# PubMed からの発展
## －CSLS Search による PubMed 情報の活用－

柳元　伸太郎*

PubMed は今日では生命科学分野の文献情報のデータベースとしてなくてはならないものとなっている。PubMed のデータベースとしてのインターフェースは年々進歩を遂げており，使い勝手もよくなってきている。われわれは，検索結果の活用の観点から，現行の PubMed のインターフェースから一歩踏み込んで，PubMed で得られる検索結果のクラスタリングを行うシステム，PubMed での検索作業をパーソナライズする機能を実装した CSLS Search を公開している。クラスタリングを活用することで，通常のキーワードによる検索結果からは見えにくい文献間のつながりや新たなキーワードが浮かび上がってくる。効率的な情報活用が期待される。

**キーワード**：PubMed，クラスタリング，文献検索，生命科学，データマイニング

## 1. はじめに

現代の科学の発展は，全て一つの研究や一つの成果で達成されているわけではなく，それまでの一つ一つの研究成果の積み重ねの結果である。今日では研究の裾野が広がり，あるテーマに取り組んでいる研究者を一つのコミュニティーでくくることは困難になってきた。同じ研究をしている研究者同士でもお互いに全く知らないことが多い。また，科学の方法の確立は，その成果公開の手段としての論文発表を促し，結果として，膨大な数の論文が日々生み出される状況となった。

今日の医学・生命科学分野における文献データベースとして，PubMed はなくてはならない存在となっている。PubMed のデータの中心は MEDLINE データベースであり，これには 1940 年代からの文献情報 1,600 万件超が含まれていて，現在も成長を続けている。PubMed には，この MEDLINE のデータを含め，2,000 万件近い文献情報が収録されている。情報が限られていた時代にはいかに多くの情報を集めるかが重要であったが，いまや，大量の情報に本当に必要な情報が埋もれてしまっている状況になっている。

例えば，「がん」を意味する「cancer」をキーワードに PubMed で文献検索を実行すると，240 万件近い文献がヒットする。特定のがんにおける，特定の分子の働きに絞って検索したければ，「colon cancer」や「p53」などをキーワードにすればよい。このキーワードでは 2,400 件あまりがヒットする。

従来はこのような形で自らの背景の知識に基づいてデータベースを利用する方法が主流であったし，また今でも多くの研究者がこのような形で情報を得ていると考えられる。しかし，近年，生命科学分野でも盛んにデータマイニングの手法が用いられるようになり，こうした考えを文献情報の利用にも持ち込むというのは自然な流れと考えられる。

われわれは，PubMed をベースにしてデータマイニングとしてのクラスタリング機能と，個別ユーザーのためのパーソナライズ機能を実装したシステム，CSLS（Center for Structuring Life Sciences）Search を構築し，平成 19 年 2 月から公開している（注 2）。本稿では，PubMed の応用例としての CSLS Search の機能，利用法を紹介して，将来の方向性を考察してきたい。

## 2. CSLS Search の概要

CSLS Search は東京大学総合文化研究科生命科学構造化センターが生命科学教育，研究支援のためのプロジェクトの一環として平成 19 年に運用を開始したシステムである。基本的には PubMed をそのまま利用して，得られた検索結果を CSLS Search のシステムで加工して表示するというものである。このため，PubMed に含まれる情報を漏らすことなく，新たな付加価値としてのクラスタリングやパーソナライズ機能を利用できるという点で，ユーザーにはメリットが大きい。

### 2.1 CSLS Search による文献検索

具体的に CSLS Search で検索を行いながらその特徴を見ていくことにする。まず，ポータル（図 1）にアクセスすると検索窓が一つあるだけである。ここにキーワードを何か入力する。この検索窓に入力した文字列はそのまま PubMed に投げられるので，入力するのは単なるキーワードに限らず，PubMed で使用可能なタグ等を用いた検索のクエリを入力することができる。（例：「watson[1au] AND 1900[PDAT]: 2000[PDAT])」など）。当然ながら日本語での検索はできない。検索窓の右側に「100 件まで」というプルダウンメニューが出ているが，これは後に詳しく説明する「クラスタリング」を行う対象とする検索結果の件数（論文の本数）である。件数を多く指定すると，検索結果の

＊やなぎもと　しんたろう　東京大学　保健・健康推進本部
〒113-0033　東京都文京区本郷 7-3-1
Tel. 03-5841-2575
（原稿受領　2010.4.27）

