既知用語辞書を用いた
情報フィルタリングによる
研究シーズ用語の抽出方法

筑波大学
図書館情報メディア研究科
2019 年 1 月

小河邦雄

概要

既知用語辞書を用いた情報フィルタリングによる
研究シーズ用語の抽出方法

【背景】

研究シーズは、研究ニーズを解決して製品開発を実現するために必要な技術である。研究シーズの探索には、研究ニーズに関する用語を使用して文献データベースを検索することが多い。しかし、大量の検索結果から研究シーズ用語を網羅的に収集することは、負担が多く、漏れが生じる危険性が懸念される。とくに、研究報告が少ない研究シーズ用語は、研究初期の新しい発見に関わる可能性があるが、見落とすことなく収集することは容易ではない。これらを抽出したリストは研究シーズの探索の有用な資料となる。とくに創薬研究では、研究シーズの情報を早い段階で見出すことが求められている。

【目的】

本研究は、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法を提案し、抽出実験によってその妥当性を明らかにすることを目的とする。研究シーズの対象は創薬研究に必要な薬理メカニズムの情報とする。

【研究課題】

目的を達成するために、次の研究課題を解決する。

研究課題 1 として、新奇な研究シーズ用語の候補を抽出する方法を提案する。

研究課題 2 として、新奇な研究シーズ用語の抽出実験により、提案した方法の妥当性を確認する。

【方法】

二つの研究課題を解決するために次の方法を用いた。

研究課題 1 の抽出方法については、専門用語辞書を利用した情報フィルタリングの方法を参考に検討した。はじめに、文献データベースを検索して得た文献レコードを使用してデータレコードを作成し、得られたデータレコードを集めて研究シーズ用語抽出のためのデータセットを作成する。つぎに、データセットと既知用語辞書との照合によりヒットしたデータレコードを既知の研究シーズ用語を含むとして削除する。残ったデータレコードのテ

キスト情報をストップワードなどの区切り文字を使用して自動分割し、選別ルールを用いて研究シーズの対象である薬理メカニズム用語や関連する標的分子用語を研究シーズ用語として選別する。選別された研究シーズ用語が文献数の少ない薬理メカニズム用語であることを確認することにより、新奇な研究シーズ用語が抽出できたとみなす。

研究課題 2 の妥当性の確認には、設定した疾病の治療薬に関係する薬理メカニズム用語を研究シーズ対象として、文献数が少ない研究シーズ用語の候補を抽出する実験をおこなう。抽出実験により得られた研究シーズ用語は、疾病名とともに文献データベースを検索し、文献レコード数を調査する。検索結果の文献レコード数の変化を調べることにより、得られた研究シーズ用語の新奇性を確認し、提案する抽出方法の妥当性を確認する。

【結果】

研究課題 1 である新奇な研究シーズ用語の抽出方法については、4 抽出段階に分けて検討した。

抽出段階(1)として、既知情報のみを除いて新奇な研究シーズ用語を得るための情報フィルタリングに必要な既知用語辞書の作成について検討した。その結果、すでに開発されている医薬品の薬理メカニズム用語とその同義語を収集して、既知用語辞書とすることにした。そのために、開発薬理メカニズム用語は医薬品の研究開発情報データベースの Pharmaprojects、Integrity、明日の新薬から収集することにした。

抽出段階(2)として、抽出段階(1)で作成した既知用語辞書とデータセットとの照合方法を検討した。その結果、データセットと既知用語辞書との照合には、論理積もしくは近接演算子による方法が使用できることがわかった。また、照合によりヒットしたデータレコードをデータセットから除くことにより、新奇な研究シーズ用語が含まれるデータレコードに絞り込む方法を使用することにした。抽出段階(2)の処理は自動化を検討した。

抽出段階(3)として、情報フィルタリングにより得られたデータレコードから研究シーズ用語を選別する方法について検討した。その結果、選別は、選別ルールを用いて手動による方法を用い、選別した用語のデータ・クリーニングについても実施することにした。また、情報フィルタリングにより得られた IT 索引が多い場合は、ストップワードにより用語に分割して得たデータから、選別ルールとして作用語を含むもしくは大文字を含むなどを用いて部分的な自動化と手動とを組み合わせて研究シーズ用語を選別することにした。

抽出段階(4)として、抽出段階(3)で多くの研究シーズ用語が選別された場合に対応する方法について検討した。その結果、研究シーズ用語を用いて文献データベースを検索して得ら

れた文献レコード数を使用して関数の STF-IDF を計算し、研究シーズ用語を削減することにした。

つぎに、研究課題 2 の提案した方法の妥当性を確認するための実験をおこなった。実験では、準備段階(1)から(3)の準備をおこない、そのあと抽出段階(1)から(4)をおこなった。

準備段階(1)として、研究ニーズとして疾病の breast cancer と lung cancer の治療薬の研究開発を選定した。また、情報源として文献データベースの Chemical Abstracts（CA）を選択した。

準備段階(2)として、疾病名と薬理メカニズムに含まれる作用語が同じ IT 索引に含まれることを指定して CA を検索した。

準備段階(3)として、検索結果の文献レコードを IT 索引単位に分割し、文献 ID、索引 ID、タイトル、統制語、テキスト説明句、切り出し語の情報要素から構成されるデータレコードを作成し、これらをまとめたものをデータセットとした。準備段階(3)は自動化した。つぎに、提案した方法を用いて新奇な研究シーズ用語の抽出実験をおこなった。

抽出段階(1)として、医薬品の研究開発情報データベースを疾病名で検索した結果を使用して既知用語辞書を手動により作成した。抽出段階(2)として、データセットと既知用語辞書との照合をおこない、照合されたデータレコードをデータセットから削除し、新奇な研究シーズ用語が含まれるデータレコードの要素である IT 索引を得た。抽出段階(2)は自動化した。抽出段階(3)として、IT 索引から研究シーズ用語を選別した。選別は選別ルールとして薬理メカニズム用語に含まれる作用語を目印にするなどを用いて手動によりおこなった。lung cancer の実験では選別された研究シーズ用語が多かったため、抽出段階(4)として、研究シーズ用語の順位付けをおこない、PubMED を検索した文献レコード数などを使用した STF-IDF 値により順位付けすることにより、研究シーズ用語を削減した。抽出段階(4)は、自動化した。

【考察】

実験の結果に基づいて、本研究で提案した抽出方法について、文献レコード数に基づく新奇性、新奇な研究シーズ用語の抽出における情報フィルタリングの役割、新奇な研究シーズ用語抽出の意義と展望の点から考察した。

はじめに、文献レコード数に基づく新奇性の確認について述べた。疾病名と共に、抽出した研究シーズ用語を用いて CA を検索し、得られた文献レコード数を評価に用いた。その結果、breast cancer および lung cancer の場合ともに、文献レコード数が少なかったことか

ら、提案した抽出方法により新奇な研究シーズ用語が得られていることが確認できた。さらに、文献の発行年代を区切った検索による評価をおこない、初期の研究段階の情報が多く得られることを確認した。

つぎに、新奇な研究シーズ用語の抽出における情報フィルタリングの役割について述べた。とくに、探索目的に適した既知の範囲の既知用語辞書の使用の必要性、論理積および近接演算子を使用する照合方法の相違点、研究シーズ用語の選別ルールの妥当性、新奇な研究シーズ用語の属性の確認、研究シーズ用語の順位付けについて述べた。

最後に、新奇な研究シーズ用語抽出の意義と展望について、イノベーションの支援、人による判断の支援の意義を述べた。

【結論】

本研究により提案した、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法により、新奇な研究シーズ用語を抽出できることを明らかにした。

本研究により提案した抽出方法は 4 抽出段階から構成される。抽出段階(1)では、疾病名を検索語として医薬品の研究開発情報データベースを検索することにより、薬理メカニズム用語を収集して既知用語辞書を作成する。抽出段階(2)では、データセットと既知用語辞書との照合をおこない、ヒットした IT 索引を含むデータレコードを削除することにより、研究シーズ用語が含まれる IT 索引を選別する。抽出段階(3)では、選別ルールおよび手動により薬理メカニズム用語を選別する。抽出段階(4)では、件数が多い場合の処理について STF-IDF 値で降順にソートすることにより、研究シーズ用語を削減する。抽出段階(2)と抽出段階(4)については、処理の自動化をおこなった。

得られた研究シーズ用語の新奇性は次の方法により確認した。すなわち、疾病名と研究シーズ用語を検索語として用いて CA を検索し、文献レコード数を調査した。その結果、検索された文献レコード数が少なかったことから、疾病との関係において新奇な研究シーズ用語が得られていることが確認できた。

本研究では、研究シーズ用語を抽出するための情報源には調査主題に関する情報を含み、詳細な索引情報が付与されている CA を使用した。しかし、MEDLINE などの他のデータベースも使用できる可能性がある。しかし、付与されている索引情報は異なっているため、文献タイトルや抄録を対象とした情報フィルタリングの方法の検討が必要である。情報源を拡大または変更する場合には、データベースの特徴を考慮して、データセットの作成方法

の調整も必要となる。

また、本研究は、対象とした疾病が breast cancer と lung cancer に限られているため、他の疾病を対象とする際には細部の調整が必要であると考える。そして、この調整により創薬テーマ以外の研究シーズ用語の探索にも適応できる可能性があり、新しい主題領域への適応を検討していきたい。

Abstract

Extracting Research Seeds Terms by Information Filtering Using a Known Term Dictionary

[Background]

Research seeds (RSs) are essential technologies for product development to address research needs. RSs are usually retrieved from bibliographic databases using RS terms (RSTs) related to research needs. However, extracting a comprehensive list of RSTs from the massive search results of known research needs is a daunting task, and there is a risk of overlooking relevant terms. In particular, RSTs contained in a small number of research articles, which may well be new findings, might slip through the net. A list of such low-frequency RSTs is potentially useful in searching RSs. Detecting RSs early is especially needed in drug discovery research.

[Objective]

This study proposes a method of extracting candidates for novel RSTs from bibliographic databases and clarifies its validity through extraction experiments. The RSs search deals with information on pharmacological action mechanisms (PAM) essential for drug discovery research.

[Research questions]

Research questions to be solved for achieving the objective.

Question 1: Propose a method of extracting candidates for novel RSTs.

Question 2: Clarify the validity of the proposed method through extraction experiments.

[Methods]

Methods proposed for solving the two questions.

Question 1. Extraction method means information filtering using a dictionary of technical terms. The first operation was to assemble data records yielded by a bibliographic database search to create a dataset for the extraction of RSTs. Data records matching the known term dictionary (KTD) were eliminated from the datasets, assuming the correlation suggests terms were known. Then, by automatically dividing the remaining text information of data records with stop words, PAM terms (PAMTs) and the target molecule terms were selected. Finally, by checking that these RSTs were the PAMTs in a small number of articles, it was assumed that novel RSTs were selected.

Question 2, the validity of the proposed method was clarified by extracting candidates for novel RSTs in a small number of articles in the case of PAMTs for the cure of certain diseases as the target of research for experiments. The number of documents with the RSTs and disease names were examined in a bibliographic database. The novelty of the RSTs was confirmed by the differentiation in the number of documents, as was the validity of the proposed extraction method.

[Results]

In Research Task 1, the extraction method for novel RSTs was examined in four steps.

Extraction step (1), a KTD was built by collecting PAMTs related to developed medicinal supplies and their synonyms from research databases such as Pharmaprojects, Integrity, and Asu-no-Shinyaku.

Extraction step (2), matching the KTD from step (1) and retrieving datasets was studied, and an AND operator and a proximity operator were found useful. A method of filtering data records that include novel RSTs was established by eliminating data records that matched the KTD.

Extraction step (3), selecting RSTs through data records was studied, and resulted in using a manual selection method according to a selection rule, and

conducting data cleaning of selected terms. When a large IT index remained after information filtering, separation of terms including action words and capital letters—after categorizing data into terms by stop words, with a functional or a manual operation—was conducted.

Extraction step (4), dealing with the massive amount of PAMTs selected in extraction stage (3) was studied, and resulted in using the number of records obtained through the bibliographic database search as indicators. Calculating the rating score with the STF-IDF function led to a smaller number of documents.

Two experiments were conducted to confirm the validity of the method proposed in Question 2.

Preparation step (1), R&D in curatives for breast and lung cancer was chosen as research needs, and Chemical Abstracts (CA) as a source.

Preparation step (2), selected disease names and action terms included in PAMTs were designated to be under the same IT index, and CA was searched.

Preparation step (3), records of search results were divided into IT index units, and data records such as document IDs, index IDs, titles, controlled terms, text modifiers, and segmented terms were compiled into a dataset, followed by extraction experiments of novel RSTs by automatization.

Extraction step (1), a KTD was created from a disease-name search of a database of medicinal supply R&D. Extraction step (2), data records matching the KTD were removed automatically. This yielded an IT index as an element of data records in which novel RSTs were included. Extraction step (3), RSTs were selected from the IT index, manually, using action terms included in the PAMTs as markers. Concerning lung cancer experiments, to deal with the massive amount of data, as Extraction step (4), RSTs were ranked by STF-IDF values calculated from the number of database records in PubMed automatically.

[Discussion]

Based on the results of the experiment, here I discuss three aspects of the validity check.

First, the validity of the proposed method was evaluated using the number of document records as an indicator of novelty. As an evaluation, the number of documents obtained through a search of CA using RSTs and disease names was counted. The result showed a high novelty of RSTs was obtained by the proposed method because the number of document records with RSTs in breast and lung cancer searches was minimal. Furthermore, novelty also was confirmed through a chronological comparison.

Second, the role of information filtering in the extraction of novel RSTs was discussed, especially with respect to the necessity of using relevant KTDs; as was the difference between matching methods using AND operators and proximity operators; the validity of selection rules; the confirmation of the novelty of RSTs; and the ranking of RSTs.

Finally, the significance and outlook of the proposed information filtering method was considered. As a method of finding knowledge, this method will support innovation through the extraction of novel RSs. Furthermore, it will support human evaluation of novelty through partial automation of processing RSs.

[Conclusion]

This study showed that novel RSTs can be retrieved by extracting candidates of novel RSTs from a bibliographic database using the proposed extraction method.

The proposed method consists of four extraction steps. Step 1, an R&D medicinal supplies database is searched using the names of diseases as search terms, PAMTs are collected, and a KTD created. Step 2, correlations are checked between a given dataset and the KTD. Then, an IT index with novel RS is used to automatically remove data records with an IT index that contain positive correlations with known terms. Step 3, using a selection rule and

manual operation, relevant PAMTs are selected. Step 4, when processing a large amount of data, candidates for novel RSTs are automatically downsized by sorting in descending order of STF-IDF value as a determiner of novelty.

The novelty of obtained RSTs was confirmed as follows. CA was searched using RSTs and disease names as research terms and the number of document records noted. The number obtained was confirmed to be small, which established the novelty of RSTs in relation to the target diseases.

In this study, CA was used because of its ample and detailed information as a main investigation subject. However, such information filtering could be applied to other databases such as MEDLINE, with such adjustments as including thesis titles and sentence units as markers. When expanding or changing data sources, an adjustment should be made in creating a dataset, according to the features of the databases.

Since this study was limited to breast and lung cancer, the method needs specific adjustments when studying other diseases. With such adjustments, it could be applied to searches for RSTs beyond drug discovery, which is a subject for future study.

