにより論文を検索するパターンマッチングの検索手法である．そのためキーワードに近い類義語的な情報は検索対象にはならない．また論文検索のアクションも利用者主導であるため，利用者の時間的都合によっては必要な論文を早急に得られない等の問題がある．そこで，我々はPubMedを利用し，医師に対して以下の三つの支援を行うPubMed Check Systemを構築する．

1. **自動論文検索支援** … 一度のキーワード登録で，日々PubMedを自動で論文検索し，結果をメール配信する．

2. **関連論文検索支援** … 医師が自動論文検索で用いたキーワードと，そのキーワードでヒットした論文のAbstractから得られた単語(索引語)でPubMedを再検索し，今までは見つからなかったが，研究を発展させる上で必要となる論文の発見を支援する．

3. **研究戦略支援** … 医師が自動論文検索で用いたキーワードと，関連論文検索で求められた索引語を，医師を表すユーザプロファイルと見立て，他の医師と比較することで医師間の距離を測り，研究戦略(研究の方向性)を支援する．

実際に自動論文検索の機能を構築し，小児歯科医などに評価実験を行った．その結果，自動論文検索の有効性が証明された．

本稿では，**2.**で自動論文検索支援のシステム構築について述べ，**3.**で関連論文検索支援，**4.**で研究戦略支援の手法について提案を行う．

2 自動論文検索支援

日々発表される最新の論文は，研究を発展させるために重要である．しかし，発表される論文の量は膨大であり，その中で必要なものを常にチェックするのは難しい．

そのため，興味のあるキーワードを一度登録するだけで，毎日論文検索を行う自動論文検索システムを構築した．

本システムを使用すれば，論文を検索する時間が短縮ができ，最新の論文の見逃しもなくなる．また，臨床や学生への教育といった日々の職務の中で疎かになりがちな研究意欲が，毎日配信されるメールによって喚起されるといったメリットもある．

2.1 システム構成

自動論文検索システム(図1 参照)はLinux上で稼働している．システム構成を以下に示す．

- OS: Red Hat Linux 7.3
- BioRuby version 0.5.1
- Http Server version: Apache/2.0.40
- DataBase version: PostgreSQL 7.2.2
- PHP version: 4.2.2

医師が検索キーワードを登録するWebページをPostgreSQL，Apache，PHPで構築した．Linuxのcronデーモンによって，検索キーワードにヒットする論文を毎日23時に自動取得し，sendmailにより医師の元へと配信している．また，PubMedから論文を取得するためにBioRubyのpubmed.rbというモジュールを使用している．

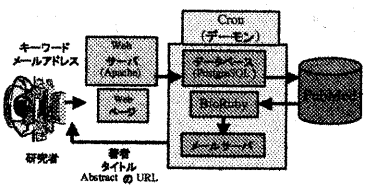

図 1: 自動論文検索の流れ

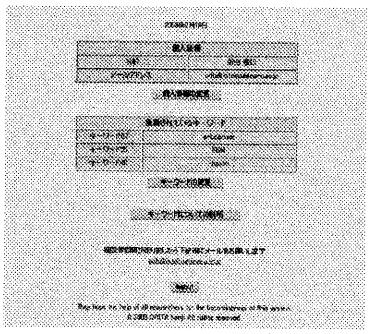

図 2: キーワード登録

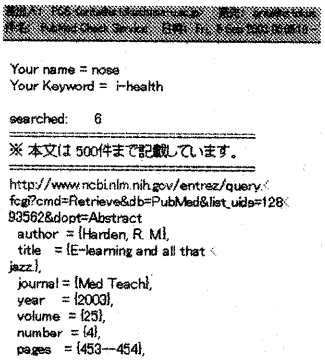

図 3: メール内容

2.2 自動論文検索の流れ

医師は，図2に示すWebページから，自分のメールアドレスと検索したいキーワードを登録する．本システムは，PubMedに日々登録される論文の中から，キーワードを含む論文を検索する．検索論文が見つかった場合は図3に示すように，論文数，著者，タイトル，掲載ジャーナル，年，巻，号，ページ情報及びAbstractと対応したURLリンクをテキスト形式で利用者にメール配信する．

図3に示すように，各論文の冒頭には，その論文のAbstractと対応したURLがリンクされているため，クリックするだけで，その論文のAbstractが見られる．

また，キーワードを含む論文がなかった場合も「キーワードを含む論文はありませんでした」という内容のメールが送信されるため，自分が投稿しようと思っている内容の論文が投稿されていないとわかる．

