E-005

生命医学文献の新たな文献探索システムの開発支援

Support for Development of a New Text Mining in Biomedical Literature

田中 一博† 大和田 勇人‡
Kazuhiro Tanaka Hayato Ohwada

1． はじめに

生命科学の分野において、研究文献は日々爆発的に増加し、膨大な数となっている。そのため、研究者が文献を読んで理解しながら対処できる限界を超えているなどの問題が挙げられている。近年、大量の文献群から効率よく関連する情報を収集する技術が切望されており、特に PubMed を対象とした文献からの情報抽出の研究が生命科学の研究分野において注目されている。PubMed とは、MEDLINE® といわれる世界で最もよく使用される生物医学系データベースを検索するための Web ツールである。生物医学系の研究分野は PubMed を利用することにより探索できる。例えばインフルエンザに関する研究文献や論文を調べたいとする。その際、テキストボックスに influenza と記入し送信すると influenza のワードが載っている研究文献や論文は探索される。また、influenza の同義語や関連語も探索したいとする。そこで、MeSH といわれる生命科学用語集を利用することにより、他の influenza に関する研究文献や論文が探索できる。

一般的な文献探索システムはほとんど全文検索になっており、検索キーワードを含む文献のみを探索する型になっている。そこで本研究では、PubMed を対象としたシソーラスを利用した新たな文献探索システムの提案を行い、検索キーワードを含まないで検索キーワードに関連する文献探索の可能性を示す。提案の目的として、研究者が知りたいキーワードの関連文献を探索するための手助けが挙げられ、また、関連語抽出やコーパス作成への発展に繋がると考えている。

2． シソーラス

シソーラス（Thesaurus）の定義は、単語の上位/下位関係、部分/全体関係、同義関係、類義関係などによって単語を分類した体系である[1]。シソーラスは自然言語処理の分野でも重要な位置にあり、電子データベース化されているものもある。データベース化されたシソーラスは木構造、または表形式で成り立っているものが多く、全文検索システムなどで利用される「あいまい検索」もシソーラスを利用して行われている。代表的な生命科学の分野のシソーラス辞書として、MeSH が挙げられる。

3． 提案手法

本研究ではシソーラス辞書を用いた文献探索手法について提案する。そこで先ほど紹介した MeSH を利用する。しかし、MeSH だけで PubMed からあるキーワードに関する研究文献や論文を全て探索することができているのだろうかという疑問がある。関連研究として、金子ら[2]は「文献情報の解析に基づく対訳シソーラスの評価」と題した研究を行っている。この研究では、テキストマイニング等に応用できるシソーラスへの発展を目標に、MeSH の評価を行った。その結果、MeSH だけでは PubMed から文献を探索しきれていないことが示唆された。そこで、他の生命科学用語集の代表的なものとして LSD と MeSH の比較を行い、双方を組み込んだ文献探索手法について提案する。また、文献探索手法にあたって、Apache Tomcat(通称 Tomcat)と言われる Java サーブレットや JSP を実行するためのサーブレットコンテナを用いる。

本研究では以下の通りのシソーラス辞書を構築していく。

① ユーザの知りたいシソーラス語のキーワードの記入欄としてテキストボックスを使用する。送信ボタンをクリックすることにより、そのシソーラス語がピックアップされている画面に遷移する。
② ピックアップされているシソーラス語の画面にとんだら、その中でユーザの知りたい研究文献や論文に関するワードを選んで、クリックする。
③ そのワードが入った研究文献や論文が探索される。

一般的な文献探索は、ほとんど全文検索となっており、ユーザの知りたい研究文献や論文のキーワードを記入することにより、そのキーワードが入った研究文献や論文だけを取り出すような探索である。本研究では、上記の②のとおり、間にシソーラス辞書を用いることにより、ユーザの知りたいキーワードがなくて、ユーザの知りたいキーワードに関する研究文献や論文を探索できる。