目次

図目次

表目次

第 1 章 序論

本研究では、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法を提案し、抽出実験によって提案した方法により新奇な研究シーズ用語の候補が抽出できることを明らかにした。本章では研究の背景（1.1 節）、研究の目的と範囲（1.2 節）、関連研究（1.3 節）および本論文の構成（1.4 節）について述べる。

1.1 研究の背景

製品やサービスとして社会から求められている機能がニーズであり、研究目的として優先される傾向があるが、それらのニーズを実現するための科学的方法や革新技術などのシーズも、製品として成功するためには非常に重要である[1]。よく知られた既知のシーズを用いてニーズを解決できる製品を開発しても、競合が多くなり市場価値の維持は難しい[2]。そのため、ニーズを実現するシーズをいかに研究テーマとして設定するかということが研究開発の大きな命題である。シーズ志向の新製品を開発するため、技術経営（MOT：management of technology）の観点からも研究と実践が進んでいる[3][4]。

日本は資源の少ない技術立国であり、高い技術を使用した付加価値の高い製品やサービスによって国力を高めていくことが求められている。科学技術を使用して、今までにない画期的な製品やサービスを生み出していくことができれば、世界的な競争力をもつことが可能となる。このようなイノベーションを実現するためには、十分に調査、分析をおこない、社会からの要望としての研究ニーズとその解決や実現を可能にする適切な研究シーズを見出す必要がある。

社会の大きなニーズがあるにもかかわらず良い解決方法がない場合、それを解決するための研究の意義は大きい。とくに医療分野においては、治療する薬剤のない疾病は現在においても多く残されているため、それらを治療する医薬品を開発する創薬研究は社会から期待されている。

ある病気の治療薬を開発する創薬研究では、その病気の治療に関係する体内タンパク質などの研究シーズを見落とすことなく調査し、選択した研究シーズを使用した実験によって医薬品として最適な候補化合物を見出す必要がある。このため、病気の発症の仕組みに関係するタンパク質である標的分子などを探す方法、標的分子に対する薬剤の作用である薬

理メカニズムを探す方法、および化合物の薬理メカニズムに着目して開発候補薬剤を探す方法を組み合わせて調査することになる。なお、本論文では、医薬品を開発するための研究を創薬研究とし、研究の目的や方法について決めたものを研究テーマ、研究の結果として社会などから求められている機能を研究ニーズ、研究ニーズを実現するためのテクノロジーを研究シーズとする。

製薬企業においても、疾病の治療に関係する多様な研究シーズの探索が不可欠となっているため、創薬テーマとして重要な研究シーズを外部の大学、公的機関の研究に求めるオープンイノベーション[5]を推進するなど、多くの研究機関との情報や人的なネットワーク構築の必要性が高まっている[6]。大学は、産学連携によって大学の基礎研究を具体的な社会における成果に結びつけるために、研究シーズ情報の公開に力を入れている。製薬企業は、これらの研究シーズ情報も参考にして研究テーマを企画し、創薬研究に取り組んでいる。

製薬企業においては、研究者などによって医薬品開発の研究テーマである創薬テーマが発案されると、情報調査部門において新規性調査を実施し、研究企画部門が企業戦略との関係や事業としての成功率などを考慮して創薬テーマ採用可否を検討する場合が多い。発案された研究テーマのうちでも、とくに医師による処方箋が必要な医薬品の開発は、多くの製薬企業において成果に結びつくことが難しい状況にあり、情報を活用して少しでも改善することが求められている。これらの創薬テーマを発案する際には、創薬テーマの発案者が文献や学会などから薬理メカニズムや疾病に関する情報を入手し、先行研究を調査している。しかし、情報調査部門において網羅的な調査をすると先行研究が多く認められる場合があり、競合性の問題から、創薬テーマとして採用されない傾向がある。提案された薬理メカニズムが著名な学術雑誌に新しく掲載された有望なものであり、それが関係する疾病が患者や医療関係者からの治療ニーズとして大きいものである場合は、新規性と情報源の確実性の観点から研究テーマとして採用されることもある。しかし、著名な雑誌に掲載された研究テーマは研究開発を進めても、世界の製薬企業との競合になる場合が多く、競争の面から研究開発を断念するケースも多い。

近年の創薬の成功率は低く、合成した物質から医薬品として承認にいたるまでの累積成功率は、日本製薬工業協会の「DATA BOOK」によると 2009 年から 2013 年で 29,140 分の 1（0.003%）であり、極めて低い[7]。成功率を高めるためには、研究開発プロセスに関係するすべての要素の改善が必要であるとされている。なかでも研究シーズとしての標的分子の選択はその後の影響が大きいため重要な要素であり、より有用な研究シーズの探索

が求められている。すなわち、創薬テーマを発案する場合は、患者や医療関係者からの治療ニーズを解決する有用な研究シーズ情報を早い段階で入手して、競合者と差別化した薬剤スクリーニングを実施することが創薬としての成功には不可欠となる。そのため、疾病に関係すると思われるタンパク質などの新しい標的分子とその機能を最新の医学基礎研究の中から探索することは、創薬研究の重要なプロセスの一つとなっている[8]。

新たな研究テーマの発案に必要な研究シーズを探索する方法として文献検索を利用する方法がある。これは、研究ニーズに関係する検索語を使用した網羅的な文献検索の結果から、研究ニーズを解決するためのヒントとなる研究シーズに関する用語リストを作成して、有望なものを確認していく方法である。しかし、このような探索的な目的に対して、網羅的で適切な検索をおこなうことは、調査担当者にとっても容易ではない。研究ニーズに関する検索語を使用して検索すると数千から数万の膨大な数の文献レコードが得られる場合がある。そのため、すべての文献レコードを読んで研究シーズ用語のリストを作成することは、情報の過負荷（information overload）[9]となり、適切に処理ができないため情報の漏れなどを生むことが問題となる。過負荷を避けるためには検索結果を追加のキーワードを使用して絞り込む方法もあるが、その結果として研究シーズとして期待できるものを失う危険性があり、追加検索語の選択も難しい。このように、検索結果が大量の場合は、目的とする情報を適切に入手することは簡単ではない。

とくに創薬研究の場合は、研究や開発が進むと、薬理メカニズムや関連する体内タンパク質、体内タンパク質を標的として開発された薬剤について記載された文献、特許や臨床報告などが大量に生産される。そのため、文献データベースを疾病名により検索すると数万件の文献レコードが得られる場合もある。その検索結果を薬理メカニズムの観点により整理して研究シーズのリストを作成することは、情報量が多いことが原因となって見落としが発生し判断の誤りを招く問題がある。文献データベースの検索結果から文献レコード数が少ない薬理メカニズム用語を抽出することができれば、それらは研究の初期段階の可能性が高い研究シーズ用語のリストになり、研究シーズの探索に役に立つ

大量の検索結果から必要な情報を見つける方法としては統計的手法があり、出現回数が多い用語を見つけるための機能を提供している検索システムがある。たとえば、SciFinder、JDreamⅢでは、情報フィルタリング機能として統制語の頻度集計結果が表示され、検索に使用できる[10][11]。しかし、研究テーマ設定のための研究シーズ探索には、検索結果の新奇性（novelty）や意外性（serendipity）、多様性（diversity）などが必要となる。そのため、

収集した情報のなかから、大量の既知情報を除いて、それ以外の低頻度な情報を抽出し、リスト化できることが望ましい。しかし、一般に低頻度の用語の抽出は異なり語数が多い傾向にあり、統計的手法による自動化の適用は難しいとの研究がある[12]。そのため、検索結果から低頻度の情報を識別するためには、人の経験と知識に大きく依存することになり、ふたたび情報の過負荷が問題となってくる。大量の既知情報を含む検索結果のなかから、自動的な処理により低頻度の情報を抽出することができれば、イノベーションにつながる画期的な研究も期待できる。

1.2 本研究の目的と範囲

本研究は、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法を提案し、抽出実験によってその妥当性を明らかにすることを目的とする。この妥当性は研究シーズ用語の候補を抽出し、その新奇性を文献レコード数により確認することで判断する。研究シーズの対象は創薬研究に必要な、疾病に関係する薬理メカニズムの情報とする。

本研究では、文献データベースを研究ニーズに関する用語を用いて検索して得られた文献レコードから、研究シーズ用語を含む索引部分を取り出し、既知の範囲をデータベースへの収載により定義して作成した既知用語辞書による情報フィルタリングによって、既知用語が含まれる情報を除いて、設定した疾病に関して文献数の少ない新奇な研究シーズ用語を抽出する方法を検討する。

そのために、研究課題 1 として、新奇な研究シーズ用語の候補を抽出する方法を提案すること、さらに研究課題 2 として、抽出実験により提案した方法の妥当性を確認すること、の二つを設定し、これらを解決する。なお、ここでは、新奇は特許で使用される過去に例が全くない意味の新規とは異なり、あまり知られていない、珍しい、目新しい、稀という意味で使用する[13]。すなわち、設定した疾病と同じ文献に出現する頻度の少ない、報告された文献数が少ない研究シーズ用語を新奇な用語とする。実際に医学的に疾病と関係がある研究シーズ用語であることの確認については、研究の範囲としない。

創薬研究の研究テーマ探索に必要となるライフサイエンス分野の文献は、年間 100 万件以上発行されていることから、効率的な調査が不可欠となっている[14]。創薬研究は、疾病の治療を研究ニーズとして、疾病の原因に関係する「標的分子（体内タンパク質など）」、薬剤の標的分子への作用である「薬理メカニズム（作用機序）」、有用な薬理メカニズムが期待できる「薬剤骨格（化学構造式）」の三つを研究シーズとして選択して、化合物のスクリーニング（選別検査）をすることが基本となっている。本研究において探索対象とする研究シーズは「薬理メカニズム」に関する情報とする。ただし、「標的分子」も薬理メカニズムに関係する情報として研究シーズに含める。

また、創薬研究では、研究開発が進むと文献が大量に生産される。新奇な研究シーズを探し出すためには、大量の文献の中から研究の初期段階の文献を探し出す必要がある。しかし、初期段階にある研究は当該疾病を創薬テーマとする研究者であればわかるが、大量の文献を処理することは負担が多く、困難を伴う。また、調査担当者がこれらの処理をおこなう場

合も、当該疾病の薬理メカニズムに関して知識が少ないと判断は難しい。既知の研究シーズを把握して、大量の文献の中から除くことができれば、新奇な研究シーズのみを取り出すことが可能になると考えられる。大量の文献から既知の研究シーズを取り除く方法として、情報フィルタリングを利用できないか検討した。

本研究では、疾病の治療薬を開発するために必要な研究シーズとしての薬理メカニズム用語や標的分子用語と、研究ニーズである疾病との意外性の高い稀な組み合わせを効率よく抽出する方法を検討する。しかし、どの程度の頻度が稀であるかは、調査者の知識、経験などによって異なる相対的なものであり、客観的な数値として設定することは難しい。また、得られた新奇な研究シーズ用語を創薬研究に用いて最終的な医薬品となるまでの結果を示すことは、医薬品の研究開発過程には非常に多くの要因が関係し、臨床試験も必要であるため困難である。本研究でおこなう創薬に関する実験においては、検索システムの評価に多く使用される適合集合を基にした再現率や精度を適用した評価はできない。そのため、本研究の範囲は、抽出した研究シーズ用語と研究ニーズである疾病との関係性が新奇であることを確認するまでとする。

1.3 関連研究

本章では本研究に関連する次の先行研究について概観する。すなわち、

(1)研究シーズの探索方法：とくに疾病の治療薬を研究ニーズとする場合の研究シーズ探索に関する研究、

(2)創薬の研究シーズ調査に関する情報源：本研究の研究シーズである薬理メカニズム情報を抽出するための情報源に関する研究、

(3)テキストマイニング：情報抽出に関係する自然言語の処理で必要となるテキストマイニングの研究、

である。

1.3.1 創薬を中心とした研究シーズの探索方法

研究テーマ探索の方法については多くの先行研究があるが、本項では研究テーマとして設定した創薬研究のニーズである疾病の治療に対して、関連する研究シーズの探索に関して概観する。この研究ニーズに対する研究シーズの「シーズ」という用語は、日本において広く使用されているが、海外では technology や idea、solution などが用語として使用されている。研究ニーズに必要な研究シーズ、とくに創薬に関係した情報を中心に探す方法を整理し、表 1 に示す。

表 1. 創薬研究における研究シーズ探索の方法

	方 法	内 容	網羅性	意外性	ノウハウ調査
①	文献調査	研究シーズが記載された文献（文献、特許など）を調べて探す。	○	△	×
②	研究シーズ保有者による情報公開	大学などにより研究シーズ情報として Web サイトなどで公開された基礎研究から探す。	×	△	×
③	研究シーズの公募	研究ニーズを公開し、それを解決できる研究シーズをもつと考える研究者からの研究シーズ情報を公募して探す。	×	○	○
④	テクノロジー・スカウティングサービス	研究ニーズと研究シーズを結びつけるマッチングの専門機関に依頼して探す。	△	○	○

※表中、○は網羅性、意外性が高いもの、ノウハウ調査が可能なもの、△は網羅性、意外性がある程度得られるもの、×は網羅性、ノウハウ調査が難しいものを指す。

①の「文献調査」の方法は、研究の成果として公表される特許、文献などを文献検索して調査する方法であり、ある程度の網羅的な調査が可能である。

この方法は、研究者が研究成果として技術を公表していることが必要である。また、文献調査の方法が利用できるためには、研究シーズとなる技術を用いて解決できる研究ニーズも記述していることが必要である。探索の目的とする研究ニーズについて記載されていない場合は検索することは難しい。この「文献調査」の方法では、目的とする研究ニーズに関する文献数が少ない場合は、研究ニーズに関係するキーワードで文献検索し、得られた全文献を読むことにより研究シーズ探索の目的は達せられる。

しかし、医学分野のように、研究ニーズとしての疾病に関する文献が非常に多い場合は、文献検索で得られた全文献を読み、研究シーズ用語を抽出することは難しい。そのため、研究ニーズに関する情報が多い場合の研究シーズ探索を支援する方法など、研究ニーズに研究シーズを関連付けて調査する方法についての研究が必要とされ、システムが提案されている[15][16]。とくに、医薬品に関係する文献は非常に多いため、研究ニーズである疾病に対して、ある程度知られた研究シーズに関しては医薬品の研究開発情報のデータベースによって整理されている。しかし、頻度の少ない情報の収録は十分ではない。これらの低頻度情報は、技術経営の観点からも競争優位のための差別化戦略に必要とされ、重要な情報と考えられている。低頻度の技術情報の入手に関する研究については、菰田らの報告がある[17]。

ここでは、情報担当者と研究者による協働した調査により、低頻度であるが重要な技術に関する情報のリストを作成する方法が提案されている。しかし、自動化された方法ではないため、大量の情報を処理することには適していない。

②の「研究シーズ保有者による情報公開」を利用する方法に関しては、基礎研究に力を入れているほとんどの日本の大学において、大学の Web サイト上で研究シーズ情報の発信をおこない、共同研究などの提案を募っている[18]。東京大学の産学協創推進本部では、「東大シーズ集 東京大学産学連携プロポーザル」として各種研究分野に関する 1,616 件の提案テーマを公開し、共同研究を推進している[19]。また、筑波大学では「筑波大学・研究シーズ検索」の Web サイトから、各種の技術分類によって研究シーズを探すことや汎用連想計算エンジン（GETA）により、派生した概念を含めた検索を可能にしている[20]。研究シーズが必要な企業は、大学の研究シーズに新奇性とともに、技術としての確実性も求めている。企業の立場からは、大学が積極的に発信する情報は信頼性の高い情報源の一つとしてとらえられている。しかし、公開される情報は特許出願されたものが中心であり、学術文献を含まないことが多く、大学がもつ研究シーズのなかでも限られた範囲の情報のみが公開されていることになる。また、調査する場合は、各大学の Web サイトを個々に見る必要があり、調査効率が良くない。このように②の「研究シーズ保有者による情報公開」を利用する方法は、研究シーズ探索としての網羅性に関しては不十分である。

③の「研究シーズの公募」に関しては、世界的なオープンイノベーションが進んでいる。米国の世界的な製薬企業であるイーライ・リリー社が取り組んでいる例がある。自社で販売する製品の半分を社外の研究シーズによるものにする必要があるとの外部コンサルティングの結果から、1999 年に外部シーズ導入の専門部門を設立し、世界の大学やベンチャー企業などから積極的な研究シーズの公募をおこなっている[21][22]。これらに対する外部からの提案は年間数千件と多く、イーライ・リリー社は文献調査をおこない、提案を採用しない場合でも、できる限り科学的な評価を付けた回答をしている。これは、採用しない研究シーズ提案の場合も、不採用の理由に関して客観的な根拠を示すことが、より多くの提案を得るためには必要との考えからである。このように、公募に応じて提案された研究シーズを利用する場合においても、他の研究シーズと比較した客観的な観点から選択するためには、他の研究シーズを含めた網羅的な調査が必要であり、①の「文献調査」による研究シーズ情報の網羅的検索は不可欠である。この文献調査の負担は、文献データベースの検索結果から研究シーズ用語を効率よく抽出する方法があれば、軽減することができる。しかし、これらの負

