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

医療文献からの類似項目発見についての考察

園田, 亮
九州大学大学院システム情報科学府

伊東, 栄典
九州大学情報基盤研究開発センター

高崎, 哲也
株式会社ワールドフュージョン

川原, 弘三
株式会社ワールドフュージョン

https://hdl.handle.net/2324/17045

出版情報：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. 電子情報通信学会
バージョン：
権利関係：

社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

医療文献からの類似項目発見についての考察

園田 亮† 伊東 栄典‡ 高崎 哲也* 川原 弘三*

†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
*株式会社ワールドフュージョン 〒103-0013 東京都中央区人形町 2-15-15
E-mail: †ryo.sonoda@i.kyushu-u.ac.jp, ‡itou@cc.kyushu-u.ac.jp, *{takasaki, kkawahara}@w-fusion.co.jp

あらまし 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．それに伴い，学術文献内に出現する項目の関係を用いて，新たな項目間の発見を目指す文献に基づく知識発見(Literature based mining)が行われている．ここでいう項目とは病名，薬名，化学物質名，遺伝子名などを指す．

我々も，文献に基づく知識発見について研究しており，特に病名・遺伝子と薬名・化学物質名との関係に着目している．新たな類似項目が発見された時には，分子シミュレーションを行い，類似関係を実証する．我々の調査対象は米国国立医学図書館が作成した医学，生命科学分野に特化した文献データベース MEDLINE である．本論文では，研究の全体像を述べると共に現在我々が保持しているデータ量について述べる．また，いくつかの検討している分析方法やアルゴリズムについて考察する．

キーワード 情報検索，文献検索，データマイニング，文献からの知識発見

A study of similar item discovery from medical literature

Ryo SONODA† Eisuke ITO‡ Tetsuya TAKASAKI* and Kozo KAWAHARA*

†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
‡World Fusion Co., Ltd. 2-15-15 Nihonbashi Ningyo-cho, Tokyo, 103-0013 Japan
E-mail: †ryo.sonoda@i.kyushu-u.ac.jp, ‡itou@cc.kyushu-u.ac.jp, *{takasaki, kkawahara}@w-fusion.co.jp

Abstract Recently, the scientific papers that describe the result of research in the medical field have been accumulated. With increasing these papers, Literature based mining have been tried to discover new knowledge by using the relationship of items in the source literature. The item here indicates the name of a disease, the medicine, the chemical, and the gene, etc.

We also research the knowledge discovery based on the literature. We especially pay attention to the relation between disease and gene, and between medicine and chemical. When we discover new relationship, we confirm it using molecular dynamics computer simulation. Our target database is the MEDLINE which is the literature database specializes in the medical and biological science, and is made by the United States National Library of Medicine. In this paper, we describe over view of our research plan and the amount of data, and consider our method of analysis.

Keyword Information retrieval, Document search, Data mining, Literature based mining,

1. はじめに

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 1996 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．また，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)[11]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．

医療分野に限らず，特定の分野でデータベース化が進んでいる．例えば，分子構造については Pubchem[8]というデータベースが構築されている．塩基に関するデータベースは，アメリカの GenBank[12]，ヨーロッパの EMBL[13]，日本の DDBJ[14]がある．これらは国際協力を行っており，３つのデータベース間でデータ交換を行っている．タンパク質データベースについは，

SwissProt, TrEMBL, PIR という３つのデータベースを統合した UniProt と，タンパク質の立体構造データベースである PDB が有名である．これら以外にも，仮想スクリーニングのためのデータベース ZINC や，薬物名称のデータベースである DrugBank などがある．従来これらのデータベースは，同一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．

近年，学術文献内やデータベースに出現する項目の関係を用いて，新たな項目間の発見を目指す文献に基づく知識発見(Literature based mining)が行われている．ここでいう項目とは病名，薬名，化学物質名，遺伝子名などを指す．