†東京理科大学院理工学研究科経営工学専攻
大和田研究室 j7410625@ed.noda.tus.ac.jp
‡東京理科大学理工学部経営工学科

表１．MeSH と LSD の比較結果の一部

	Mesh	LSD	共通	Meshのみ	LSDのみ
influenza	61	33	21	40	12
atopic	24	4	3	21	1
rhinitis	7	7	7	0	0
asthma	6	8	3	3	5
antidote	10	1	1	9	0
antipyretic	2	1	1	1	0
antineoplastic agent	92	6	6	86	0
respiratory distress	22	3	2	20	1
hypnotic	13	1	1	12	0
fomentation	0	0	0	0	0
esophagitis	5	2	2	3	0
myocardial	26	23	20	6	3
heart failure	3	3	3	0	0
urticaria	3	3	3	0	0
tranquilizer	16	3	3	13	0
enteritis	10	16	10	0	6
gout	3	8	2	1	6
epilepsy	27	22	21	6	1
diabetes mellitus	7	7	7	0	0
cerebral infarction	3	2	2	1	0
cerebral hemorrhage	13	3	3	10	0
pneumonia	29	42	22	7	20
lung cancer	8	3	1	7	2
obesity	8	4	4	4	0
peritonitis	6	5	5	1	0

4． 結果

4.1 比較結果

本研究では、よく使われる生命科学用語 100 語[3]を用いて調査を行った。各語のシソーラス語の数を MeSH と LSD で調査し、グラフにまとめた。MeSH と LSD の比較の調査結果は表 1 に示す。よく使われる生命科学用語 100 語の全ての結果は、MeSH では合計 951 語発見され、LSD では合計 587 語発見された。その中で MeSH と LSD で共通して発見されたシソーラス語は合計 413 語あった。そのことから、MeSH 独自のシソーラス語は合計 538 語あり、LSD 独自のシソーラス語は合計 174 語あった。

4.2 文献探索システムの結果

上記の提案の通り、シソーラス辞書を構築することにより、ユーザの知りたいキーワードがなくても、そのキーワードに関する研究文献を探索できるようになる。図 1 は MeSH と LSD を組み込んだシソーラス辞書の influenza のシソーラス語の出力例である。上記の②に述べたように、図 1 のようなシソーラス語の出力画面に遷移する。

4.3 考察

文献探索の際にシソーラス辞書を組み込ませることにより、文献探索を助けると考えられる。例として、influenza の関連した研究文献や論文を調べたいとき、先に influenza のシソーラス語を挙げることにより、今まで発見できなかった新たな研究文献や論文を発見することができると考えられる。また、本研究の提案は文献探索だけでなく、Web 検索にも応用できると考えられる。例として、google サイトのテキスト欄に「influenza」と入力する。そのときに、

図１．シソーラス語の出力画面の例

同時に「influenza wiki」などの influenza という語を含む近接語が検索の手助けとして出力される。しかし、influenza という語を含まないで influenza に関する語までは出力されない（「Thogotovirus」など）。しかし、シソーラス辞書を組み込ませることにより、influenza という語を含まない influenza の関連語も出力できるようになり、Web 検索においてさらなる検索の幅が広がると考えられる。

5． 結論

本研究では、PubMed を対象としたシソーラスを利用した新たな文献探索システムの提案を行い、検索キーワードを含まないで検索キーワードに関連する文献探索の可能性を示した。その際に MeSH と LSD の比較を行い、結果により、双方に独自のシソーラス語が発見されたため、PubMed から新たな研究文献や論文の探索が可能ということも示された。また、本研究では、MeSH と LSD を組み合わせた辞書を用いた文献探索手法について述べた。Tomcat とサーブレットを用いて、文献探索のためのシソーラス辞書の構築を行った。文献検索の間にシソーラス辞書を組み込ませることにより、ユーザの知りたいキーワードがなくて、ユーザの知りたいキーワードに関する研究文献や論文を探索することを助けると考えられる。

参考文献

[1] シソーラスとオントロジー：http://www.gengokk.co.jp/thebun.htm

[2] 金子周司、藤田信之：文献情報の解析に基づく対訳シソーラスの評価、医療情報学 2006

[3] TOP/NET 東海四県薬剤師会：http://topnet.gr.jp/index.html