担を改善するための情報処理の方法として、研究シーズ探索の文献調査をするために文献データベースの検索結果から必要な用語を抽出する方法についての研究は少ない。

この流れのなかで、日本でもオープンイノベーションへの期待が大きく[23][24]、そのための人材育成も進んでいる[25]。日本の大手製薬企業も数社が既に実施し、国内外の大学やベンチャー企業との共同研究の実績も報告されている[26][27][28]。しかし、研究シーズの保持者が研究シーズを募集している製薬企業の Web サイトを見るとともに、自分の研究が研究ニーズを解決する可能性があることを認識する必要がある。

④の「テクノロジー・スカウンティングサービス」に関しては、企業から依頼を受けた研究ニーズと、大学などの研究機関が保有する研究シーズ情報とのマッチングを専門におこなう機関や企業が世界に複数存在し、研究開発が重要な産業である製薬企業でも多く利用されている[29][30]。主な利用形態としては、利用者の製薬企業から研究シーズが必要な疾病をマッチング機関に登録し、マッチング機関は世界の基礎研究機関との間に構築した情報ネットワークを介して、その疾病についての研究シーズ募集をおこなう。応募された複数のシーズの中から利用者の製薬企業は必要な研究シーズ選択をする。海外のマッチングをおこなう例としては、英国ではテクノロジー・スカウンティングサービスの名称で 10 年ほど前からはじめられている[31]。オックスフォード大学では、2007 年にこのサービスを始め、欧州地区の大手企業を中心にその利用が拡大している。とくに、クライアントの企業の研究開発ニーズを十分に理解するために事前のインタビューに力を入れ、それに適した技術を評価して絞り込む、また選択したコア技術が研究シーズとして発展する可能性のある研究ニーズについて予測するなどのサービスが高い評価を得ている。実際に、このサービスにより創薬の分野において複数の医薬品が市販され、10 種以上の薬剤の臨床開発が進められている。

テクノロジー・スカウンティングサービスのプロセスには 4 段階がある[31]。そのなかの広範囲な技術調査と評価をおこなう 2 段階目の「ロング・リスティング」では広範囲な調査によって網羅的な研究シーズのリストを作成する。このリストにより、研究シーズの選択に網羅性と客観性をもたらすことが可能になるとしている。このテクノロジー・スカウンティングサービスは、Allen[32]が主張する企業におけるゲートキーパーの機能を組織により段階的に実施できるようにしたものであると考えることができる。しかし、マッチング機関の意向により情報を収集・分析しているため限定的であり、得られる情報の網羅性は高くはないと考えられる。

以上のように、創薬に必要な研究シーズを探索する必要性は高く、独自の研究シーズを得るために多様な方法が使用されているが、それぞれ長所短所を有している。しかし、基本となる研究シーズ探索の方法は、研究成果が記載された文献を調査することである。そして、その結果から得られた研究シーズのリストを作成することにより、最終的には創薬の専門家である研究者が独自の感性により選択することが可能になる。

1.3.2 創薬の研究シーズ調査に関する情報源

医薬品の創薬に関する知識発見のため、文献に記載された遺伝子やタンパク質の情報から、疾病に関係するタンパク質の機能を予測する研究が多くおこなわれている。これらの遺伝子やタンパク質などの情報解析の研究には、医学領域の主要な雑誌を収録し、文献レコードには疾病などに関する統制された索引語が医学の視点から付与されているため、医学文献データベースの MEDLINE が使用されることが多い[33][34][35]。その他の文献データベースを使用した情報解析として、Chemical Abstracts（以下、CA）を使用した例がある。CA の収録範囲は化学だけではなく、生化学領域を含めた応用分野までを含み、医薬品に関しては非臨床の情報を多く収録している。そのため、小島らは化学構造と薬理活性との関係に着目し、CA の索引から特定の疾病に関係する薬剤の薬理メカニズムを機械抽出した研究結果を報告している[36]。薬理メカニズム用語は、CA の化合物の役割を示すロールとして uses（用途）と、合成を示す用語である prepn、機能語である as を目印として正規表現で記載したルールを利用し自動抽出をおこなっているが、再現率は高くはないため、手動によって補完している。また、岡らは、治験薬情報や市販化合物情報に注目し、CA の索引から薬剤の中間体情報を機械抽出する方法を提案し、試薬合成の事業に活用している[37]。

また、データベースの索引部分ではなく、電子ジャーナルなどの文献の全文データを対象としたテキストマイニングもおこなわれているが、大量の文献の本文テキストを使用できるのは出版者[38]などに限られ、通常の調査においてデータセットとして使用することは著作権の問題から難しい。

1.3.3 テキストマイニング

テキストマイニングは、手動では処理することが難しい大量の文書から情報や知識を探し出すことを目的とした研究分野である。第一に、既知用語を含む索引の除去に関係する情報フィルタリングの研究について述べる。第二に、自然言語の索引から専門用語を抽出する際に必要となる固有表現抽出に関する研究について述べる。第三に、抽出した用語リストを順位付けする際に使用する用語重み付けに関する研究について述べる。第四に、本研究で重視する低頻度情報の分析についてテキストマイニングを使用した研究について述べる。第五に、本研究の結果の評価に関連する分析の技術としてのテキストマイニングの特徴に関する研究について述べる。

1.3.3.1 情報フィルタリング

徳永[39]は、情報フィルタリングは情報探索者の興味や関心を記述したプロファイルを参照して、流れてくる情報のなかで、探索者の関心があるものだけを取り出すことである、としている。そして、情報フィルタリングの特徴としては、以下のようなものが考えられるとして Belkin[40]の文献を引用している。

- 扱う情報があまり構造化されていない
- 扱う情報はテキストが主であるが音声や画像などの情報を含む場合もある
- 扱う情報の量が大規模である
- 入力が情報のストリームである
- ユーザの長期的な嗜好を表現するプロファイルと呼ばれる情報を使う
- フィルタリングは必要なものを探すのではなく、不要なものを削除する

これらの特徴のいくつかは情報検索の技術、情報抽出の技術の特徴と重複する部分があるとしている。すなわち、必要な情報のみを取り出すという意味では情報抽出と似ているが、情報抽出は文書の内容を解析し、個別の事実情報を扱うことに重点が置かれている点が情報フィルタリングと異なっているとしている。また、情報検索と情報フィルタリングはどちらもユーザの情報要求を満たす情報をユーザに提供するという目的では同じであるが、い

くつかの相違点もあるとしている。とくに、情報検索は適合文書を選択することが重要であるが、情報フィルタリングは不要な文書を削除することが重要としている。このような違いはあるが、情報フィルタリングには情報検索のほとんどの基本技術を利用することができるとしている。

情報フィルタリングを文献検索と組み合わせて、不適な情報の除去に使用する研究がある。Kitamura らは、臨床データのデータマイニングにより発見された臨床データに関するルールの有効性を確認するため、文献の検索結果と組み合わせた情報フィルタリングをおこない、他の主題に適応可能な、価値のあるルールを見出す方法を提案している[41]。ここでは、肝炎データを事例として、肝炎の進行具合を示す血液データ GPT（glutamic pyruvic transaminase）が他の二つのデータである TTT（thymol turbidity test）と D-BIL（direct bilirubin）の変化によって予測できるという発見ルールについて報告している。実験例として、発見ルールに関係する用語である GPT、TTT、D-BIL と調査領域に関係する用語である hepatitis（肝炎）を使用して PubMED を検索し、その検索された文献レコード数を昇順にソートし、閾値を設定して文献レコード数の多いものを不要なものとして除去することによる情報フィルタリングをおこなっている。文献検索結果とソートの対応付けに関しては、

・文献レコード数が多ければ、既知といえる、

・出版時期の新しい文献が多ければ、時期的に新しい話題を扱っているといえる

ことを前提としている。

このように、文献レコード数からは既知かどうかを、文献発行年代のデータからは時期的な新しさを判断し、ルールの特性を確認している。文献検索と情報フィルタリングを組み合わせた研究は多くはないが、有用であると考えられるため、より発展させた研究が期待される。なお、文献検索による文献レコード数を使用して、よく知られている情報の確認をおこなう研究はおこなわれているが、よく知られているものを不要とする研究は多くない。

一方、適合性フィードバック[42]などを使用して不要とするテキストの集合を選択して分類に使用する方法がある。すなわち、ベクトル空間モデルの考え方を使用して不要とするテキストの類似テキストを機械学習によって判断し、分離する情報フィルタリングの研究であり、電子メールのスパムメール検出などに実用化されている[43]。ベクトル空間モデルによる検索機能の改善は有用であるとされているが、検索対象が文献タイトルのみや短い文章のような場合には効果が少ないとの限界も指摘されている[39]。これは、文書を表すベク

トルの有効な長さが短くなるためであり、フィードバックが必ずしも十分な信頼性を持ち、有効な方向に働かないためと考えられている。経験的に文書中の単語が 25 以上あることが適合性フィードバックとして機能する目安とされている。このため、抽出をルールでおこなうか、機械学習でおこなうかは、抽出対象によって選択すべきとの報告もある[44]。

1.3.3.2 固有表現抽出

自然言語から情報を抽出して分析するためには、専門用語や固有表現の抽出が必要であり、テキストマイニングの前処理として固有表現抽出（named entity extraction）に関する多くの研究がおこなわれている。用語抽出の研究は、自動化によるコスト低減を目的とした場合も多い。それは、専門用語を確実に抽出するためには、分析する領域の専門家が手動によりタグ付けすることが不可欠であるからである。しかし、増加する情報量に対応するためには、コストの問題からすべてを手動に依存するのではなく、用語を自動抽出する研究も必要である。用語抽出は用語辞書やシソーラスの作成、索引付けなどの自動化や自動翻訳への利用価値も高い。

用語抽出の方法として大きく三つの方法がある。一つ目は用語辞書を使用する方法である[45][46]。二つ目は、ルールを作成して抽出する方法である[47][48]。三つ目は、統計や確率的な考え方を使用する方法で、意味を用いるのではなく、語が含まれる文章単位や前後の語との距離や頻度を数値化して、機械学習により判断する方法である。これには、タグ付けされた学習用データを用いた教師あり学習の方法[49][50][51][52]や人の知識や認知的な判断など人が優位な部分を指定し、それを補足する形で機械学習をおこなうアクティブ・ラーニングを用いる方法[53]、半教師あり学習の方法[54]がある。確実な方法は辞書を作成する方法であるとされているが、辞書を作成すること自体にもコストが必要であり、領域を特化しない場合は、汎用性に問題がある。辞書の作成が難しい場合は、学習する方法が使用できるが、学習するためにも、コーパスとして正解の情報を判定した学習用のテキストが必要であり、これ自体にも手動によるコストが必要になる。これに対して、半教師あり用語抽出として、少数の用語辞書から出発して、これらを検索で用いることにより、その結果から用語を集計して、辞書に加えることを繰り返すという方法もある。このように、自然言語で書かれたテキストから、各分野で使用される専門用語を抽出することを目的として多くの方法

が研究されている。しかし、とくに優れた方法を選択するのは難しい状況でもあるとされ、既存の研究を基にさらに改良された方法が望まれている[55]。

1.3.3.3 用語重み付け

自然言語で記載された索引から、形態素解析の使用なしに複合語や略語を抽出すると、専門用語と一般用語が混在するため、それらを区別することができない。そのため、専門用語を一般用語と区別する方法が必要となる。この方法の一つとして、文書中の用語を重み付けする TF-IDF（term frequency-inverted document frequency）[56]の考え方がある。基本的な TF-IDF の式は、個々の文書における用語の重み付けに使用する経験則の関数（関数式 1）である。

$$\text{TF-IDF} = \text{tf} \times \left(\log\left(\frac{\text{n}}{\text{df}}\right) + 1\right) \quad \cdots \text{（関数式 1）}$$

tf：対象とする文書内における当該用語の出現頻度

df：文献データベース全体における当該用語が使用されている文献レコード数

n：文献データベースの全文献レコード数

対象とする文書内において出現回数 tf（term frequency）の多い用語は、その文書の重要な概念であると仮定して、関数の値が高くなる。一方、文献データベース全体の文献レコード数 df（document frequency）が多い語は、頻出語として特異性は低く、関数式 1 中では df の値が分母にあり、TF-IDF 値を小さくする方向へ働く。そのため、tf 値が同じ場合は、文献データベース全体において頻度の大きい用語の TF-IDF 値が小さくなる。この IDF は大域的重み（global weight）とされ、精度の向上を目的とする重みであり、特定の文書に集中して出現する用語に対して大きな値が与えられる[42]。n は文献レコード数（df）の値を得るために使用した文献データベースの全文献レコード数である。また、全ての文献に含まれる用語がある場合は、対数の計算をおこなうと 0 になる（log(1)=0）ことを防止するため、その調整として 1 を加える。

この TF-IDF を使用した研究として、辻[57]は、重要な専門語となる新語を特定・予測するために、雑誌の文献テキストを、生み出された時点に基づいて新旧の二つに分け、用語抽出手法として TF-IDF やその他の手法[58][59]を使用して比較し、予測の有効性を検討して

いる。また、この TF-IDF を応用した研究として、食事レシピ食材の希少性と一般性に基づいた、レシピに使用される意外性のある食材の組合せのパターンを抽出する研究[60]がある。ここでは、TF-IDF の考え方を応用した方法として食材の意外度を示す RF-IIF（recipe frequency-inverse ingredient frequency）を提案している。RF は全レシピなかで当該食材が出現するレシピ数を全レシピ数によって除したもので、RF が大きいことが食材に関する一般性を示し、IIF は当該料理のレシピ数を当該料理のレシピのなかで当該食材が出現するレシピ数で除した値の対数をとったもので、IIF が大きいことが当該料理に使用する食材の希少度を示している。さらに、RF と IIF の最適なバランスを調整するために、IIF については指数計算によって 1、3、5 乗した値を比較している。数値の選定については、より希少な食材の値を大きくしたい場合は、IIF を 5 乗した数値を用いている。このように、調査目的によって RF-IIF のパラメーターの重み付けに変化をつけている。得られた値は、順序尺度により大小を判断するだけに使用されるため、このような式の修正が可能になっている。

また、TF-IDF を創薬のための標的遺伝子の探索に応用した研究がある[34]。ここでは、局所的重み[42]として、関数式 2 のように tf の代わりに、gtf（gene term frequency）を用い、MEDLINE 抄録の各センテンス単位に、ある疾病（d）に共起する特定の遺伝子（g）の数（$n_d(g)$）を、（d）に関連するすべての遺伝子の数（$\sum_{i=1}^{n} n_d(g_i)$）で除したものを使用している。

$$as_d(g) = gtf_d(g) \times adf_d(g) \quad \cdots（関数式\ 2）$$

$$gtf_d(g) = \frac{n_d(g)}{\sum_{i=1}^{n} n_d(g_i)}$$

$$adf_d(g) = \log \frac{m}{ad_d(g)}$$

$$drel(d_1, d_2) = \frac{||G(d_1) \cap G(d_2)||}{||G(d_1) \cup G(d_2)||}$$

$ad_d(g)$:すべての関連疾病 di に関する drel(d,di)の合計

m:異なる疾病数

そのため、gtf が大きい遺伝子は、疾病（d）との関係が深いことになる。また、大域的重み[42]として idf の代わりには、adf（association disease frequency）を用い、ある疾病(d_1)とその他の疾病(d_i)において、d_1 と共通の遺伝子（G(d)）をもつものに関して、関連性の深さの値を関数（$ad_d(g)$）によって求め、その逆数と疾病数（m）の積に関する対数を使用し

ている。そのため、多くの疾病と関連する遺伝子ほど関数の値を小さくすることになる。これは、多くの疾病に関係している遺伝子を標的として選択すると、選択性が低く影響が多いため、副作用の観点から不利であることが理由である。その結果、$as_d(g)$（association score of g to d）は、当該疾病に関して、相対的に多く報告された遺伝子であり、かつ、多くの疾病に関係していない遺伝子の値が大きくなることになる。この adf の積をとることにより、副作用の可能性が少ない遺伝子の値を高くしている。

このように、用語の特徴量を示す関数としての TF-IDF は、多くの異なる定義の関数が提案され、本来の定義とは少し異なっているが、局所的な重みと大域的な重みの積によって用語の重要度を示すことが共通している。IDF に関しても、確率的 IDF、大域的頻度 IDF、エントロピーを使用した方法などが知られている[42]。