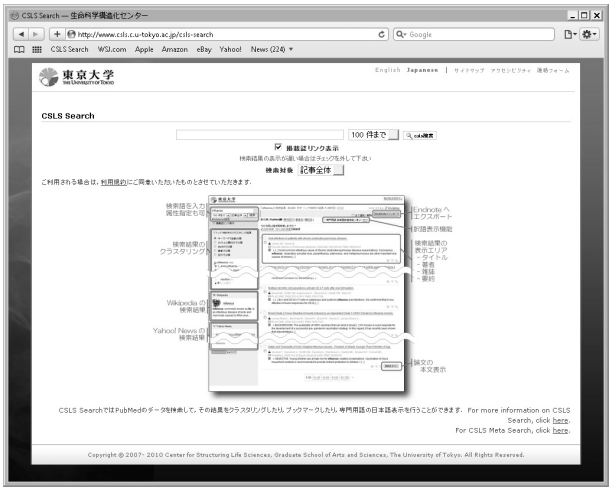


**図 1　CSLS Search のポータル**

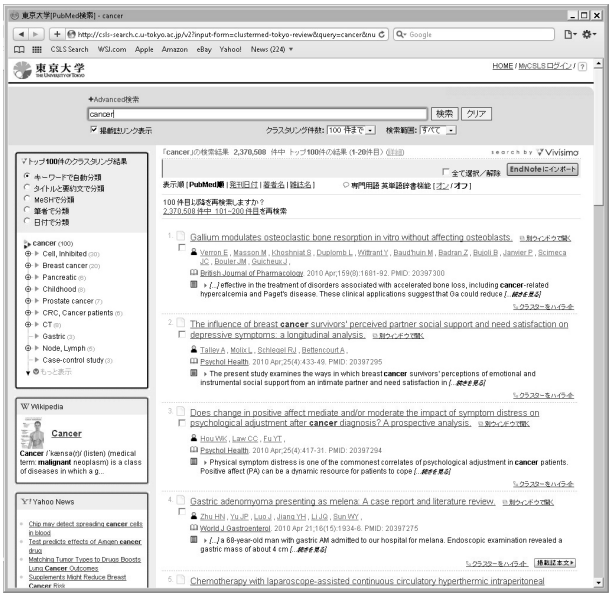


**図 2　CSLS Search の検索結果画面**

処理に時間がかかるため，表示に若干時間がかかる。(500 件で 15 秒程度)。

「csls 検索」ボタンを押して検索を実行すると，結果の画面が表示される（図 2)。

画面の構成としては，上部には検索窓があり，ポータルと同じインターフェースが提供されている。メイン（画面右側) の検索結果の一覧は PubMed での検索結果と同じ論文が同じ順序で表示されている。

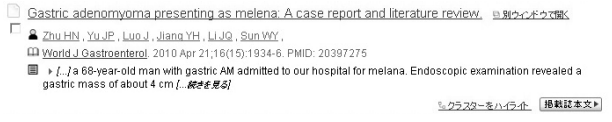


**図 3　検索結果**

タイトルをクリックすると該当する論文の PubMed でのページにジャンプする。アブストラクトは一部分しか表示されていないが，クリックすると展開してアブストラクトの全文が表示される。「掲載誌本文」ボタンは，その論文が掲載されている雑誌のサイトにある，その論文の全文へのリンクとなっている。ユーザーの出版社との契約状況によって，直接本文全文が表示されることもあるが，ログイン情報の入力を求められたり，論文購入のための表示になったりする場合がある。

メイン画面左側には，クラスタリングの結果と，キーワードに対応した Wikipedia と Yahoo! news での検索の結果が表示されている。クラスタリングについては後ほど解説する。

基本的な検索画面のほかに，検索窓の上に「Advanced 検索」のリンクがある。ここをクリックすると次のようなウインドウが開く（図 4)。


**図 4　Advanced 検索**

これは，ユーザーが PubMed の検索用のタグや構文の文法を覚えていなくても，ある程度複雑な検索ができるようにするためのものである。プルダウンメニューで項目を選びながら検索条件を加えていく。