この手法は製薬分野においては大きな影響を与えると期待される．製薬分野で用いる化学物質の数や，それらの組み合わせパターンは膨大である．新薬発見のために物質の組み合わせ候補を提案する手法の１つとして文献からの知識発見は有用であると考えられる．製薬以外の医学や生物学分野においても，病気の遺伝子や化合物との関連性を知る一助になり，また斬新な視点からの新知識獲得につながると期待できる．

我々も，文献に基づく知識発見について研究しており，特に病名・遺伝子と薬名・化学物質名との関係に着目している．対象とする文献は先行研究と同じ MEDLINE を用いる．先行研究との違いは計算機でのシミュレーションとの連携である．発見した関連項目（関連物質）を，計算機を用いて分子動力学シミュレーションにより関連性を検証する．本論文では，研究の全体像を述べると共に，現在我々が保持しているデータについて述べる．また，検討している関連項目発見アルゴリズムが，保持データに対して現実的に適用可能かどうかについて考察する．

本論文の構成を述べる．２節では関連研究について述べる．３節では我々が分析に用いるデータについて詳細を説明する．４節では，収集したデータの処理方法と適用を想定しているアルゴリズムの計算量について述べる．最後に５節で本論文のまとめを述べる．

2. 関連研究

文献に基づく知識発見の関連研究として，２つの類似項目発見について述べる．

2.1. Swanson の手法

Swanson[1]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．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．しかしながら,現在の Arrowsmith の算出する B の候補リスト (B-list) には単語が多数含まれるため，科学的に興味深い関係を導き出すのは困難である．

そこで文献[2]で，MeSH を用いて B-list の単語を順位付けし，B 候補単語の品質を保ちつつ候補単語数を制限する方法を提案している．MeSH (MEDLINE Subject Headings)とは，MEDLINE 収録論文を特徴付けるカテゴリ分類語彙であり，各収録論文には MeSH 語彙が 12 個程度付与されている．提案している B 候補の単語を順位付ける評価尺度は以下の２つである．

(i) sh-wt (subject heading weight)

(ii) coh (cohesiveness)

上記二つの評価尺度により，文献間の関係の強さが算出でき，関係の強さに基づいて B 候補の単語を制限する方法を提案している．また，提案した評価尺度を検証するために，あらかじめ Swanson らが人手で調査していた migraine（偏頭痛）と magnesium（マグネシウム）の結果と，Swanson らの手法で算出した結果との比較を行った．Swanson らの手法は，人手による調査内容を比較的再現するものとなり，より効果的かつ計算機に適した知識発見の方法であると Swanson らは述べている．

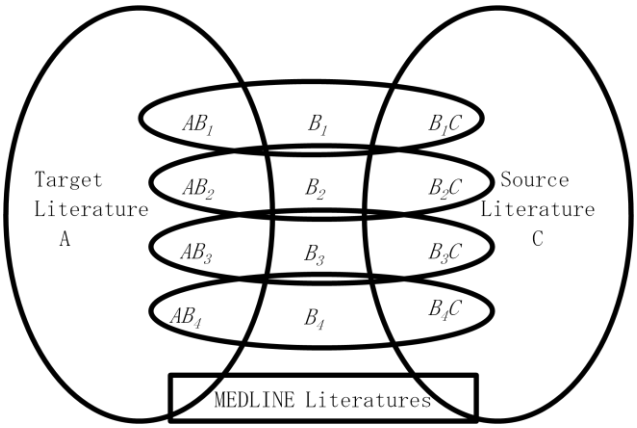

図 1. Swanson らの ABC モデル
(文献[1]を参考に作成)

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．

このアルゴリズムは表１のように(1)~(3)の情報を入

力する.

表 1:入力する情報

(1)	A	物質名
(2)	ST-B, ST-C	B,C を定める際に条件となる UMLS の意味クラス名
(3)	M	MeSH 用語候補数の制限

上記(2)の UMLS(Unified Medical Language System)とは米国国立図書館が作成したシソーラスである. ST-B,ST-C は UMLS 上の意味クラスを指定することになる. これによって, Aから導かれる, MeSH 用語を指定した意味クラスのみに制限する. 実験の例として B を定める際の条件として ST-B を"Gene or genome", "Enzyme", "Amino Acid, Peptide or Protein"３つの意味クラスと定めた. M は, B のリストを算出する際に出力する MeSH 用語候補数である.