TF-IDF の計算には、TF 値として調査対象の文章に現れるキーワード数と、DF 値として文献データベースなどの全文書中に当該用語がどの程度含まれているかを調べる必要がある。この数値を検索エンジン（Yahoo!）の連携により自動的に取得して、重要単語のリスト作成に使用する研究[61]がある。この方法は一般的なテーマに対しては適用可能と思われるが、一般的な検索エンジンの対象である Web 情報源はブログなどを含めて雑多な情報が混在しているため、医療などの専門的なテーマにおける適用には適していないと考える。

1.3.3.4 低頻度情報の分析

文献データベースなどから得られる膨大なテキスト情報から、低頻度の有用な情報を入手するために、テキストマイニングを使用した研究が多くおこなわれている。それらの中から、ここでは研究シーズ情報として低頻度情報を入手するための先行研究に関して概観する。

菰田は、3 種類のテキストデータ（特許公報、技術論文、プレスリリース）を対象としてテキストマイニングをおこない、各分野において新規事業の企画に有用な情報へ絞り込んでいく方法を提案している[62][63]。この方法では、調査目的に関係する重要な技術用語をその分野の専門家が判断して手動による抽出をおこない、これを手がかりとしてテキストデータ中で共起する語の集合を作成し、それに対してアソシエーション分析やクラスター分析によりテキストマイニングをしている。この分析結果から、さらに重要な技術用語、とくに低頻度の用語については手動により抽出し、リストを作成することを繰り返すことで

絞り込んでいく方法を提案している。ここでは、情報分析をおこなう調査担当者と研究者や技術者などの調査主題の専門家とが協働し、人による選択とテキストマイニングの分析を繰り返すことにより、調査目的に関して共起する技術用語のクロス分析やネットワーク分析をおこない、関連する文献を評価している。これにより、調査目的に関係する重要な技術用語を抽出している。このように、適切に人の判断を取り入れることは、低頻度であるが多様な属性をもつデータの中から調査目的と背景を理解した有用性の高い用語を抽出する際には有用と考える。しかし、ここで研究事例として紹介されている分析は、数百件程度のデータセットを対象にしたものであり、数千件、数万件のデータセットを分析する場合に、初期の処理から人の判断を取り入れることは、分析における負荷が大きく、現実的な方法ではない。人が判断できる程度に調査対象の情報量を絞り込むためには、既知情報との比較や用語の文献レコード数を使用するなど、ある程度の情報の絞込みの自動化が必要である。

小池は、潜在的な知識発見支援として医学生物分野の文献のテキストマイニングによって、異なる概念に関して同一テキストには記載がなくても、共起する概念を調べることにより、新しい関連性の仮説を自動的に発見する方法を提案している[64][65]。このなかで、潜在的な知識発見に必要な検討項目として、①語彙の問題、②低頻度概念の問題、③評価の難しさについて言及している。

①の語彙の問題に関しては、多くの概念は複数の同義語をもつため、これらを考慮した処理をおこなわない場合は適切な関係性は抽出できないと指摘している。これらは自然言語を対象としたテキストマイニングの基本であり、分析対象とする概念の同義語を収集して照合処理に使用する用語辞書を適切に作成することが必要としている。とくに、本研究でも対象としている標的分子名に関係した略語や遺伝子名称に関しては語彙的な曖昧性解消が必須であると指摘している。

②の低頻度概念の問題に関しては、潜在的な知識発見のためには、低頻度の出現ではあるが重要な概念を抽出する必要があるとしている。しかし、どのような統計的手法を利用しても低頻度概念の関係性抽出は困難であるともしている。それは、一般的な統計手法は、標本数の規模がある程度の大きさをもつ場合に有効であるが、希少な事象に対して仮説を検証して抽出に適応するのは難しいからであるとしている。

③の評価の難しさについては、ある現象からそれを解決する方法、もしくはそれが解決する方法の関係性を発見する場合に、複数のパターンが見出されることがあるが、これらの中からどの関係性がより重要であるかを検証することは難しいとしている。

1.3.3.5 分析の技術としてのテキストマイニングの特徴

那須川のテキストマイニングに関する著書[66]では、テキストマイングの特徴について理解するために、膨大な量の文書データを活用する場合に最も使われる技術である「検索」、それらを分類して整理する「分類整理」、さらに知識を発見するためにテキストマイングによる「分析」の技術について比較している（表 2）。

表 2. 大量文書を活用する技術の比較

処理の深さ	処理の概要	機能（用途）	技術的要素	処理対象	自然言語処理の内容
レベル 1	検索	目を通す対象の絞り込み	情報検索	文字列 単語	単語の抽出（語の原型・基本形への置換）
レベル 2	分類整理	文書の振り分けおよび全体的にどのような内容が含まれているかの把握	クラスタリング クラシフィケーション	単語の集まり (Vector Space Model)	単語分布 状況の分析
レベル 3	分析（知識発見）	面白い内容・役に立つ知見の抽出	自然言語処理 データマイニング 視覚化	意味的概念の集まり	意味の分析 関係の分析

※本表は[66]から引用

ここでは、処理の深さがレベル 1 の「検索」の技術は欲しい文書を集めることが基本的な目的であり、レベル 2 の「分類整理」の技術は類似の文書をまとめることが基本的な目的であるのに対して、処理の深さが最も深いレベル 3 の「分析」の技術としてのテキストマイニングは、文書内容に書かれた内容の傾向や特徴を調べることを目的とし、目的が異なることと、処理単位が文章ではなく内容であるところに根本的な違いがあるとしている。

また、人工知能の研究者である松尾は、大澤との共著書[67]の中において、テキストマイニングなどの「分析」により発見した問題について、「『当たるかどうか』は解に対して聞くべき問いであり、チャンス発見が求める『問題』に対して聞くべきことではない。その問いに答える責任は、現場で未来を創る人にある」と述べている。このように、本研究の情報フィルタリングもチャンス発見の「問題」の提示に近い効果を目指しているため、その評価に

関しても、検索とは異なる視点による着想の検討が必要と考える。

1.4 本論文の構成

第 1 章を序論として、本研究の背景、目的と範囲、関連研究について述べる。

第 2 章では、研究課題 1 である本研究で提案する新奇な研究シーズ用語の抽出方法を 4 抽出段階に分けて、それぞれの段階における処理方法について述べる。

第 3 章では、研究課題 2 である本研究の提案方法の妥当性を確認するために、提案した方法により創薬探索のための研究シーズ探索の実験について述べる。

第 4 章では、実験結果に基づいて本研究において提案した方法について論じ、その妥当性、役割および可能性について述べる。

第 5 章では、本研究の結果から導かれる結論を述べる。

1.5 第 1 章のまとめ

第 1 章では、本研究の背景、研究の目的と研究対象とする範囲、関連する研究、本論文の構成について述べた。

はじめに、研究の背景として新奇な研究シーズ探索の必要性を述べた。

つぎに、研究目的として研究ニーズの実現に必要とされる適切な研究シーズを見つけるために、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法を提案し、実験により提案した方法の妥当性を確認することについて述べた。研究の範囲は、得られた研究シーズ用語の新奇性を確認するまでとした。目的を達成するために、研究課題 1 として、新奇な研究シーズ用語の候補を抽出する方法を提案すること、さらに研究課題 2 として、抽出実験により提案した方法の妥当性を確認すること、の二つを設定した。

最後に、関連する先行研究として、第一に創薬を中心とした研究シーズの探索方法、第二に創薬の研究シーズ調査に関する情報源、第三に情報抽出に関係する各種のテキストマイニングの研究（情報フィルタリングの研究、固有表現抽出に関する研究、用語重み付けに関する研究、低頻度情報の分析に関する研究、テキストマイニングの特徴に関する研究）について述べた。

第 2 章 研究シーズ用語の抽出方法の検討

研究課題 1 の研究シーズ用語の抽出方法については、専門用語辞書を利用した情報フィルタリングの方法を参考に検討した。まず、新奇な研究シーズ用語を含むデータレコードを文献データベースより収集する。つぎに、既知用語辞書との照合によりヒットしたデータレコードを削除する。残ったデータレコードを区切り文字としてストップワードで分割し、選別ルールを用いて薬理メカニズム用語や関連する標的分子用語を選別する。既知用語辞書を使用した情報フィルタリングにより抽出される文献数の少ない薬理メカニズム用語を、新奇な研究シーズ用語とみなす。

2.1 基本方針の検討

本節では、研究シーズ用語を探索するために、情報フィルタリングの考え方を用いて、データセットに含まれる大量の情報から研究シーズ用語の候補を抽出する方法を検討する。

研究シーズ用語を探索する方法の基本は、文献検索により文献を調査することであり、文献に記載された研究シーズ用語を収集することが重要である。用語の収集には人による高いコストが必要とされるため、収集方法を自動化する研究が多くおこなわれている[45]。しかし、ライフサイエンス分野の文献からテキストマイニングにより専門用語を抽出する方法は、大部分が専門用語辞書を使用する方法であるため、抽出する用語を予め指定する必要があり、さらにその同義語についても広く収集するなど語彙に関するこれらの課題を解決する必要がある[46]。そのため、新しく文献に記載され始めた用語など低頻度に出現する用語の必要性が高いにも関わらず、それらを収集して辞書を作成し、抽出することは難しい。そのため、統計的な手法を用いて抽出を自動化する研究もおこなわれているが、学習用のデータが少ないこともあり困難な面が多いとされている[65]。とくに、バイオインフォマティックスの分野では用語抽出の自動化が難しいため、専門用語を手動により文献情報から抽出してデータベース化するマニュアル・キュレーションが重要な方法として現在でも多く実施されている。

本研究の探索の目的として収集する研究シーズ用語は文献中での出現頻度が低いため、あらかじめ辞書として準備することは難しい。一方、本研究の探索の目的ではない不要な情報である既知の用語については、収集してリストを作成することは可能である。このため、開発が進んだ医薬品の薬理メカニズム用語を収集できれば、それらを集めて既知用語辞書を作成し、既知用語辞書との照合により不要な既知の情報を判別し、それらを除去する方法、すなわち情報フィルタリングの方法が使用できるのではないかと考え、検討した。はじめに、既知用語辞書と抽出対象の用語が含まれるデータセットとの照合方法について検討する。つぎに既知用語を含むとして照合されたデータレコードを削除することにより、残されたデータレコードに新奇な研究シーズ用語が含まれていることが期待できる。そのため、この方法をデータセットと既知用語辞書との照合方法として検討する。

情報フィルタリングで得られたデータレコードから研究シーズ用語を抽出する方法を検討する必要がある。データ量が多い場合に備えて、自然言語で記載されたデータレコードについては用語への分解、研究シーズ用語の選別、さらに研究シーズ用語の順位付けについて

も検討する。

新奇な研究シーズ用語の抽出方法を 4 抽出段階に分け、それぞれの段階における処理方法を明らかにする。各段階における検討内容について述べる。

抽出段階(1)では、情報フィルタリングに使用する既知用語辞書の作成について検討する。薬理メカニズム用語を収集して既知用語辞書を作成するために、薬理メカニズム用語を収録している情報源を選択し、既知として使用する用語の範囲について検討する。この抽出段階(1)については、調査の領域特有の判断が必要であり自動化は難しく、自動化のメリットも大きくないため自動化は検討しない。

抽出段階(2)では、データセットに含まれる薬理メカニズム用語と既知用語辞書との照合方法について検討し、さらに、既知の用語を含むデータレコードをデータセットから除く情報フィルタリングの方法について検討する。この抽出段階(2)については、定型的な処理が可能であり処理プログラムによる大量の情報処理など自動化のメリットが高いため、自動化を検討する。

抽出段階(3)では、データセットから薬理メカニズム用語を選別する方法について検討する。この抽出段階(3)については、選別対象のデータレコードが大量にある場合は手動では処理負担が大きいこと、さらにある程度の定型化が可能であると予測されることから部分的に処理プログラムによる自動化を検討する。

また、抽出段階(3)で多くの薬理メカニズムが選別された場合に対応するため、抽出段階(4)として、出現頻度による順位付けについて検討する。この抽出段階(4)については、データベースの検索および順位付けの計算については手動による処理は困難であり、処理プログラムによる自動化を検討する。

すなわち、第 2 章では、研究シーズ用語の抽出方法を、既知用語辞書の作成、データセットと既知用語辞書との照合、研究シーズ用語の選別、出現頻度による順位付けの 4 抽出段階に分けて部分的な自動化を検討することにした。そして、検討した結果を研究課題 1 の新奇な研究シーズ用語の候補を抽出する方法として提案する。

2.2 既知用語辞書の作成

本節では、研究シーズ用語の抽出に使用する既知用語辞書の作成過程を抽出段階(1)として、その方法を検討する。

本研究においては、創薬テーマを事例として、疾病に対する新奇な薬理メカニズム関連情報を入手するために、情報フィルタリングの機能を用いて既知の情報を除去することにした。除去する既知の情報を設定するためには、医薬品の薬理メカニズム用語として収集する既知の範囲を定義し、これらを集めた用語辞書を作成する必要がある。

既知用語辞書を作成するためには、薬理メカニズム用語を収集する情報源が必要である。薬理メカニズム用語は、疾病の原因に関係する体内タンパク質などの標的分子用語と組み合わせて、阻害、拮抗などの薬剤の作用を示す単語である作用語と組み合わせて作られる用語である。そのため、作用語を目印に用いれば、薬理メカニズム用語を収集することができることになる。しかし、医薬品の研究開発を報告する文献などに書かれているすべての既知の薬理メカニズム用語を収集するためには、手動による大量のテキストデータの処理が必要となる。疾病名と薬理メカニズム用語の情報を含むデータベースがあれば、既知の薬理メカニズム用語の情報源の候補となる。

創薬研究をおこなう製薬企業や大学などにおいて、創薬を目的に世界中で研究開発されている薬剤名とその薬理メカニズムや臨床試験などの開発段階に関する情報の必要性は非常に高い。そのため、専門のアナライザーが薬剤ごとにそれらの内容をまとめたデータベースが、医薬品の研究開発情報データベースとして複数作成されている。これらの医薬品の研究開発情報データベースを疾病名を使用して検索することにより、当該疾病の治療を目的として研究開発されている薬剤名とその薬剤が有する薬理メカニズムが得られる。そのため、医薬品の研究開発情報データベースが、既知の薬理メカニズム用語を収集するための情報源として適していると考えた。これらの医薬品の研究開発情報データベースに記載された薬理メカニズムは、創薬に関係する多くの研究者が参照する可能性が高いため、研究開発の競合を避けるために、競合が発生する可能性が高い既知の情報の範囲を決める資料としても適当であると考える。

Pharmaprojects[68]などの医薬品の研究開発情報データベースには、実際に人を対象とした臨床試験がおこなわれている治験薬や市販され医療機関で使用されている薬剤だけではなく、非臨床試験としておこなわれている細胞や動物を用いた実験段階の研究に関して

も、ある程度収集している。これらの情報は、基本的に文献、新聞、企業から直接公開されるニュースリリースや Web サイト、株主向けのアニュアルレポートなど、公開された情報から手動により収集している。収集した情報のうち、医薬品として研究開発が進められる可能性が高い情報を専門のアナライザーが判断してデータベースに収録している。そのため、基礎医学的な段階の研究開発が実施されるかどうか判断できない新しい薬理メカニズムを有する薬剤についての情報は収録されない場合が多い。薬理メカニズム用語の収集に使用する医薬品の研究開発情報データベースとして、世界で標準的に使用されているのは Pharmaprojects[68]や Integrity[69]などである。これらのデータベースを使用して収集できる薬理メカニズム名とその同義語の例を表 3 に示す。

表 3. 薬理メカニズム用語とその同義語

薬理メカニズム用語	同義語
epidermal growth factor receptor inhibitor	EGFR inhibitor ERBB inhibitor ERBB1 inhibitor HER1 inhibitor mENA inhibitor NISBD2 inhibitor PIG61 inhibitor avian erythroblastic leukemia viral oncogene homolog inhibitor cell growth inhibiting protein 40 inhibitor cell proliferation inducing protein 61 inhibitor epidermal growth factor receptor tyrosine kinase domain inhibitor erb-b2 receptor tyrosine kinase 1 inhibitor proto-oncogene c-ErbB-1 inhibitor receptor tyrosine-protein kinase erbB-1 inhibitor EGFRK inhibitor epidermal growth factor receptor kinase inhibitor