もう一つ，付加機能として説明しておきたいのは，専門用語の日本語表示機能である。これは，意味を知りたい英単語の上にマウスのカーソルを持って行くと，対応した日本語が表示されるというものである（図 5)。日本語に変換するための辞書はライフサイエンス辞書プロジェクト（注 3）から提供を受けており，生命科学分野の専門用語について，かなりあたらしいものまで的確な日本語が表示されるようになっている。この機能はユーザーがいつでもオン／オフ可能で，必要のないユーザーは機能をオフにすることでサイトの不要な動作を軽減し，表示を速くすることができる。多くの研究者は少なくとも自分の専門領域の用語


**図 5　専門用語英単語辞書機能**

に関しては，英語で表記されていても不自由はないと考えられるが，クラスタリングの結果を活用して，少し通常触れないような領域の論文の内容を見たりする場合には有用であることが期待される。また，これから研究生活に入ろうというユーザーは，PubMed での文献検索の大きな壁の一つが言葉の問題である。これを少しでも解消することも，データベースを有効に活用して情報収集をするという目的の一助になると考えている。

このほか，本稿での詳細な説明は省略するが，代表的な文献データ整理ソフトウェアの一つである EndNote への検索結果のエクスポート機能，検索キーワードのスペルミスが疑われるときにほかの候補を指摘する機能など，ユーザーの要望を受けて追加してきた機能もある。

ここまで，CSLS Search の基本的な使い方を解説してきた。PubMed と同じ検索結果が得られるという点では，PubMed と大きな違いはない。しかし，検索結果にたどり着くまでのプロセスを，少しでも容易にしたこと，また結果の利用に関しても，論文全文へのアクセスの改善や，専門用語の日本語表示など，ユーザーがなるべく容易に検索結果を利用できるようにしたことで，PubMed に比較して利便性の向上を図った。ユーザーが増えるに従い，様々な要望のよされるようになっており，今後もユーザーサイドの視点からシステムの更新を進めていく必要がある。

### 2.2 MyCSLS－CSLS Search のパーソナライゼーション

CSLS Search はユーザー登録（無料）をしてログインすることで MyCSLS というサービスを利用できるようになる。これは，一言で言うと CSLS Search をパーソナライズする機能である。

主な機能としては，検索履歴の保存，よく実行する検索条件の保存，検索結果の文献の選択保存，保存する文献へのラベリングとメモ機能，などである。インターフェースは，少なくとも運営者側としては，自明だと考えているので，その詳細については自らがふれて確かめていただきたい（図 6）。


図 6　MyCSLS

## 3. クラスタリングから見える可能性

### 3.1 クラスタリングの仕組み

クラスタリングとは，データ解析の手法の一つで，多数のデータを何らかの特徴に従って分類するものである。生命科学の分野でも，新たな研究手法の導入やコンピュータの発達によって大量のデータが生み出されるようになり，その解析でもクラスタリングが広く行われるようになっている。CSLS Search では，PubMed から得られた文献情報をクラスタリングして，その結果をユーザーに提供している。PubMed から得られる情報は基本的にはテキストデータであり，CSLS Search で行っているクラスタリングはテキストマイニングの一種と言える。

CSLS Search のクラスタリング機能は，アメリカの Vivisimo 社（国内総代理店グループネット株式会社）の Vivisimo Velocity というシステムによって実現されている。CSLS Search での情報の流れは，次のようになっている。CSLS Search に入力された検索のためのクエリを PubMed に送信すると検索結果のデータが返ってくる。CSLS Search では，これをメイン画面右側で表示するために加工するとともに，データを Vivisimo Velocity に渡して解析（クラスタリング）する。クラスタリングの結果は CSLS Search の検索結果画面の左側に表示される（図 7）。