具体的な実験例として, アジアで香辛料として使われ, さまざまな治療効用があるといわれている Curcuma longa（別名 turmeric）を物質名(A)と選定している. PubMed に入力する検索語は Turmeric または Curcumin または Curcuma である. これら三つは同じ物質を指す言葉の別名（エイリアス）である.

前述した導出アルゴリズムを用いて Curcuma が有効に作用すると考えられ, かついままで Curcuma との関係が見つかっていない病気を探した. 実験の結果, 網膜の病気(retinal diseases), クローン病(Crohn's disease), 脊髄に関係する病気(disorders related to the spinal cord.)に Curcuma が有効に働くと導出された. これらの病気は, Curcuma が強く関与していると考えられた TNF-α, MAPK, NF-κB, COX-2 のような遺伝子やサイトカイン, インターロイキンといったタンパク質との関係に基づいて導出された.

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. 基礎データ

この節では, 我々が用いる LSKB データベースについて述べる. 次に, 我々が収集した XML 形式の文献データと, 関連ファイルについて述べる.

3.1. LSKB データベース

ワールドフュージョン社では LSKB(Life Science Knowledge Bank)と名付けたデータベースを作成している. LSKB は, 遺伝子名やシンボル, キーワードなど, 遺伝子に関する情報をデータベース化したシノニム辞書と, 相同性検索により同定された遺伝子, 関連するタンパク質の機能辞書を搭載している. さらにはシノニム辞書を利用して行った文献マイニングのデータを保持しており, これによって遺伝子と疾患, 化合物との関連性を検索表示が可能である. 表 2,3,4,5 に詳細を示す.

表 2:化学物質に関するマイニング情報

DBの名前		エントリー数
化学物質(薬)		
PubChem	全化合物エントリー数	10,300,474
	全物質エントリー数	17,665,282
	文献に利用されている化合物の数	92,086
	文献に利用されている物質の数	231,264
ZINC	エントリー総数	5,627,809
	このうちPubChemに登録されている数	3,813,892
FDA Product		10,602
FDA 化合物	エントリー総数	1,858
	このうちPubChemに登録されている数	1,855
MeSH	エントリー総数	169,174
	PubChemに登録されていないMeSH	122,132

表 3: 遺伝子に関するマイニング情報

DBの名前		エントリー数
遺伝子		
UniProt/SwissProt	エントリー総数(Human,Mouse,Rat)	35,727
	PDBありエントリー(Human,Mouse,Rat)	3,524
UniProt/TrEMBL	エントリー総数(Human,Mouse,Rat)	112,918
	PDBありエントリー(Human,Mouse,Rat)	100
PDB	全エントリー	47,137
UniProt/SwissProt	エントリー総数(全生物種)	267,354
	PDBありエントリー(全生物種)	10,526
UniProt/TrEMBL	エントリー総数(全生物種)	4,361,897
	PDBありエントリー(全生物種)	2,612
Gene	Human(9606)(Active)	38,549
	エイリアス(類義語)を区別しない場合	253,466
	Human(9606)(NCBIでReplace)	9,313
	エイリアス(類義語)を区別しない場合	12,553
	Human(9606)(NCBIで削除)	114,951
	エイリアス(類義語)を区別しない場合	116,567
	Mouse(10090,10092,57486)(Active)	62,194
	エイリアス(類義語)を区別しない場合	227,940
	Mouse(10090,10092,57486)(NCBIでReplace	26,921
	エイリアス(類義語)を区別しない場合	34,120
	Mouse(10090,10092,57486)(NCBIで削除)	76,823
	エイリアス(類義語)を区別しない場合	77,284
	Rat(10116)(Active)	38,063
	エイリアス(類義語)を区別しない場合	90,023
	Rat(10116)(NCBIでReplace)	1,707
	エイリアス(類義語)を区別しない場合	2,214
	Rat(10116)(NCBIで削除)	27,173
	エイリアス(類義語)を区別しない場合	27,196
GENE-GO_Product		16,831
GENE-Go_term		358,879