これらの情報源を疾病名により検索して、得られたデータに含まれる薬理メカニズム用語とその同義語の情報を参照して、既知用語リストである既知用語辞書を作成することにした。方法は、全疾病に関する既知用語辞書を作成する場合は医薬品の研究開発情報データベースのシソーラス情報を参照して、薬理メカニズム用語とその同義語情報のリストを作成する。得られた同義語をその代表的な薬理メカニズム用語と組み合わせて同義語数分のデータを作成し、既知用語辞書とする。一方、特定の疾病に関する既知用語辞書を作成する場合は、疾病名で医薬品の研究開発情報データベースを検索し、得られた結果から薬理メカニズム用語を含む情報を出力する。つぎに、得られた薬理メカニズム用語を全疾病の既知用語辞書に含まれる薬理メカニズム用語と照合して、該当する同義語のデータを収集することにより既知用語辞書を作成する。これらの処理は定型化が難しく、内容を確認

して進める必要があるため Excel とその関数を使用する方法が適していると判断し、自動化については検討しない。

2.3 データセットと既知用語辞書との照合

本節では、情報フィルタリングの核心となる機能であるデータセットと既知用語辞書との照合方法とヒットしたデータレコードを除く方法を抽出段階(2)として、その方法を検討する。この段階は、定型的な処理が可能であり、大量の情報を処理するためには処理プログラムによる自動化のメリットが大きいため、自動化の検討をおこなう。

照合方法については、照合対象のデータセットに対して既知用語辞書に含まれる全用語をデータレコードの先頭から順番に照合を繰り返していくことにより、データセットの全データレコードとの照合をおこなう方法が考えられる（図 1)。この方法では、同じデータレコードの情報要素に複数の既知用語辞書の用語が重複してヒットする場合もあるが、本研究では、ヒットしないデータレコードの情報が必要であるため支障はないと考える。

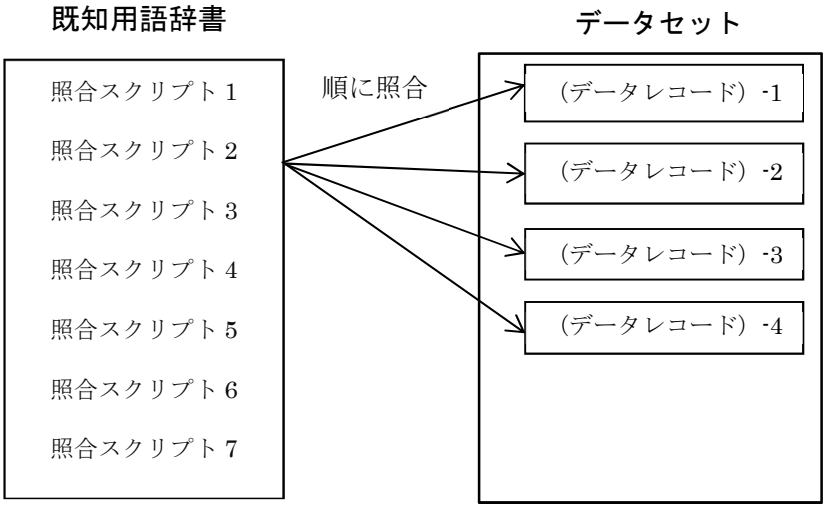

図 1. データセットと既知用語辞書との照合方法

既知用語辞書に収載されている既知用語が単語の場合は、その単語がデータレコードに含まれる場合をヒットとする。しかし、既知用語辞書に収載されている既知用語は複数の単語から構成される複合語の形式の場合が多い。複合語の場合は、構成する単語がデータセットにすべて含まれる場合にヒットとする論理積による照合方法 I を使用することが考えられる。この方法は、データレコード中の薬理メカニズム用語を構成する単語との間に形容詞などその他の単語が複数入った場合や、構成する単語順が倒置して使用されているような

場合にもヒットすることになる。このように、論理積による照合方法Ⅰは、ヒットの再現性は高いことが期待できるが、その反面、誤ヒットの可能性も高くなると考えられる。

一方、照合の精度を重視したい場合に使用できる近接演算子の機能を有する照合方法Ⅱについても検討する。この照合方法Ⅱは、複合語の薬理メカニズムを構成する単語が順不同で、各単語間に指定した語数以下のその他の単語を含んでもよい条件などを設定できる。これは、薬理メカニズムを構成する単語の並びには倒置などを含めて多様性があり、単語間に不規則に記号や単語が入る場合が多いことに対応して照合するためである。複合語の各単語間に許容する単語数は、複合語を構成する単語数により適切な値が変化することも考えられるため、複合語の構成単語数によって許容する単語数が自動変動する機能も検討する。精度を重視して 2 単語に固定した設定で近接演算子の関数を使用した場合の、照合処理によってヒットする例を図 2 に示す。

ここでは、4 単語に near2 の近接演算子の関数を使用し、機能としては 4 単語すべてが順不同で含まれ、同時に単語間に 2 個以下の指定外の単語が含まれるパターンに対してはヒットすることになる。

＜近接演算子の機能を有する関数で作成した near2 の場合の機能＞

(near2) A B C D

→各単語（A、B、C、D）間に 2 個以内の他の単語が入ってもよい
→語順は問わない

ヒット	照合対象とする複合語
○	A B C D
○	A C B D
○	D **E** B **F** C A
×	A B C A
×	A B **E** **F** **G** C D

図 2. 近接演算子の機能の例

つぎに、既知用語辞書でヒットしたデータレコードを除くには、照合によりヒットしたデータレコードの識別番号をファイルとして保存し、全部の照合が終了した後に、照合でヒットしたデータレコードの識別番号を目印に当初のデータセットから除く方法が考えられる。これにより、既知用語を持たないデータレコードの集合体であるデータセットが作成できる。抽出段階(2)のデータセットと既知用語辞書との照合方法とヒットしたデータレコードを除く方法については、perl 言語によるプログラミングをおこなうことにより、短時間に処理が可能になるための自動化をおこなう。

2.4 研究シーズ用語の選別

既知用語を含まない自然言語で記載されたデータレコードの索引から新奇性が期待される研究シーズ用語を選別する過程については、抽出段階(3)として、その方法を検討する。この抽出段階(3)については、索引からの用語選別について大量の情報の処理が必要な場合に対応できるように、部分的に処理プログラムによる自動化を検討する。

既知用語辞書による情報フィルタリングによって得られた新奇な研究シーズ用語が含まれる可能性が高いデータレコードの索引には、薬理メカニズム用語や薬理メカニズム用語に含まれる標的分子用語の単名詞や複合名詞が含まれている。とくに、データレコードに自然言語により記述された索引が含まれている場合は、その他の主題の専門用語や一般用語も含まれている可能性が高い。そのため、薬理メカニズム用語のみを区別して効率よく選別する方法が必要である。

薬理メカニズム用語を再現性よく選別するためには、この分野の専門用語の知識を有する専門家が判断して手動により選別する方法が確実である。このため、情報フィルタリングによる絞り込みの結果として、選別対象の情報が新奇な薬理メカニズム用語を含む索引に絞り込まれ、処理件数も少なくなった場合は、作用語などを目印として、手動により薬理メカニズム用語を選別する方法を使用することができる。しかし、情報フィルタリングによって文献検索の結果を目的とする新奇な薬理メカニズム用語を含む情報に絞り込んだ結果、絞り込みの効果が十分ではなく数百件から数千件の結果が得られる場合も考えられる。このような場合は情報の過負荷となるため、手動による適切な選別が困難となることが予測される。また、処理する情報量が少ない場合でも、この分野の専門用語の知識が少ない人の場合は、手動による選別は難しいことが予想される。

自然言語の文章からテキストマイニングによって専門用語を自動選別する際に、用語に関する情報が収載された辞書が用意できる場合は比較的問題は少ない。しかし、本研究が選別する用語は新奇の研究シーズ用語であり、それを含む辞書を用意することはできない。専門用語などの複合語をテキストから収集する技術は、探索対象とする専門領域の分野固有の知識に依存する部分が多く、当該分野の知識が必要であるとされている。本研究では、創薬テーマを企画するために必要な新奇な薬理メカニズム用語や標的分子用語を選別することが目的である。そのため、薬理メカニズム用語や標的分子用語の形態的な特徴を利用した選別ルールを用いた方法について検討した。

本研究では、当初の研究において薬理メカニズム用語の形態的な特徴を利用した正規表現によるパターンマッチングによる選別についても検討した。すなわち、自然言語で書かれたテキストからスクリプトを使用して直接、薬理メカニズムを選別する方法である。しかし、正規表現のみを使用して自然言語中の薬理メカニズムの表記の多様性をカバーすることは困難であることがわかった。それは、スクリプトのパターンに使用した作用語と標的分子用語が異なる文脈で使用される場合もあり、標的分子用語を再現性よく選別することができなかった。先行研究においても、正規表現による機械的な薬理メカニズム用語の収集[36][70]がおこなわれているが、同様に低い再現性に課題があると指摘している。

このように、薬理メカニズム用語をパターンマッチングにより選別することは困難であるため、それ以外の方法を検討する。すなわち、自然言語で書かれたテキストに含まれるストップワードを取り除くことにより、薬理メカニズム用語を選別するための単名詞、複合名詞のリストを作成する方法である。はじめに、代表的な前置詞や冠詞などをストップワードとして、テキスト文を区切る自動処理の方法を作成する。つぎに、分離されたすべての単名詞と複合名詞などを集計して用語リストを作成し、その中から薬理メカニズム用語および薬理メカニズム用語に含まれる標的分子用語をルールにより選択する。このストップワードによる分割により、複数の単語数からなる複合名詞も選別が可能になると考えた。

ストップワードを使用して用語へ分解するためには、不要な分割を防止するための前処理が必要になる。データレコードに含まれる阻害剤の薬理メカニズムは、“XXX inhibitor”と“inhibitor of XXX”の記載形式が多い。“inhibitor of XXX”のような倒置表現の場合にストップワードとして使用する“of”によって分割されることを防ぐためには、“inhibitor-of XXX”のようにストップワードの前にスクリプトによる一括処理でハイフンを入れることにより“of”の前のスペースをなくし、分割を防止するための自動処理をおこなう。

ストップワードにより分解して得られた単語と複合語からなる用語リストの中から薬理メカニズム用語や関連する標的分子用語を選別するためにつぎに示す選別ルールを考えた。

(1) 薬理メカニズム用語を構成する作用語を含むルール

薬理メカニズム用語に含まれる用語として、薬剤の標的分子への作用を示す作用語がある。そのため、薬理メカニズムに共起する作用語を含む用語は、薬理メカニズムの可能性が最も高いため、inhibitor、antagonist または agonist を含む用語を選別する。

(2) 標的分子関連語を含むルール

作用語を含まない標的分子用語やその略語などがストップワードによる分割の結果とし

て選別される場合がある。これらを含めて抽出する場合は、薬理メカニズムに含まれる標的分子に関連する用語として target などを含む用語について選別する。また、標的分子関連語を選別する方法として大文字を含む用語を選別する方法がある。この大文字を含むルールに関しては、標的分子名が大文字を含む略語で記載されることが多いとの研究[71]があり、使用できると考えた。

これらのルールにより得られた研究シーズ用語の表記のゆれを正規化するための処理は定型的な対応が難しいことから自動化は難しい。そのため手動によるデータ・クリーニングを考えた。すなわち、ピリオド、ハイフンなどの記号、動詞、形容詞などを手動により除去した後に、得られた用語をアルファベット順ソートし、同じ語幹の用語を名寄して頻度集計をおこない、選別された研究シーズ用語のリストを作成することにした。

2.5 研究シーズ用語の順位付け

選別により得られた研究シーズ用語が多い場合には、新奇な用語に絞り切れていない場合がある。それらには、薬理メカニズムと関係のない専門用語や一般用語、既知の薬理メカニズム用語も含まれていると考えられる。そのため、これらの用語が使用されている文献の数により順位付けをすることができれば、大量の情報の処理が可能になる。そのため、研究シーズ用語の順位付けの過程として、抽出段階(4)を設定し、その方法について検討する。この抽出段階(4)については、データベースの検索および順位付けの計算の自動化を検討する。

用語の順位付けスコアを計算する方法について検討した結果、TF-IDF[56]を利用することにした。TF-IDF は、対象とする文書内における用語の出現頻度と、文献データベース全体における当該用語が使用された文献レコード数を用いて、用語の重み付けをおこなう関数である。この関数は、情報検索における索引付けやテキストマイニングの分野でも用いられている。このように、用語の特徴や重要性を表す尺度を与えることは用語の重み付けと呼ばれる。本研究では、特定の疾病に関して文献による使用頻度の少ない薬理メカニズム用語の抽出を目的とするため、TF-IDF が使用できると考えた。

すなわち、ルールよって選別した研究シーズ用語のリストをプログラムによる自動処理により得られた TF-IDF 値を用いて降順にソートし、調査目的に応じて適当な位置で区切り、上位の用語に限定することにより、研究シーズ用語の削減が可能になると考えた。

本研究においては、TF-IDF の定義が通常の文書を対象とした場合と異なるため、適切な式への修正が必要と考えた。tf は局所的重み（local weight）とされ、特定の文書における当該用語の出現頻度に基づき計算される重みである[42]。しかし、本研究で使用する tf は、文献検索で得られた 1 年分の文献に含まれる複数の IT 索引中における用語数である。これは、文献検索で得られた 1 年分の文献レコードを索引単位で分解して作成したデータレコードを使用しているためである。そのため、基となる情報が 1 年分で大きいことと同じ索引語が複数付与されることもあるため、通常の短い文書に含まれる用語頻度である tf 値より値が大きくなる可能性がある。そのため、tf と異なる stf（seeds term frequency）を定義し、値の差が大きい場合は、その差を少なくすることを目的に加算値として a を加えることを検討することにした（関数式 3)。

$$\text{STF-IDF} = (\text{stf} + \text{a}) \times \left(\log\left(\frac{\text{n}}{\text{df}}\right) + 1\right) \quad \cdots \text{（関数式 3）}$$

stf：指定期間（1 年など）における当該用語の出現頻度

a：stf 値を調整するための値

df：文献データベース全体における当該用語が使用されている文献レコード数

n：文献データベースの全文献レコード数

IDF に関しては、文書集合全体における当該用語の分布を考慮して決定される重みとしての df に相当する値を定義する。

以上のように、出現頻度による順位付けにより研究シーズ用語を削減するために、文献データベースの当該用語の文献レコード数取得について自動化し、さらに、その値を使用して STF-IDF の値を自動計算するための方法を検討する。

2.6 研究シーズ用語の新奇性の確認

提案した方法を用いた情報フィルタリングにより、疾病による文献検索で得られた文献レコードから新奇な研究シーズ用語の候補が抽出できると考えられる。ここで得られた研究シーズ用語の新奇性を確認する方法を検討する。

抽出した研究シーズ用語を設定した疾病名とともに文献データベースの検索をおこなうことにより、その文献レコード数から、新奇性を推測できると考えた。すなわち、文献レコード数が少ない場合は、抽出した研究シーズ用語と設定した疾病の組み合わせにおいて、研究された報告が少ないことから、新奇な研究シーズ用語であると考えられる。

2.7 第 2 章のまとめ

第 2 章では、研究課題 1 である、情報フィルタリングによる研究シーズ用語の抽出の方法を検討した。検討結果を表 4 に示す。

表 4. 研究シーズ用語の抽出方法

段階		過程の内容	検討結果
抽出段階	(1)	情報フィルタリングに使用する既知用語辞書の作成	既知用語辞書は、医薬品の研究開発情報データベースを疾病名を検索語として検索し、検索結果から薬理メカニズム用語を収集して作成する。処理は Excel を使用し、自動化はおこなわない。
	(2)	データセットと既知用語辞書との照合	照合方法は、論理積による方法および近接演算子による方法を作成する。照合されたデータレコードをデータセットから除去し、新奇な研究シーズ用語を含むデータレコードを絞り込む。処理は、定型的な処理が可能であり処理プログラムによる自動化のメリットが高いため、自動化をおこなう。
	(3)	研究シーズ用語の選別	情報フィルタリングにより絞り込まれたデータレコードから、選別ルールを用いて手動により研究シーズ用語を選別し、データ・クリーニングをおこなう。データレコードが多い場合は、ストップワードにより用語に分解して得たデータから、選別ルールを用いて関数および手動により研究シーズ用語を選別する。処理は、データレコードが多い場合は、索引からの研究シーズ用語選別は手動では処理負担が大きいこと、さらにある程度の定型化が可能であることから部分的に処理プログラムによる自動化をおこなう。
	(4)	研究シーズ用語の順位付け	選別された研究シーズ用語が多い場合は、順位付けのスコアを計算することにより、順位付けされた研究シーズ用語から選択できるようにする。処理は、データベースの検索および順位付けの計算については、処理プログラムによる自動化をおこなう。