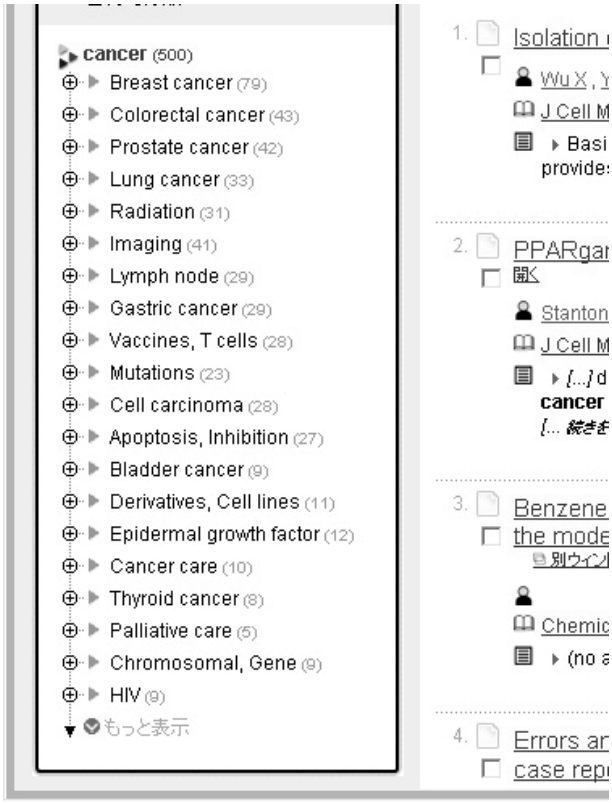


図 7　クラスタリングの結果

「cancer」をキーワードに CSLS Search で検索を行い，検索結果のうち，先頭の 500 件のデータをクラスタリングした結果が表示されている。

クラスタリングによってクラスターが生成されるロジックの詳細については Vivisimo 社は公開していないが，概略は次のようになっている。まず，検索結果などのテキストデータは単語やフレーズに分解される。この段階では全ての単語やフレーズはクラスターのキーワードの候補である。個々に分解された単語やフレーズの中から，クラスタリングには採用されない，ストップワードと呼ばれる単語はのぞかれる。ストップワードは予め登録されており，例えば，どのような検索結果にもほとんど常に現れるような重要度の低い単語が含まれている。

続いて，分解された個々の単語／フレーズが出現頻度，検索キーワードへの依存度などによってスコアリングされる。基本的には，出現位置が検索キーワードに近い単語ほど検索キーワードとの関連においては重要度が高い傾向にあると評価される。また，タイトルに含まれる単語の方が重要度が高いと判断されるなど，単語の使われ方によっても異なる重み付けがされている。一方，ストップワードに含まれていなくても，あまりにも出現頻度が多いものについては，かえって価値がないと判断し，クラスタリングの単語としては除外される。

このようにして，各単語やフレーズの重要度が数値化された結果，スコアの高いものはクラスタリングのキーワードとしてピックアップされる。最初の検索結果がこのキーワードに基づいてまとめられることで，一つのクラスターができる。CSLS Search ではクラスタリングのキーワードは重要度順に，メイン画面左側最上段に表示される。

クラスター名（キーワード）をクリックすることで，最初の検索結果のうち，そのクラスターに含まれる文献がメイン画面右に一覧として表示される。また，画面右最上段の「もっと詳しく検索しませんか？」をクリックすることで，クラスター名をキーワードとして PubMed 全体を再検索することもできる。また，キーワードによってはクラスターの中でサブクラスターが生成されることもある。これも，クラスタリングの結果の中に表示されている。

このように，CSLS Search ではクラスタリングは自動的に行われており，キーワードを指定する必要はない（指定できない）。しかし，CSLS Search のクラスタリング結果の表示を見ると明らかなように（図 8），CSLS Search ではクラスタリング対象を指定することができる。これによって，例えば，論文の著者でクラスタリングすることも可能になる。

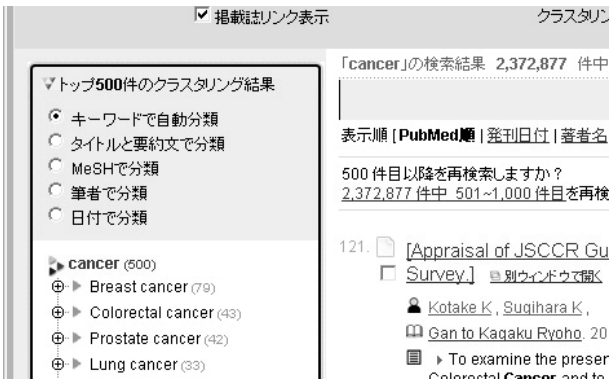


**図 8**　クラスタリング対象の指定

### 3.2　クラスタリングの利点

ここまで解説してきたように，クラスタリングによってクラスターが形成される際に，もとのキーワードに関連するキーワードが機械的にピックアップされる。当然ながらこのクラスターのキーワードはもとの検索のキーワードと深い関連が明らかなことが多い。例えば，「cancer」をキーワードに検索すると，クラスターのキーワードとして「breast cancer」がピックアップされてくる。しかし，「機械的」なクラスタリングは，時に，検索者が意識しないキーワードをピックアップしてくる場合がある。これはもちろん，偶然の結果で意味がないこともあり得るが，ある程度以上の量のデータの解析から求められたキーワードは多くの場合，何らかの関連があるものである。こうした思いもかけないキーワードは，情報検索の範囲を広げ，これまで見えていなかった関係を見いだす上できわめて重要な役割を果たす。