2.3 自動論文検索の評価実験

実際に，小児歯科医など9名に10日間本システムを使用してもらい，システムの利便性および要望等についてのアンケートを実施した．

アンケート回答数は9名となり，「今後も本システムを使用したいか」という問いに対して，五段階評価一位の「大変思う」が5名，二位の「やや思う」が4名と良好な結果が得られた．「自分が投稿した論文がPubMedに登録されたのがわかって感激した」といった意見も聞かれた．また，実際に配信されたメール内容を見ると，登録した検索キーワードが不適切であったため，全くヒットしないまま評価実験が終わってしまった医師がいた．その医師の意見としては「キーワードに近い情報をシステムが自動的に検索できないか？」「キーワードに近い情報を辞書としてシステムが持っていないのか?」等があった．そこで我々は利用者の意見・要望を基に，関連論文検索支援を提案する．

3 関連論文検索

例えば，歯科用語で虫歯を意味するcaries(齲蝕)を検索キーワードとして入力したとする．しかし，PubMedで全文検索を行っても，同じ虫歯を意味するdecayed toothを含む論文が見つからないことがある．

このように，パターンマッチングの検索手法では見つからなかったが，研究を発展させる上で必要となるであろう論文を医師へ提示することを，関連論文検索支援の目的とする．

3.1 関連論文の検索手法

まず，検索にヒットした論文のAbstractをデータベースに保存し，論文の内容を特徴付けるような単語(索引語)を見つける．この時の手法は，単語頻度と，文書頻度の逆数から索引語を見つけるTF*IDF法[3]を用いる．こうして見つけた索引語で再びPubMedを検索し，今までのキーワードでは見つからなかった新しい論文を医師へ提示する．

これにより，今まで自分とは無関係と思っていた単語や，思いつくことができなかったキーワードでPubMedを検索するため，研究を発展させる上で重要な論文が見つかる可能性がある．

例えば，ある医師が検索キーワードに，EBM(evidence-based medicine)を登録したとする．自動論文検索システムをある程度の期間使用すると，EBMを含むAbstract群をデータベースに保存できる．

このAbstract群の中から"Brill's tagger"[4]を使用して名詞のみを抽出する．こうして得られた単語(名詞)の頻度と，文書頻度の逆数から求められた索引語は，EBMと関連性が強い単語になる．

3.2 TFの実証と改良

実際に"EBM"でPubMedを検索してAbstract群を取得し，まずは，TF(Term Frequency; 索引語頻度)によって索引語候補を見つけた．

この時，当然ながら一つの論文から一つのAbstractが取得できるが，Abstractは数文程度の文章量なので，頻度で索引語候補を見つけるには短すぎる．そのため，一つのキーワードでヒットした論文から取得した全てのAbstractを一つの文書とみなす．

表 1: EBMでヒットした文書の索引語候補

順位	索引語候補	TF
1	evidence / NN	8
2	practice / NN	8
3	Infection / NN	7
4	control / NN	6
5	EBM / NNP	4

表 2: 品詞記号とその意味

品詞記号	意味
NN	名詞(単数，質量)
NNS	名詞(複数)
NNP	固有名詞(単数)
NNPS	固有名詞(複数)

こうしてEBMでヒットしたAbstract群の中から，TFによって見つけた索引語候補が表1になる．第一位の索引語候補としてevidenceを抽出することができた．システム自体にEBMの意味となるevidence-based medicineという用語を入力せずとも，その意味の単語が結果としてわかったことになり，Abstract群を対象としたTFの有効性が示された．索引語候補につけられた品詞記号の意味を表2に示す．

3.3 IDFの導入と実証

次に，索引語には出現頻度の偏りがあるため，大局的重みを計るIDF(Inverse Document Frequency; 文書頻度の逆数)を考慮する．これは，例え高頻度の語であっても，どの文書にも現れるような語は文書を特徴付ける語にはならず，文書集合の中で偏って現れる語の方が，文書を特徴付ける単語になるであろうという考えからである．

実際にhealth，child，familyの三つのキーワードでPubMedを検索して，TF*IDFの値を求め，各々の上位5個をまとめたのが表3，4，

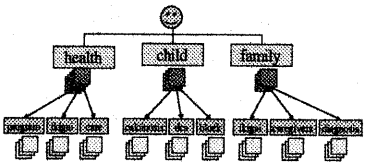

図 4: ユーザプロファイル

5になる．health，child，familyで検索すると，各々14，6，6本の論文がヒットした[1]．これらの合計数26本を文書総数 n と見立てる．

表の中で示している f_{ij} は出現頻度であり，Abstract群の中で単語が出現した回数である．TFは索引語頻度であり，出現頻度の高い索引語に過大な重みを与えないようにするために対数化を行った対数化索引語頻度(logarithmic term frequency)を使用する．$\log(1+f_{ij})$ という計算式によって求める．n_i は文書頻度であり，単語を含む文書の数である．IDFは文書頻度の逆数であり，$\log(\frac{n}{n_i})$ によって求める．これらの表は，TF*IDFの項目をキーとして，降順に並んでいる．

表5より，familyというキーワードは，ikigai(生きがい)，caregivers(介護人)，cancer(癌)などといった単語が近いとわかる．こうして見つけた索引語によってPubMedを再検索する．これにより，今までは発見することが出来なかった新たな論文が得られる．