抽出段階(1)として、情報フィルタリングに使用する既知用語辞書の作成について検討した。その結果、情報源として医薬品の研究開発情報データベースが適していると判断し、疾病名で検索した結果から薬理メカニズム用語を収集して既知用語辞書を作成することにした。医薬品の研究開発情報データベースとしては、Pharmaprojects、Integrity、明日の新薬を候補とした。処理は Excel を使用し、自動化はおこなわないことにした。

抽出段階(2)として、既知用語辞書を使用してデータセットに含まれる薬理メカニズム用語を照合する方法について検討した。その結果、データセットと既知用語辞書との照合は、論理積もしくは近接演算子による方法を使用することにした。また、研究シーズ用語が含ま

れたデータレコードを絞り込むために、照合でヒットしたデータレコードをデータセットから除く方法を使用することにした。処理は、定型的な処理が可能であり処理プログラムによる自動化のメリットが高いため、自動化をおこなうことにした。

抽出段階(3)として、データレコードの索引から研究シーズ用語を選別する方法について検討した。薬理メカニズム用語を再現性よく選別するためには、この分野の知識を有する専門家による選別が適しているため、索引数が絞り込まれた場合は、作用語を含むなどの選別ルールを用いて手動により薬理メカニズム用語を選別することにした。索引数が多い場合は手動による選別が難しいため、ストップワードを区切りとして得られた用語を複数のルールによって選別し、さらに、名寄せなどのデータ・クリーニング後に頻度集計したリストを作成することにした。処理は、データレコードが多い場合は、索引からの研究シーズ用語選別は手動では処理負担が大きいこと、さらにある程度の定型化が可能であることから部分的に処理プログラムによる自動化をおこなうことにした。

抽出段階(4)として、選別した用語が多い場合に対応するために、選別で得られた研究シーズ用語を文献レコード数を使用した関数により順位付けする方法を検討し、使用することにした。処理は、データベースの検索および順位付けの計算について処理プログラムによる自動化をおこなうことにした。

提案した方法により抽出された研究シーズ用語の新奇性は、設定した疾病名と研究シーズ用語を使用して文献データベースを検索することにより確認する。すなわち、文献レコード数が少ない場合は研究された報告が少ないことから、新奇性がある研究シーズ用語であると考える。

第 3 章 研究シーズ用語の抽出実験

研究課題 2 については、breast cancer と lung cancer の治療薬開発に必要な研究シーズを対象として、新奇な研究シーズ用語の抽出実験をおこなう。得られた研究シーズ用語の新奇性を確認することにより、提案する抽出方法の妥当性を確認する。抽出方法を検討する際には、新奇な研究シーズ用語を抽出するデータはデータセットとし、その作成方法については検討に含めなかった。抽出方法は汎用性がある方法として検討した。しかし、データセットの作成方法は情報源として使用する文献データベースによって調整が必要となるため、実際に実験をおこなう際に検討することにした。そのため、実験に先立ち、準備段階としてデータセットの作成方法について検討する。

3.1 準備段階

情報フィルタリングによる研究シーズ用語の抽出実験の準備段階を、(1)研究ニーズの選定と抽出に使用する情報源の選択、(2)情報検索による文献レコードの収集、(3)文献レコードを用いたデータセットの作成の 3 段階で考える。

3.1.1 研究ニーズの選定と情報源の選択

本研究では、研究シーズの対象を創薬研究に必要な、疾病に関係する薬理メカニズムとした。そこで、対象とする研究ニーズとして breast cancer と lung cancer の二つの疾病を定めた。

breast cancer は、患者のがん細胞の遺伝子やタンパク質によって種々の病気のタイプに分類され、近年、タイプによっては治療成績が劇的に向上する薬剤が開発されている。このため、がん細胞に関係する遺伝子やタンパク質などの新しい標的分子に基づく、さらなる治療薬の開発が期待されている[72]。

lung cancer は、日本人の死亡原因として最も多い疾病である。近年、肺がん細胞の増殖に関係する多くの標的分子が明らかになり、それに作用する抗体や低分子の治療薬の研究が進んでいるため[73][74]、医薬品として承認され、臨床においてすぐれた効果を発揮する薬剤も知られている。より安全で高い効果が得られる新しい治療薬への社会からの期待も大きく、大学や製薬企業の研究者によって基礎研究から臨床研究まで活発におこなわれている[75]。肺がんの治療薬については、高い研究ニーズがあるため大量の文献が報告されている。そのため、すべての文献を調査することは困難な面があり、文献を対象としたテキストマイングによる標的探索の研究の進展が期待されている[76][77]。

つぎに、疾病に関係する研究シーズの探索に適した情報源を選択する。必要な情報を収集するためには探索主題に関する情報を多く含む情報源を選択することが重要である。文献情報からの情報収集には情報源として文献データベースが適しているが、文献データベースにより、対象とする学術雑誌の範囲、文献の収録基準、付与されている統制語と非統制語などが異なっているため、これらを確認して文献データベースを選択する必要がある。本研究では、疾病に関係する標的分子の研究、とくに研究初期段階の薬理メカニズムについての情報を入手することを探索主題とした。新しく医薬品を作るための研究である創薬研究で

は、標的分子への薬剤の作用である薬理メカニズムの研究情報が重要となる。これらの研究成果は、医学、薬学、生化学などの分野の学術雑誌に掲載される。そのため、それらの領域の文献を収録するライフサイエンス分野の文献データベースを比較検討し、適する情報源を選択した。

3.1.1.1 医学、薬学分野の文献データベースの比較

医学、薬学文献を調査する際には、それらに関係する文献を多く収録するライフサイエンス分野の文献データベースである MEDLINE、EMBASE、BIOSIS が使用されることが多い[78]。そのほかにも、文献データベースとして医薬品の研究開発に関係する文献を中心に収録する DDF、非臨床試験として基礎医学や生化学の文献を多く収録する CA がある。これらの 5 データベースが医学、薬学文献の収録数が多く、信頼性の高い文献データベースとして認識されているため、情報源の候補とした。これらのデータベースを比較した結果を表 5 に示す（この結果は、以前に報告[78]したものを、2018 年 3 月に再調査して修正したものである）。

表 5. 医学･薬学分野の文献データベースの比較

	BIOSIS	CA	DDF	EMBASE	MEDLINE
収録情報	生物、生物医学分野の広範囲な文献情報	化学、生化学、化学工学、医学の非臨床分野を中心とした文献情報、特許情報	医薬品の合成、開発、評価、製造、使用などの文献情報、会議録情報	生物医学、薬学領域の文献情報	生物医学、薬学、歯科学などの幅広い文献情報
収録開始期間	1926 年-	1907 年-	1983 年-	1947 年-	1946 年-
収録件数	1980 万件	3170 万件	130 万件	1310 万件	1840 万件
作成機関	Clarivate Analytics	CAS	Clarivate Analytics	Elsevier	National Library of Medicine
特徴	会議資料も多数収録している。概念コードで研究分野を限定できる。生物系統分類コードを使用できる。	化学構造やタンパク配列から検索できる。物質ごとに詳細な索引を付与している。	医薬品研究のための明確な選択方針に従って文献を収録している。薬剤毎に統制語を関連付けて付与し、第三者抄録を付与している。	医薬品の統制語が充実している。MEDLINE に比べて欧州の文献を多く収録している。	看護学、栄養学、獣医学などの文献も収録している。MeSH というシソーラスを持ち、最新の統制語で過去に遡った検索ができる。

これらの 5 データベースの収録情報を比較すると、BIOSIS、CA、EMBASE、MEDLINE の 4 データベースは、本研究の探索主題に関係の深い基礎医学文献も収録対象となっていた。一方、DDF はある程度研究が進んだ医薬品の研究開発に関する文献収録が中心であり、基礎的な研究の文献の収録は少ない。さらに、DDF は年間の収録文献数が約 5 万件と少ない。そのため、DDF は本研究の探索主題である研究初期の標的分子やその薬理メカニズムに関する情報は少ないと考え、それ以外の 4 データベースを情報源の候補として比較検討した。

ライフサイエンスに関する研究情報を解析して新しい知見を得る学問や技術はバイオインフォマティクスと呼ばれ、情報源としては MEDLINE の情報を含む PubMED が使用されることが多い[78]。この理由としては

・PubMED は世界の主要なライフサイエンス分野の学術雑誌情報を収録している、

・索引が付与されていない最新の情報も収録している、

・使用料金が不要であるため情報収集のコストが低い、

・規定遵守したダウンロードによりデータを使用する場合は、作成者である米国国立医学図書館（NLM）に許諾が不要である、

などが考えられる。その他のライフサイエンス分野の文献データベースである BIOSIS や EMBASE は、検索と出力のための使用料金が高額であり、さらに検索結果をテキストマイニングなどの分析に使用する場合は、データ使用に関する契約と使用料金が必要となる。そのため、BIOSIS や EMBASE のデータを使用した情報分析の研究に関する報告は少ない。このような理由もあり、バイオインフォマティクスなどのライフサイエンスに関する文献情報を分析する情報源として PubMED が使用される傾向がある。

さらにこれらの文献データベースの索引を比較すると、BIOSIS、EMBASE、MEDLINE の索引語は、統制語が中心であり、それ以外の索引語として、化学物質名や著者キーワードなどが付与されている。このなかで BIOSIS は、生物の分類を表す統制語が付与されていることに特徴があるが、他のデータベースと比較して本研究の薬理メカニズムに関する索引語の付与数は多くはない[79]。そのため、研究初期の薬理メカニズムに関する用語を抽出する目的には十分ではないと考えられることから、BIOSIS は本研究の情報源として適切ではないと判断し、候補から除外する。そのため、残りの CA、EMBASE、MEDLINE の 3 データベースについてさらに内容を検討した。

3.1.1.2 薬理メカニズム収録件数の比較

情報源の候補として絞られた CA、EMBASE、MEDLINE の 3 データベースについて、2016 年発行の収録文献を STN 情報検索システム（以下、STN システムと略す）[80]で検索（2018 年 5 月 8 日）して、文献レコード数を確認した。結果を図 3 に示す。

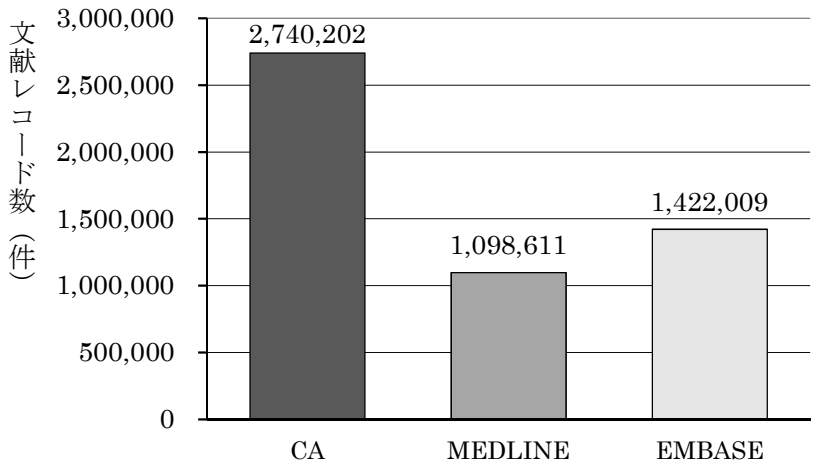

図 3. 文献レコード数の比較

年間収録件数を比較すると、CA は約 274 万件と最も多く、2 番目は約 142 万件の EMBASE、3 番目は約 110 万件の MEDLINE であった。データベースにより収録対象の情報源が異なるため、単純に収録件数のみでは比較できないが、3 データベースのなかで CA が最も大規模な文献データベースであることがわかった。

つぎに、CA、EMBASE、MEDLINE における、標的分子への薬剤の作用を示す薬理メカニズム用語が使用されている文献レコード数を比較した。本研究では研究初期の薬理メカニズムを調査の対象としたために、研究初期の非臨床試験の薬剤しか報告されていない薬理メカニズム用語を調査して選択した。調査には Pharmaprojects データベース[68]を使用し、人での試験がおこなわれていない非臨床試験段階の薬剤 3,867 件の中から薬理メカニズム 351 種を収集し、市販、臨床試験、中止などの開発段階にある薬剤の薬理メカニズムを除き、さらに CA を予備検索して件数が非常に多いものと件数の少ないものを除いて 30 種を得た。つぎに、これらの薬理メカニズム用語を検索語として 3 データベースを検索し、その文献レコード数を比較した。フレーズ検索式（例：検索式 1）と近接演算子“(6A)”を用いた検索式（例：検索式 2）で検索した結果の文献レコード数を表 6 に示す[33]。近接演算子“(6A)”は、検索に使用した複合語を構成する単語の語順に関係なく、単語間に 6 語までの他の単語が含まれても検索する近接演算子である。

S “THROMBOXANE A 2 ANTAGONIST?”　・・・（検索式 1）

S THROMBOXANE(6A)A(6A)2(6A)ANTAGONIST?　・・・（検索式 2）

表 6. 非臨床試験のみの薬剤に関わる薬理メカニズム用語の文献レコード数の比較

No.	薬理メカニズム用語	A:フレーズ検索			B:近接演算子(6A)検索		
		MED	EM	CA	MED	EM	CA
1	THROMBOXANE A 2 ANTAGONIST?	103	107	375	837	815	1,182
2	PHOSPHODIESTERASE V INHIBITOR?	22	960	173	123	1,063	401
3	PROSTAGLANDIN AGONIST?	16	17	64	752	732	956
4	SODIUM ANTAGONIST?	8	11	17	584	601	978
5	HISTAMINE RELEASE INHIBITOR?	26	136	100	483	590	796
6	POTASSIUM ANTAGONIST?	5	5	28	407	435	786
7	GASTRIN INHIBITOR?	9	11	55	434	439	582
8	SELECTIN ANTAGONIST?	56	136	126	98	175	218
9	TESTOSTERONE AGONIST?	1	1	9	201	189	349
10	ENDOTHELIN B RECEPTOR AGONIST?	25	208	24	98	295	72
11	GLUTAMATE 3 RECEPTOR ANTAGONIST?	0	0	0	155	159	191
12	INTERLEUKIN 10 ANTAGONIST?	0	1	11	124	124	206
13	FACTOR XIA INHIBITOR?	5	4	31	66	59	111
14	INTERFERON AGONIST?	0	0	11	38	34	176
15	TOLL LIKE RECEPTOR ANTAGONIST?	2	7	7	17	26	46
16	UROTENSIN II RECEPTOR AGONIST?	2	6	4	8	9	15
17	PHOSPHOFRUCTOKINASE 2 INHIBITOR?	0	0	1	12	6	14
18	PANTOTHENATE KINASE INHIBITOR?	1	2	5	4	2	9
19	SERINE RACEMASE INHIBITOR?	2	2	2	4	3	8
20	BMX TYROSINE KINASE INHIBITOR?	0	0	4	0	0	11
21	LEUKOTRIENE D 4 ANTAGONIST?	1	2	2	1	5	4
22	TIE 1TYROSINE KINASE INHIBITOR?	0	0	0	0	0	15
23	PROLACTIN RELEASE STIMULANT?	0	0	0	5	5	3
24	EPHRIN B 2 INHIBITOR?	0	0	0	3	3	6
25	FIBROBLAST GROWTH FACTOR RECEPTOR 2 ANTAGONIST?	0	0	0	1	3	8
26	HYALURONAN SYNTHASE INHIBITOR?	1	1	2	2	1	5
27	HYALURONIC ACID AGONIST?	0	0	0	1	1	10
28	INTERLEUKIN 22 RECEPTOR ANTAGONIST?	0	0	0	2	2	7
29	INTERLEUKIN 31 ANTAGONIST?	0	0	0	4	3	4
30	PROTEASE ACTIVATED RECEPTOR 2 ANTAGONIST?	1	0	2	2	1	5