しかし，より単純に，この機能は検索スキルが低いユーザーや，関心のあるキーワードの分野での知識や経験が十分ではないユーザーにとって大変有用である。例えば，さきの「cancer」をキーワードに検索した際の例では，クラスターに「breast cancer」や「lung cancer」があがってきた。これは，がんの研究をしている研究者にとっては何も目新しいものではないが，初めてがんの研究に取りかかろうという研究者にとっては，クラスターの上位に上がってくる乳がんや肺がんがそのフィールドで重要である（あろう）ということを示唆してくれるので，当初の検索結果の絞り込みや，新たな検索キーワードとしてさらに情報収集を進めること可能にしてくれる。

### 3.3　クラスタリングの応用例

CSLS Search のクラスタリングのロジックはブラックボックスとなっているので，クラスタリングの結果をそのまま研究成果として扱うことは難しいかもしれないが，クラスタリングで，キーワードが浮かび上がる，という例を示してみたい。

表 1 は，「influenza x: x+1[dp]」（x は実際には西暦の年）というキーワードで CSLS Search による検索を実行した結果生成されたクラスターをまとめたものである（上位 500 件のクラスタリング）。当時インフルエンザを研究していた研究者には当然のキーワードかもしれないが，それを知るよしもない現代のわれわれでも，「influenza」というキーワードに深い関連のある項目が，おおまかにではあるが，年代毎に知ることができる。

表で，1970 年になると，「Hong Kong influenza」がクラスターとして上がってくるが，これは，その前々年から前年に流行したいわゆる香港かぜの影響と考えられる。その後，1980 年代からはインフルエンザのワクチンによる予

表 1 「influenza」をキーワードにした CSLS Search の年代別クラスター結果一覧

| x:x+1 | 1960~61 | 1970~71 | 1980~81 | 1990~91 | 2000~01 | 2007~08 | 2009~2010 |
|---|---|---|---|---|---|---|---|
| 1 | immunology | therapeutic use, drug therapy | influenza vaccines | influenza vaccines | influenza immunization | avian influenza | Avian influenza |
| 2 | epidemiology | epidemic | complications, etiology | T cell | T cells | pandemic, epidemic | Pandemic H1N1 |
| 3 | epidemic | cells, virus | T-lymphocytes | Respiratory, Children | epidemic, influenza pandemic | oseltamivir, neuraminidase | oseltamivir, treatment |
| 4 | inhibitors, hemagglutination | virus infection | RNA | Class I | physicians, influenza vaccine | RSV, respiratory syncytial virus | acute respiratory |
| 5 | Asian influenza | pneumonia, complication | Hong Kong | Fusion, pH | zanamivir, sialic acids | prevention, and treatment | CD, T cells |
| 6 | therapy | influenza A virus, Turkeys | Avian | RNA, Genetics | CTL cytotoxic T lymphocytes | healthcare workers | 2009 pandemic influenza A |
| 7 | cells, infection | Hong Kong influenza | influenza virus hemagglutinin | influenza activity | HIV | MDCK cells | Healthcare workers |
| 8 | cultures, tissue | Clinical, Picture | antibody response | HIV | influenza surveillance | CD,T cell | pregnant women |

防が盛んに進められていたこと，免疫学の研究の発展が著しかったことが，クラスターからは見て取れる。抗インフルエンザ薬が臨床で一般に使われるようになり，2000 年，2007 年に抗ウイルス薬の zanamivir（リレンザ®），oseltamivir（タミフル®）がクラスターにも登場するようになった。2009 年はまだ記憶に新しいところであるが，新型インフルエンザが出現し，クラスターからもそれが読み取れる（表記のぶれが同じことについて複数のクラスターを生成させているようである）。

これは，過去のデータを遡及的に，ある意味，答えを先に見て，眺めているわけであるが，クラスタリングの結果には，このように，研究のトレンドを如実に反映する性質があり，そこから将来への研究の方向性を見定めるための材料が得られる可能性もある。

## 4. PubMed を中心としたデータベース利用の広がり

### 4.1 PubMed に見られる検索機能の充実

PubMed が 1997 年に公開された当初のユーザーインターフェースは非常にシンプルであった。データベースとしての基本的な検索機能は備えていたが，複雑な検索を実行するにはある程度以上の PubMed 利用のスキルが必要であった。