4 研究戦略支援

医師の研究レベル向上のため，**2.**で述べた自動論文検索，および**3.**で述べた関連論文検索を基として，医師の研究戦略(方向性)を支援する．

自分と他の医師との距離が具体的な数値で比較できると，自分の研究の位置付けがわかり，研究を進める方向性の指針になるであろうという考えに基づいている．

[1]ヒットする論文数を絞るため，日本の雑誌に投稿して，2003年8月中にPubMedに登録された論文の中から検索した

表 3: healthでヒットした文書の索引語

単語	f_{ij}	TF	n_i	IDF	TF*IDF
program	22	1.36	2	1.11	1.52
ikigai	15	1.20	2	1.11	1.34
time	14	1.18	6	0.64	0.75
support	16	1.23	8	0.51	0.63
care	22	1.36	9	0.46	0.63

表 4: childでヒットした文書の索引語

単語	f_{ij}	TF	n_i	IDF	TF*IDF
calcinosis	11	1.08	1	1.41	1.53
dca	9	1.00	1	1.41	1.41
block	7	0.90	2	1.11	1.01
valve	6	0.85	2	1.11	0.94
pain	7	0.90	4	0.81	0.73

表 5: familyでヒットした文書の索引語

単語	f_{ij}	TF	n_i	IDF	TF*IDF
ikigai	15	1.20	2	1.11	1.34
caregivers	8	0.95	2	1.11	1.06
diagnosis	11	1.08	3	0.94	1.01
cancer	11	1.08	5	0.72	0.77
life	8	0.95	5	0.72	0.68

4.1 医師間の距離

図4に示したように，TF*IDF法によって求められた索引語は，医師が登録したキーワードでヒットした論文のAbstractから取得したため，医師を表す一種のユーザプロファイルと考えられる．そのため，医師に付属している索引語のTF*IDFの値を比べることで，医師間の距離を求めることが出来る．

距離がわかったら，医師はこの情報を基にして，以下の二つの考え方から，他の医師の情報を利用し，自分の研究戦略を決定する．

1. 医師Aが，自分と近い研究を行っている医師Bを見つけ，医師Bの情報を見る．そこ

には，医師Bが，過去にどんな論文を投稿したのか，現在どのような分野を取り扱っているのかがわかる情報があるとする．これを見た医師Aは，自分の研究の位置づけや，今後進む方向といった戦略を立てることができる．

2. 自分と似ている研究を行っている医師を見つけるということは，研究ライバルを見つけることにもなる．ゲノム，創薬などの分野は，特許を一日でも早く出そうと競争をしている．そのような状況の中で，ライバルの研究グループがわかることの意味は大きい．そのグループが過去にどんな論文を投稿しているのか，また，投稿された論文を即座に知ることができると，これも今後の研究の戦略を立てる上で重要な情報になる．

4.2 理論の実証

実際のデータを用いて，医師間の距離を計ってみる．

小児歯科系の医師Aがキーワードとして用いそうなhealth，child，familyを基としてPubMedを検索し，ヒットしたAbstractから索引語を求めた．これらの索引語は，小児歯科系の医師を表すユーザプロファイルとなる．

同様にして，工学系の医師Bがキーワードとして用いそうなonline，database，webから索引語を求め，薬学系の医師Cがキーワードとして用いそうなdrug，apothecary(薬剤師)，pharmacology(薬物学)から索引語を求めた．

医師AとBはcareという索引語が一致した．医師Aのcareの TF*IDF の値は0.63になり，医師Bのcareの TF*IDF の値は1.05だったので，二つの積の0.6615が医師AとBの近さになる．したがって，距離はこの値の逆数の1.51になる．

次に，医師AとCはcancerという索引語が一致した．ここで，医師Cの索引語の中にはcancerが2個(共に TF*IDF の値は0.5)あったので，医師Aのcancerの TF*IDF の値0.77を二つに掛けて，足した0.77が医師AとCの近さになる．したがって距離はこの値の逆数の1.29になる．

この結果より，医師A(小児歯科系)から見て，医師B(工学系)より医師C(薬学系)の方が自分に近いとわかった．

5 結論

本稿では，多忙な医師を支援するために，自動論文検索支援，関連論文検索支援，研究戦略支援が行えるPubMed Check Systemについて述べた．今後は，求められた医師間の距離を使って，研究グループ分けを行い，医師同士が協調研究を行えるシステムを構築したいと考えている．

参考文献

[1] PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[2] 「PubMed活用マニュアル」編著 縣 俊彦，南江堂，2000

[3] 「情報検索アルゴリズム」北 研二，津田 和彦，獅々堀 正幹，共立出版，2002

[4] Eric Brill's.: Brill's Tagger, http://www.cs.jhu.edu/ brill/