フレーズ検索式を用いると検索に使用した複合語と完全一致の条件で検索され、近接演算子“(6A)”を用いた検索式では、検索語間に 6 単語以内の単語の存在が可能となり単語の配置や距離の変化に対応して検索される。3 種のデータベースで検索された文献レコード数のなかで最も多い文献レコード数の背景を赤で示した。フレーズ検索では、CA が 17 個（57%）、EMBASE が 8 個（27%）、MEDLINE が 1 個（3%）であった。EMBASE では、5 個の薬理メカニズムについては CA よりも多くの文献が検索されていた。一方、近接演算子検索に関しては CA が 25 個（83%）の薬理メカニズム用語で最も多い文献レコード数であった。EMBASE は 3 個のみで文献レコード数が最も多かった。近接演算子を使用すると、検索される文献レコード数は増加し、フレーズ検索と比較した増加率は、MEDLINE は平均 29.0 倍、EMBASE は平均 26.5 倍、CA は平均 16.3 倍であった。文献中での薬理メカニズムの記載は、..inhibitor などの記述が inhibitor of..などと倒置されたり、途中の語句の順番が前後したりすることも多く、近接演算子の使用が有効であったと考えられる。非臨床試験のみ

の薬剤に関わる薬理メカニズム用語の文献レコード数の比較結果によると、CA は、研究初期の薬理メカニズムについては、他のデータベースよりも文献レコード数が多い傾向にあることがわかったため、情報源の候補とすることにした。

しかし MEDLINE は医学分野の文献検索によく使用されているため、念のため特定の疾病に関係する具体的な薬理メカニズム用語を含む文献レコード数を調査し、CA と MEDLINE とを比較した。疾病の breast cancer を治療対象とした初期の臨床試験段階にある薬剤の Roscovitine を選択し、その薬剤が有する薬理メカニズム用語である Cyclin dependent kinase 2 inhibitor について、MEDLINE および CA を検索（2014 年 11 月 11 日検索）し、雑誌の発行年代別に文献レコード数を比較した結果を図 4 に示す。

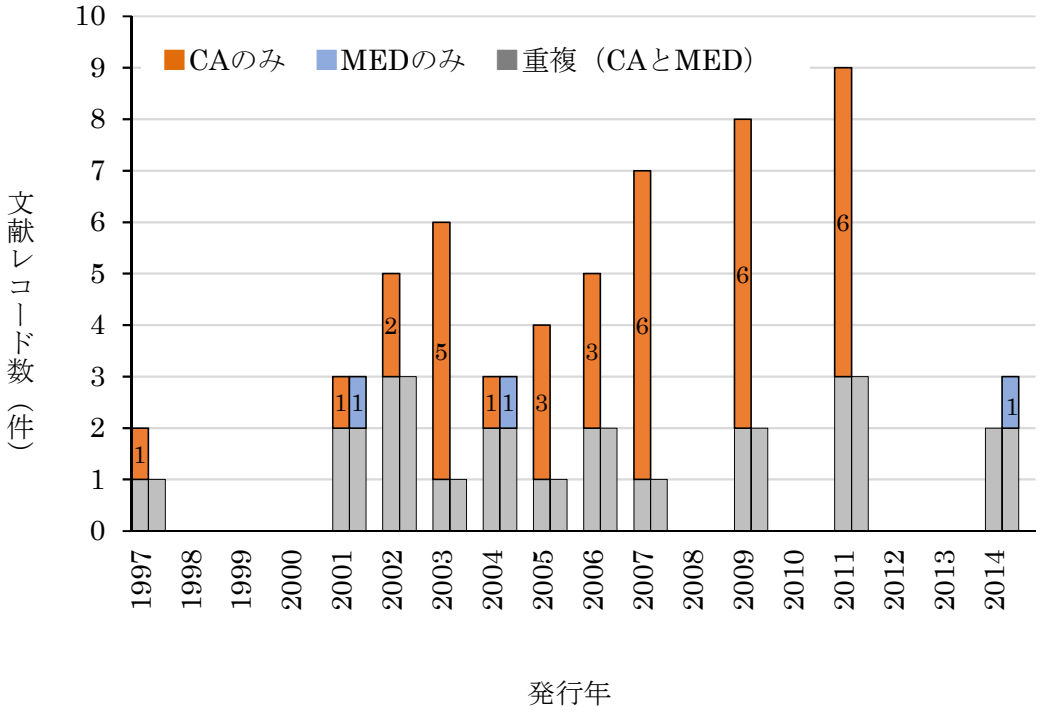

図 4. 研究初期の薬理メカニズム用語を含む文献レコード数の比較

図 4 に示すように、青色で示した MEDLINE のみで検索された文献レコード数より、オレンジ色で示した CA のみで検索された文献レコード数の方が多く、MEDLINE のみで検索された文献レコード数は 2001 年、2004 年、2014 年の計 3 件のみであった。医薬品の研究開発期間は、最初に研究テーマが設定されてから 10 年前後が必要であり、MEDLINE は、最後の方で実施される臨床試験に関する文献が多いことが理由として考えられる。このよ

うに、収録分野、索引方法、収録データ数を比較することにより、研究初期の薬理メカニズムに関する情報源として CA が適していると判断した。

3.1.1.3 CA の特徴

CA を情報源にするにあたり、収録される文献の専門分野について確認した。CA に収録される文献は、その内容により図 5 に示す大分類に分けられている。CA の 2016 年発行の文献数を、大分類コードを使用した STN システムで検索（2018 年 3 月 13 日）により調査した。結果を図 5 に示す。

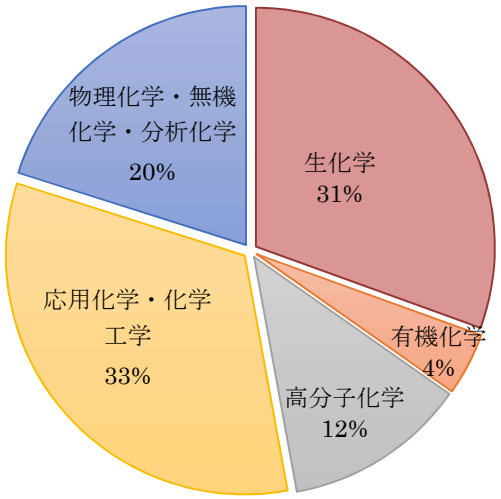

図 5. CA の収録分野

CA は、ライフサイエンス分野として重要な生化学以外にも有機化学、応用化学など多くの創薬に関係の深い分野を収録対象としている。これらのなかで医学、薬学に関係する情報として重要な生化学の分野の文献レコード数が全体の 31%と多く占めていることがわかった。さらに、CA に収録される文献はその内容により 80 のセクション（分野）に分類され、生化学分野は表 7 に示すように、20 のセクション[81]から構成されている。

表 7. 生化学分野のセクション

セクションの名称
1. Pharmacology （薬理学）
2. Mammalian Hormones （哺乳動物ホルモン）
3. Biochemical Genetics （生化学的遺伝学）
4. Toxicology （毒物学）
5. Agrochemical Bioregulators （農芸化学的生物調整剤）
6. General Biochemistry （生化学一般）
7. Enzymes （酵素）
8. Radiation Biochemistry （放射線生化学）
9. Biochemical Methods （生化学の方法）
10. Microbial, Algal, and Fungal Biochemistry （微生物生化学）
11. Plant Biochemistry （植物生化学）
12. Nonmammalian Biochemistry （非哺乳類生化学）
13. Mammalian Biochemistry （哺乳類生化学）
14. Mammalian Pathological Biochemistry （哺乳類病理生化学）
15. Immunochemistry （免疫化学）
16. Fermentation and Bioindustrial Biochemistry （発酵，工業生物化学）
17. Food and Feed Chemistry （食品，飼料化学）
18. Animal Nutrition （動物栄養学）
19. Fertilizers, Soils, and Plant Nutrition （肥料，土壌，植物栄養学）
20. History, Education, and Documentation （歴史，教育およびドクメンテーション）

このうち、本研究の探索主題に関係すると考えられるセクションの背景を赤で示した。このなかでも、セクション 1 の薬理学とセクション 13 の哺乳類生化学には、本研究の探索主題に関係する薬理メカニズム情報が多く含まれていることが期待できる。

CA の索引語に着目すると、CA には統制語と自然言語による索引があり、索引者が文献の全文を読んで付与している[82]。STN システムの場合は、検索対象として命名された索引語フィールド（以下、IT フィールドと略す）に索引が配置されている。IT フィールドに含まれる一つの索引（以下、IT 索引と略す）は、基本的に一つの主題を表し、1 件の文献レコードには、複数の IT 索引が付与されている。各 IT 索引には、統制語、ロール、統制語を補足説明するために自然言語で記載されたテキスト説明句がセットで配置されている[83]。このテキスト説明句には文献中で使用されている新しい主題の用語についてもそのまま記載されるため、新しい主題の用語について検索することが可能となる。

3.1.2 情報検索の方法

準備段階(2)の、データベースの CA を用いて情報検索による文献レコードの収集につい

て検討した。情報源として選択した CA は、SciFinder や STN システムなどを使用して検索することができる。SciFinder は、キーワード検索だけではなく、化学構造検索や反応検索機能が利用できる。キーワード検索では、使用した検索語と関連するキーワードを含めて自動的に検索されるが、検索に使用するキーワードが記載される位置を厳密に指定することはできないため、検索の精度が STN システムと比べて不利である。本実験では、検索機能および検索結果の表示機能の必要性に基づいて、STN システムを使用することにした。STN システムを使用することにより、

・IT フィールドを指定した検索、

・語尾変化を許容する前方一致検索、

・近接演算子を使用した、検索語の位置関係を指定した検索、

・特許を除く検索、

・文献の発行年を指定した検索、

・検索結果の索引を確認するための SCAN 形式による表示、

が利用できる。

疾病名を検索語として CA を検索すると、検索結果には基礎研究から臨床試験に近い応用研究まで、様々な研究成果に関する文献が含まれている。この検索結果から新奇な研究シーズ用語を効率よく抽出するためには、新奇な研究シーズ用語が含まれている可能性が高い文献レコードのみに絞り込みできることが望ましい。しかし、探索目的としている新奇な研究シーズ用語は不明であり、検索語として用いることはできない。研究シーズ用語の薬理メカニズム用語に含まれる inhibitor や antagonist などの作用語を検索語として使用することにした。

そのため、CA を検索する際の検索対象であるフィールドの指定方法や探索主題の情報入手に適した検索式の作成などの検索戦略が重要となる。

STN システムでは、検索する際に検索対象フィールドを指定しない場合は、基本索引に含まれる TI（標題）、AB（抄録）、ST（補遺語）、IT（索引語）のフィールドすべてが検索対象になる。基本索引全体を対象として論理演算子の論理積を使用して検索すると、検索語が異なるフィールドに使用されている文献レコードも検索される。これらは、検索語間に意味の上での関係はなく、本研究で探索目的としているものではない。本研究においては、疾病の原因に関係する標的分子への薬剤の作用である薬理メカニズム情報を研究シーズとして、疾病（breast cancer、lung cancer）と薬理メカニズムに含まれる作用語の二つの用語

が関連性をもつ文献を検索する必要がある。そのため、検索対象の範囲は、索引語である統制語とその統制語についての説明を記載したテキスト説明句から構成されるITフィールドを指定し、検索に使用する用語間の関係を限定する方法を使用することにした。

薬理メカニズム用語には標的分子に対する薬剤の作用を示す単語が含まれる場合が多いことが知られている。本研究で探索をおこなう薬理メカニズムの場合には、inhibitor（阻害)、antagonist（拮抗)、agonist（作動）などの作用を表す単語が、標的分子名と組み合わされて薬理メカニズム用語として使用されている。そのため、これらの作用語と疾病名とが同じIT索引に存在することを指定する検索式により､その関係性を指定する。これにより、疾病名と作用語を含む薬理メカニズムが同じIT索引中に存在する場合は検索することができる。そのため、ITフィールドを指定して、一つのIT索引に疾病名と薬理メカニズム用語に含まれる作用語の二つの用語が存在する文献を検索する。また、疾病名は一般事項索引（general subject index）として統制語の部分に記載される場合が多く、一方、薬理メカニズム用語は統制語ではなく、原文で使用された自然言語としてテキスト説明句の部分に記載される場合が多いため、これらの情報を関連付けて検索する指定が必要となる。

検索語間の関係を厳密に指定して検索するために適切な近接演算子を使用する。IT フィールドを指定して、同じIT索引に疾病名と薬理メカニズム用語に含まれる標的分子への作用を示す作用語が含まれている文献を検索するために、STN システムの近接演算子として同じ検索対象フィールド内に前後の検索語が存在することを指定する“(L)”を使用する。検索語の単語間に“(L)”を使用する検索により、近接演算子の前後の単語が同一の情報単位内に存在することが指定できる。そのため、検索式3を使用してCAを検索した。

S (研究ニーズ用語)(L)(作用語)/IT　・・・（検索式3）

研究ニーズ用語：breast cancer、lung cancer

作用語：agonist、antagonist、inhibitor

情報単位の指定としてIT フィールドを指定した場合は、異なる主題の複数のIT 索引が付与されている場合にも、一つのIT索引に検索に使用した複数の用語が含まれているものに限定して検索できる。このように、主題ごとに付与されたIT索引を使用して検索することにより、必要な情報を得ることが期待できる。

検索結果の索引を確認する出力形式であるSCAN形式は、特許分類、CC（CA分類コー

ド)、TI、ST、IT、RL（ロール）を出力することができる。出力指定のオプションコマンドとして TI、HITIND（hit index の略語）を指定することにより、検索でヒットした IT 索引のみを出力することが可能になるため、出力内容の指定に使用した。SCAN 形式では出典の確認に必要な雑誌名などの書誌事項は出力されない。本研究では、検索結果を出力した文献レコードを研究シーズ用語の収集に利用した。この利用目的は、探索調査のために CA の検索に使用する検索語の発見である。検索した文献レコードの使用については、STN システムの代理店である一般社団法人化学情報協会に利用目的と方法について説明し、予め使用許諾を得た。

breast cancer(L)agonist?/IT の検索式により breast cancer と agonist が一つの IT 索引にある文献レコードを検索し、SCAN TI HITIND 形式により出力した例を図 6 に示す。一つの IT フィールド中に検索で使用したキーワードが含まれていることが確認できる。

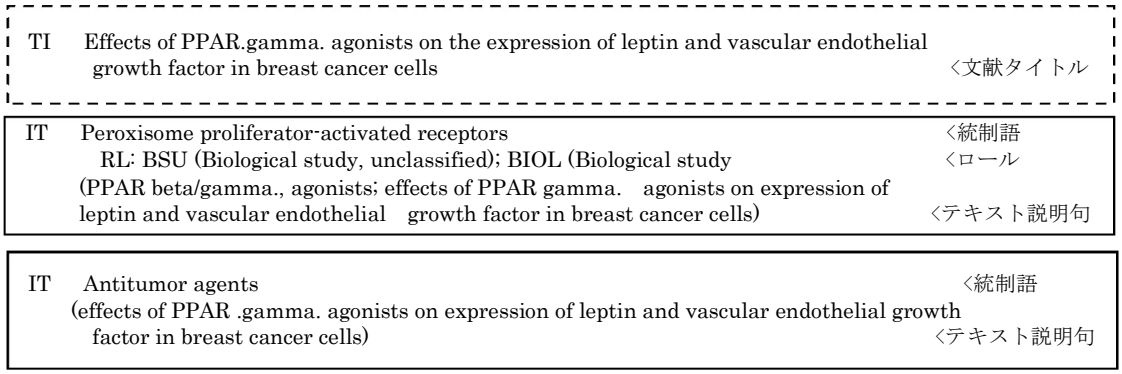

図 6. CA の検索結果の出力例

検索対象の期間に関しては、索引の部分を対象に検索する必要があることを考慮して決定した。すなわち、CA に索引が付与されるまでに、文献発行から長い場合は 2～3 ヶ月を要することがある。そのため、索引が付与された文献を検索するために、発行日については余裕をもたせて検索時より半年以上前に限定した。また、比較的新しい年代の文献を分析することにより、新しい年代に発表された新奇な研究シーズ情報を探索するため発行期間は 1 年間に限定することにした。