現在提供されている PubMed の機能の詳細は他稿に譲るが，今日の PubMed は情報検索のしやすさが格段に向上している。検索窓にキーワードを入れると，他のキーワードの候補が自動的にリストアップされたり，比較的参照頻度が高いと思われる review や無料で本文にアクセスできる論文への絞り込みのリンクが予め生成されたりするなど，新たな機能が追加されている。また，個別の論文情報にアクセスすると関連論文が表示されるという機能も便利である。

このほか「MyNCBI」というサービスを利用すると，検索文を保存したり，特定のキーワードに該当する論文が新たに登録されると電子メールを受け取ったりすることができる機能も提供されている。

### 4.2 東京大学附属図書館の文献検索サービス

東京大学附属図書館では東京大学学術論文横断検索というサービスを提供している。パブリックに提供されているものではないので簡単に紹介として触れることにする。基本的には，PubMed を含む，複数のデータベースを横断的に検索できる，というもので自然科学に限らず，人文，社会科学，海外，国内のデータベースを 70 以上対象とすることができる。クラスタリングの機能が搭載されているが，汎用性を重視したためか，クラスターの範囲が非常に大きなくくりになっている傾向がある。

## 5. おわりに

医学生命科学の情報をデータベースにまとめる試みは古くから行われてきた。Index Medicus はその代表的な例であろう。書籍の形態をとっていたデータベースは情報の取り出し方そのものにも様々な問題があり，今日，デジタル化されたデータベースでは考えられないような苦労があった。また，情報・通信の発達も今日とは比べものにならず，いかに網羅的に，いかに多くの情報を収集するか，がデータベースの整備の上では重要な課題であった。

しかし，今日のように，情報量が非常に多くなり，また蓄積していく速度も，個々の研究者が対応できる範囲を大きく超えてしまった状況において，データベースのユー

ザーである研究者は，大量にあるデータをどのように有効活用していくのか，大量のデータから，本当に有用なものを如何に効率的に見いだすのか，という課題に直面している。今後，PubMed をはじめとする種々のデータベースは，単なる検索機能だけではなく，情報の発見や整理を支援するような機能を備えながら発展していくものと考えられる。

### 注

1) CSLS Search は Microsoft Internet Explorer での動作確認を行っている。図中，Safari のスクリーンショットが使われているが，Safari では一部の機能が正常に動作しないことが判明している。
2) http://www.csls.c.u-tokyo.ac.jp/csls-search，東京大学総合文化研究科生命科学構造化センターサイト内。なお，生命科学構造化センターは平成 22 年 3 月でプロジェクトが終了し，CSLS Search の管理・運営は東京大学教養学部附属教養教育高度化機構・生命科学高度化部門が引き継いでいる。
3) http://lsd.pharm.kyoto-u.ac.jp/ja/index.html

### 参 考 文 献


1) PubMed
http://www.ncbi.nlm.nih.gov/pubmed
[accessed 2010-04-19].
2) MEDLINE fact
http://www.nlm.nih.gov/pubs/factsheets/medline.html
[accessed 2010-04-19].
3) Vivisimo
http://vivisimo.com/ [accessed 2010-04-19].
4) Groupnet
http://www.groupnet.co.jp/products/velocity/index.html
[accessed 2010-04-25].
5) Wikipedia
http://en.wikipedia.org/wiki/Main_Page
[accessed 2010-04-25].
6) Yahoo! news.
http://news.yahoo.com/
[accessed 2010-04-25].
7) WebLSD
http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html
[accessed 2010-04-25].



**Special feature**: Advanced PubMed. Advanced application of PubMed -CSLS Search for better use of information in PubMed-.


Shintaro YANAGIMOTO (University of Tokyo, Division for Health Service Promotion, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 JAPAN)



**Abstract**: PubMed has become an indispensable database in the field of life sciences. Its user interface is improving steadily, and it is becoming increasingly user friendly. Beyond its own function as a database, we have developed a system, named CSLS Search, that enables users to process the search results by “clustering” according to keywords, which elucidates unseen relationship between articles in a search result. Our system also allows users to personalize their searches by storing results or adding notes on an article. We are expecting more effective use of information.


**Keywords**: PubMed / clustering / journal search / life sciences / data mining