表 4:病気に関するマイニング情報

項目	数
ICD 疾患数	1,194
MeSH 疾患数	9,749
論文上の連携疾患数	6,528

LSKB では, 分子やタンパク質などのキーワードからの標準遺伝子シンボル検索や, ２項関係を用いた遺伝子と関連するタンパク質の一発表示, さらには MEDLINE データベースを利用した, 遺伝子・タンパク質・疾患および化合物の関連性検索が可能である.

表 5:関係に関するマイニング情報

関係の種類	数
遺伝子_遺伝子	78,015,926
遺伝子_疾患名	13,021,513
遺伝子_化学物質名	125,013,446
遺伝子_組織名	17,980,430
疾患名_疾患名	1,574,641
疾患名_化学物質名	12,349,883
疾患名_組織名	2,354,034
化学物質名_化学物質名	71,053,960
化学物質名_組織名	17,298,467
組織名_組織名	1,542,453

3.2. 解析対象データ

知識発見の解析対象とする文献データについて述べる．本研究では，LSKB において疾患名が「肺癌」(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検索により抽出した．

NCBI(National Center for Biotechnology Information)に Batch Entrez[15]という Web 検索ツールがある．PubMed の文献 ID リストを入力することにより， 文献の二次情報（題名，著者名，概要，その他）を XML 形式で取得することができる．このツールを利用して前述の条件に当てはまる文献データを入手した．Batch Entrez によって取得を試みた 全ての PubMed ID のリストの数は 23,467 件であった．

4. 類似項目発見についての考察

この節では，収集した文献データからの類似項目発見についての考察を行う．

4.1. データにおよび分析手法の考察

前節で述べたように，文献データは XML 形式をしている．そこで，文献毎に付属している XML 中で特定のタグで囲まれた属性情報を抜き出す．抽出結果をデータベースに入力しておき，後からの項目関係発見アルゴリズムの適用を簡易にする．

次に，新関係発見の基本方針を述べる．同一文献内だけの関係のみで新たな項目間の関係を発見することは困難であるため，複数文献からの新たな項目間関係を発見することが重要である．そのためには，異なる文献をキーワードや分野といった何らかの指標でで関連付ける必要がある．

現在，疾患名「肺癌」を含む文献を検索して解析対象と考えている．従って，疾患名(A)をスタート地点として，A と同文献内で共起する遺伝子名，MeSH 用語，疾患名，化学物質名等の中間語句(B)を経由して有効に働きそうな物質(C)を A が出現しない文献内で B との共起関係を用いて導出する。換言すれば，A と C との関係を複数文献間で調査する．(図 2)

そのためには，まず疾患名を含む文献内に出現する遺伝子等の名前を，類語辞書や別名辞書を参考にしつつ抽出する．さらに，Batch Entrez を用いて遺伝子等が出現する文献データを取得する．取得した文献データを分析することで，各中間語句となる物質の順位付を行う．この手順で,「肺癌」に作用する薬名と，それを導き出した原因となった遺伝子名や分野名を導き出せると考えている．

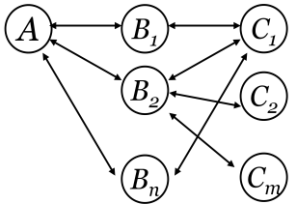

図 2. 共起関係項目の関連

4.2. 発見アルゴリズムの信憑性の考察

一方，発見アルゴリズムの信憑性も問題になる．アルゴリズムの信憑性を評価するために，既知の成果を利用する．もし，適用したアルゴリズムが導出する関係が今までに有効と知られている関係を多く含む(再現率が高い)ならば，そのアルゴリズムの信憑性が高いと考えられる．同時に，アルゴリズムは，未だ知られていない関係(¬X∩Y)も導出すると考えられる．図 3 に示すように，既知の項目間関係と，導出された項目間関係の割合で信憑性が評価可能と考えている．

$$\text{適合率}\quad precision = \frac{X \cap Y}{Y}$$

$$\text{再現率}\quad recall = \frac{X \cap Y}{X}$$

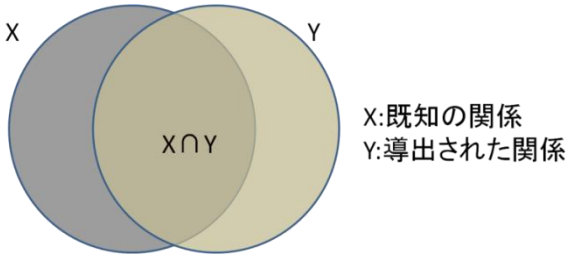

図 3. 発見アルゴリズムの信憑性評価

4.3. 候補の重み付けの考察

候補に重みを付ける方法の代表的な手法として TF-IDF 法がある.全てのドキュメント数を N，あるキーワード A の出現回数を TF_A あるキーワード A が出てきたドキュメントの数を DF_A (document frequency)としたとき，IDF_A (inverse document frequency)は以下

のように定義される

$$IDF_A = \log \frac{N}{DF_A}$$

ある語 A の重み($weight_A$)は

$$weight_A = TF_A * IDF_A = TF_A * \log \frac{N}{DF_A}$$

と書ける.TF-IDF 法を用いれば高い頻度で少数の文書に現れる特徴的なキーワードを発見することができる.

コサイン類似度を用いる場合. キーワード A が, 文献中に出現したか否かを 1,0 で表現する.

$$\vec{A} = (a_1, a_2, a_3 .. a_n) \text{ ただし } a_i = \{0,1\}$$

2 つのキーワード A, B のコサイン類似度は以下のように計算できる.

$$\cos(\vec{A}, \vec{B}) = \frac{Co(A,B)}{\sqrt{DF(A) * DF(B)}}$$

ただし Co(A,B)とは A と B が共起した回数である. コサイン類似度を用いればキーワード A,B のベクトル空間に基づいた類似性を計算可能である

２つのキーワードの共起関係に注目する. 共起関係を用いた場合偶然共起した場合と, ２つのキーワードが強固に関係している場合の区別をつける必要がある. まず, 共起回数で関係の強弱を判断する. 何度も共起する２つのキーワードは強い関係があるといえる. この場合, 少数の共起例からの新しい関係を発見する可能性を潰すかもしれない。しかし, 少数の共起関係を考慮に入れると, そこから関連のありそうな項目の候補数が膨大になるため, 本当に関係のある項目の見つけることが困難になる. 前述したように導出した関係のうち, 既知の関係（正解）の割合が高く, 一部に未知の関係を含む場合が新発見への信憑性が高いと考えれば, 少数の共起関係は無視してもよいだろう.

専門家が人手で文献に割り当てた MeSH 語彙は, 精度の高い情報といえるため, 同じ MeSH 用語が割り当てられている文献間の共起関係は重視する. MeSH を使えば,同一分野での文献における共起が分かるため, 有効である.

なお, 関連項目発見の初期段階では, 中間語句として遺伝子名, MeSH 用語, 疾患名, 化学物質名に限定する. これは, LSKB のデータを有効に使い正確な結果を出すためである.

最終的には,「肺癌」(A),中間物質名(B),薬名(C)をそれぞれノードとみなしたときに,各 C に対して,A からその C のすべてのパスを合計する. ただし 2 つ以上の中間語句から同一の論文間にパスができている場合は 2 回以上カウントしない(1 回とカウントする). 同じ MeSH 用語が同じ２つの論文間の「肺癌」と薬名の関係の場合は重みを付け, 通常よりも重視する.

前述したいくつかの重みを合計した値で順序付けを行う.特に有効であると考えられる組み合わせについては, 分子シミュレーションを行い, 実証を行う予定である.

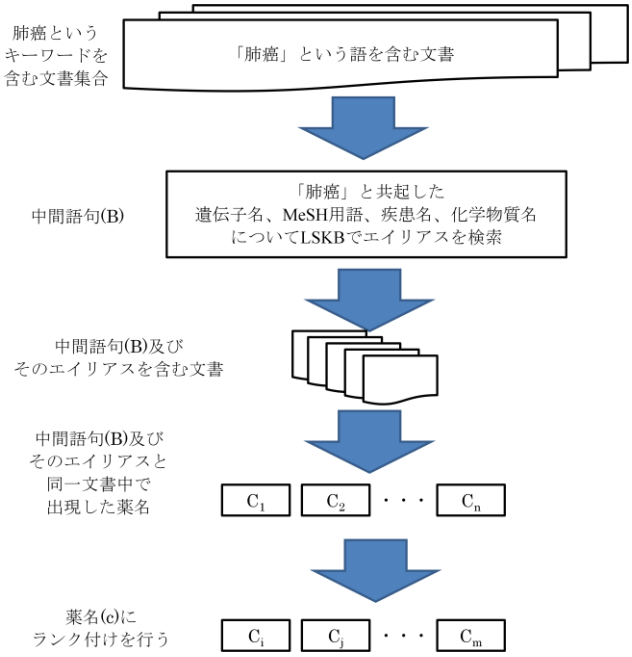

図 2:研究の全体像

4.4. 今後の課題

今後の課題としては, 上記のアルゴリズムを実現するためのプログラムを作成し, 収集した「肺癌」が記載されている文献に対してアルゴリズムを適用する。得られた実験結果の問題点を見つけ, プログラムに改良を加える. 実験結果がアルゴリズムそのものに問題があると考えられる場合には, アルゴリズムに修正を加えて, 科学的に意味のある結果を算出する. アルゴリズムを適用することで「肺癌」と関係があるとされた薬名が大量に出てくることが予想される. このため候補の導出条件を厳しくする手法を見つけ出し, 洗練された結果を出す必要があると考えられる.

収集しているデータは疾患名「肺癌」の記載されている文献で, なおかつ掲載年が 2003 年～2007 年となるものを検索した結果に対する考察であったが, 前述した方法は, 他の疾患名に対しても有効であると考えられる. 今後は,「肺癌」だけではなく他のデータについても収集および分析を行い, 各々のデータについてアルゴリズムを適用していきたいと考えている.

5. おわりに

本論文では, 文献に基づく知識発見について, 我々が検討している内容について述べた。医療分野で蓄積

されている学術論文から，学術文献内に出現する項目の関係を用いて，新たな項目間の発見を目指す。ここでいう項目とは病名，薬名，化学物質名，遺伝子名などを指し医学，生物学的に興味深い新たな発見をすることが期待されている．我々は特に病名・遺伝子と薬名･化学物質名との関係に着目している．我々はLSKBデータベースを利用して，MEDLINE の文献解析を検討している。

まず，現在我々が保持しているLKSBデータベースのデータ量について述べた。また，収集した文献データについても述べた。研究の初期段階では「肺癌」を含む文献集合を解析対象としており，文献データはNCBI のツールを用いて収集している。また，研究の全体像を述べ分析方法について考察した．

今後は，計画に基づいて新たな類似項目発見のシステムを構築する予定である。また，既知の関係を用いて適用アルゴリズムの評価を行う。今後，有用そうな関係が導出された場合には計算機による分子動力学シミュレーションを行い，類似関係の検証を行う予定である．

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] NCBI,PubChem,<http://pubchem.ncbi.nlm.nih.gov/> 2004

[9] 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>

[10] NLM. UMLS (Unified Medical Language System); 2004. National Library of Medicine.

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

[12] NCBL, GenBank

[13] EBI(European Bioinformatics Institute), EMBL

[14] 国立遺伝学研究所, DDBJ

[15] Batchentrez<http://www.ncbi.nlm.nih.gov/sites/batchentrez>