3.1.3 データセットの作成

研究シーズ用語の抽出に使用する情報フィルタリングに適したデータセットの作成について検討した。ここにおけるデータセットとは、設定した処理目的に必要な情報要素を含むデータレコードの集合であり、処理プログラムで使用するために処理可能な形態のテキストファイルである。データセットは、データの追加や加工に適している形式である CSV ファイルや TSV ファイルのようなデータレコードに含まれる複数の情報要素をカンマやタブで区切り、データレコードを改行により区切ったテキストファイルで作成することにした。これにより、処理プログラムから直接読み込むことや、データ加工の際に Excel を使用することができるようになる。この処理は、文献レコードとデータセットが定型的な情報であることからテキストマイニングによる自動処理が適しているため、perl 言語で作成したプログラムによるデータセット作成の自動化をおこなった。

文献を検索した結果は文献レコード単位で出力されるため、文献レコードに含まれる主題の概念ごとにレコードを分割する必要性がある。疾病名と作用語を使用して CA を検索した結果には、既知および新奇な薬理メカニズムの情報が混在する可能性がある。この理由は、既知もしくは新奇な薬理メカニズムについてのみ記述された文献があるだけではなく、一つの文献中において、新奇な情報が既知情報と対比して記述されることがあるからである。そのため、研究シーズ用語として新奇な薬理メカニズム用語のみを入手するためには、それらを分離する必要がある。不要な既知情報を除くために、既知情報のキーワードを論理演算子の NOT を使用して検索することは適切ではない。それは、情報検索では文献レコード単位の検索となるため、新奇な情報が含まれている文献も同時に除去される可能性があるからである。

CA の文献レコードには、主題ごとに数種の IT 索引が付与され、上述の様な例では、新奇な情報と既知な情報は、異なる IT 索引として作成すると考えられる。そこで、得られた文献レコードを IT 索引単位に分割した。すなわち、はじめに文献レコードに文献 ID を付与し、つぎに IT 索引単位に文献レコードを分割し、索引 ID を付与して、データレコードを作成するための自動化をおこなった（図 7）。

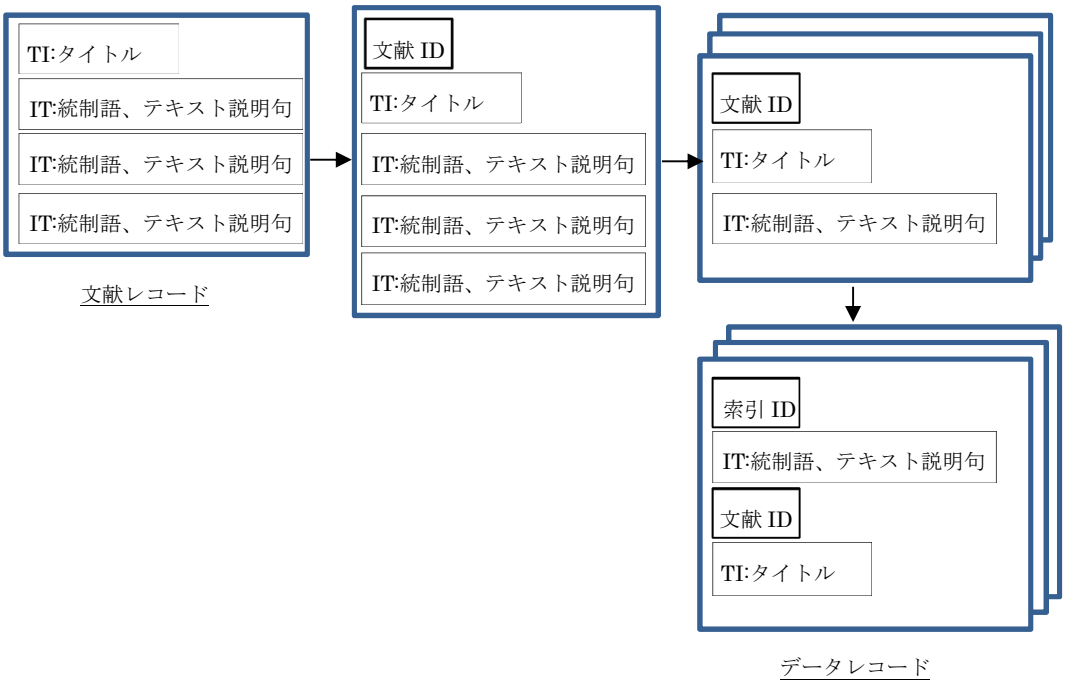

図 7. 文献レコードの分割によるデータレコードの自動作成

新奇な薬理メカニズム用語は、IT 索引の、自然言語で記載されたテキスト説明句に含まれる可能性が高いため、IT 索引のテキスト説明句をデータレコードの情報要素とした。また、IT 索引の統制語に関しても、情報フィルタリングで照合対象とする場合も考えられるため、情報要素として加えた。その他にも文献タイトルは、情報フィルタリングの結果を文献の内容から評価するために必要と考え情報要素の一つとして加えた。さらに、情報フィルタリングで照合してヒットした IT 索引を分離するためには、個々の IT 索引を識別する必要があるため、文献レコードごとの文献 ID の付与とともに、IT 索引単位に分けられたデータレコードごとに索引 ID を付与した。

データレコードには、さらに、テキスト説明句から正規表現によるパターンマッチングにより自動的に切り出した薬理メカニズム用語を情報要素として追加した。パターンマッチングのパターンは正規表現により記述した。すなわち、文頭やカンマなどの文の切れ目もしくは機能語として前置詞、接続詞など（表 8）を目印にし、それらとその後ろにある薬理メカニズムの作用語である inhibitor、antagonist、agonit に挟まれた単語を切り出した（付録１-②)。また、作用語が先に記載される倒置の形態については、予め自動修正する処理をおこなった。

表 8. パターンマッチングの目印

分類	内容
前置詞	at、in、of、from、by、to、on、with
接続詞	and、or、as
その他	:、,

図 8 の場合は、カンマの後ろから inhibitor までの “next-generation ALK inhibitor”を切り出している。

テキスト説明句

alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer

↓切り出しスクリプト

切り出し語

next-generation ALK inhibitor

図 8. テキスト説明句からの薬理メカニズム用語の切り出し

6 個の情報要素をからなる処理用のデータレコードを作成するために、検索結果を出力して得られたデータを perl 言語で作成した正規表現のパターンマッチングにより文献レコード間を区切り、さらに IT 索引単位に分割した。つぎに、文献レコードのタイトル、IT 索引の統制語、IT 索引のテキスト説明句に対して perl 言語で作成したデータレコード作成スクリプト（付録 1-①）によって、文献 ID と索引 ID を付与してデータレコードを作成した（図 9)。図 9 に示すように、データレコードの情報要素は、①文献 ID、②索引 ID、③文献タイトル、④統制語、⑤テキスト説明句、⑥切り出し語とし、タブで区切られた形式とした。また、IT 索引に含まれる CAS ロール（RL）は、物質の文献中の役割を示す情報であるが、研究シーズ情報収集に不要であるため使用しなかった。

＜文献レコード＞

L4　630 ANSWERS　ZCAPLUS　COPYRIGHT 2016 ACS on STN

TI　Alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer　←文献タイトル

IT　Antitumor agents　←統制語

Homo sapiens

Human

Non-small-cell lung carcinoma　↓テキスト説明句

(alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer)

IT　166433-56-3, Anaplastic lymphoma kinase

RL: BSU (Biological study, unclassified); BIOL (Biological study)

(alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer)

IT 索引 → 索引 ID 付与

IT 索引 → 索引 ID 付与

一つの文献レコード→文献 ID 付与

＜データレコード＞

① 文献 ID	② 索引 ID	③ 文献タイトル	IT 索引 ④ 統制語	IT 索引 ⑤ テキスト説明句	⑥ 切り出し語
208	993	Alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer	Antitumor agents Homo sapiens Human Non-small-cell lung carcinoma	alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer	next-generation ALK inhibitor
208	994	Alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer	166433-56-3, Anaplastic lymphoma kinase	alectinib: a selective, next-generation ALK inhibitor for treatment of ALK-rearranged non-small-cell lung cancer	next-generation ALK inhibitor

図 9. 文献レコードの出力情報からのデータレコードの作成例

3.2 breast cancer の治療薬開発に関わる研究シーズ用語の抽出実験

breast cancer の治療薬開発に関わる新奇な研究シーズ用語の候補を抽出するために、第 2 章で検討した方法を用いた実験をおこなった。

3.2.1 実験の準備段階

CA（STN システムの ZCAPLUS ファイル）を情報源とし、疾病名の breast cancer と薬理メカニズムに含まれる作用を示す用語が同じ IT 索引にあることを指定する検索式 4 を使用して文献検索をおこなった。

S (breast cancer)(L)(inhibitor? or antagonist? or agonist?)/IT　・・・（検索式 4）

作用語としては、代表的な inhibitor（阻害作用に使用）、antagonist（拮抗作用に使用）、agonist（作動作用に使用）の 3 語に前方一致の指定“?”を使用して、語尾変化を許容した検索をおこなった（2014 年 9 月 25 日検索）。

検索の結果、9,993 件の文献レコードが得られた。検索結果には特許文献や学術文献が含まれているため、ここから特許を除いて 9,381 件とした。最近の 1 年間を対象とするために資料発行年を 2013 年に限定して 742 件の文献レコードを得た。

CA の検索で得られた 742 件の文献レコードについて、SCAN TI HITIND 形式を使用して文献タイトルおよびヒットした IT 索引を出力し、テキストデータとして保存した。出力したテキストデータは、perl 言語で作成したデータセット作成スクリプトによって自動加工し、個々の IT 索引単位に分割したものをデータレコードとし、その集合体をデータセットとして作成した。この結果、検索で得られた 742 件の文献に付与された IT 索引から、検索でヒットした IT 索引として 2,974 個のデータレコードからなるデータセットが得られた。表 9 にデータレコードの構成と例を示す。

表 9. データレコードの構成および例

① 文献 ID	② 索引 ID	③ 文献タイトル	IT 索引 ④ 統制語	IT 索引 ⑤ テキスト説明句	⑥ 切り出し語
3	10	MiR-221/-222 differentiate prognostic groups in advanced breast cancers and influence cell invasion	Cyclin-dependent kinase inhibitor proteins	CDKN1B; quant. RT-PCR reveals that miR-221/-222 overexpression have high prognostic impact for identification of significantly varied groups in advanced human breast cancers and strongly increases cell proliferation, invasion in vitro	-
4	11	Topoisomerase Inhibitors in Metastatic Breast Cancer: Overview of Current Practice and Future Development	Metastasis	mammary gland neoplasm; topoisomerase inhibitors in metastatic breast cancer	mammary gland neoplasm; topoisomerase inhibitor
4	12	Topoisomerase Inhibitors in Metastatic Breast Cancer: Overview of Current Practice and Future Development	Mammary gland neoplasm	metastasis; topoisomerase inhibitors in metastatic breast cancer	metastasis; topoisomerase inhibitor
4	13	Topoisomerase Inhibitors in Metastatic Breast Cancer: Overview of Current Practice and Future Development	Antitumor agents Combination chemotherapy Drug toxicity Human Topoisomerase inhibitors	topoisomerase inhibitors in metastatic breast cancer	topoisomerase inhibitor

表 9 のとおり、文献 ID が 3 では一つの IT 索引が、文献 ID が 4 では三つの IT 索引が出力されているため、この例では 2 件の文献から、4 個のデータレコードが得られている。索引 ID が 13 では、統制語として複数の異なる統制語がまとめて付与され、共通するテキスト説明句が付与されている。テキスト説明句から正規表現により切り出した薬理メカニズム用語を右端列に示した。索引 ID が 10 のデータレコードについては、テキスト説明句からの切り出し語はなかった。

研究ニーズ、情報源、情報検索、データセットの作成について表 10 にまとめる。

表 10. breast cancer の実験の準備

準備段階	内容	方法	結果
(1)	研究ニーズの選定と情報源の選択	breast cancer CA データベース	
(2)	情報検索	CA データベース／STN システム 検索式：S (breast cancer)(L)(inhibitor? or antagonist? or agonist?)/IT 特許を除き、資料の発行年を 2013 年に限定	742 件の文献レコード
(3)	データセットの作成	SCAN TI HITIND 形式を使用して検索結果を出力し、テキストデータとして保存 テキストデータは、perl 言語で作成したデータセット作成スクリプトによって自動加工 IT 索引単位に分割したものをデータレコードとし、その集合体をデータセットとして作成 データレコードの情報要素は、①文献 ID、②索引 ID、③文献タイトル、④統制語、⑤テキスト説明句、⑥切り出し語	2,974 個のデータレコードからなるデータセット

3.2.2 既知用語辞書の作成

breast cancer の実験における抽出段階(1)の既知用語辞書の作成について述べる。ここでは、多様な疾病に使用できる汎用性の高い辞書を作成するために、全疾病を治療対象とした薬剤の薬理メカニズム用語を収集し、使用することにした。全疾病を対象とした既知用語辞書を使用することにより他の疾病で既に知られている薬理メカニズムも既知として照合されて除かれる。しかし、breast cancer の発症に関係する標的分子は他の部位の癌と共通していない場合が多いことから、全疾病を対象とした既知用語辞書を使用することにした。

既知用語の情報源として、薬剤の適用疾病や開発段階についての情報が記載されている医薬品の研究開発情報データベースの中から世界で標準的に使用されている Pharmaprojects[68] の情報を参照した。さらに同義語の情報を補足するために Integrity[69]を使用した。これらの資料に含まれる薬理メカニズム用語とその同義語、異表記の用語を含めて、6,613 語からなる既知用語の用語リストを作成した。作成した用語リストの例を表 11 に示す。

表 11. 既知用語のリスト

既知用語（薬理メカニズム用語）	薬剤数
HER2 antagonist EGFR 2 antagonist EGFR2 antagonist Epidermal growth factor receptor 2 antagonist ErbB 2 antagonist ErbB2 antagonist	52
EGFR3 antagonist Epidermal growth factor receptor 3 antagonist ErbB-3 antagonist HER3 antagonist antagonist	15

薬理メカニズム用語は、標的分子への薬剤の作用を示す用語であるが、標的分子の用語には多くの異表記の用語や略語があり、文献中でも多くの異表記の用語が使用されている。表 11 に示した HER2 antagonist の場合は、6 種類の異表記の用語が認められる。そのため、文献中で使用される薬理メカニズム用語を認識するためには、よく使用される同義語を構成する用語を集めて表 12 に示す辞書を作成して照合した。

表 12. 既知用語辞書の例

薬理メカニズム用語	薬剤数	異表記の薬理メカニズム用語を構成する単語					
HER2 antagonist	52	EGFR	2	antagonist			
HER2 antagonist	52	EGFR2	antagonist				
HER2 antagonist	52	Epidermal	growth	factor	receptor	2	antagonist
HER2 antagonist	52	ErbB2	antagonist				
HER2 antagonist	52	ErbB	2	antagonist			

また、薬理メカニズム用語には複合語が多く存在し、文献中ではそれらの複合語を構成する単語の順番が変化することや単語間にその他の単語が入ることがわかった。

3.2.3 データセットと既知用語辞書との照合

breast cancer の実験における抽出段階(2)の情報フィルタリングについて述べる。既知の薬理メカニズム用語を検出してデータレコードを絞り込むために、データセットと既知用語辞書との照合をおこなった。

CA を検索して作成したデータセットは、検索結果のレコードを IT 索引ごとのデータレコードに分解して作成した（図 10)。このデータレコードの IT 索引の部分を既知用語辞書と照合して、ヒットしたデータレコードを除去することにより、新奇な研究シーズ用語が含まれている可能性がある IT 索引を得ることにした。検索結果のレコードを IT 索引ごとのデータレコードに分解することが、本研究が提案する方法の特徴であり、既知用語辞書との照合によりヒットした IT 索引に限定して除去することが可能になるため、新奇な研究シーズ用語を含む IT 索引を得ることが期待できる。

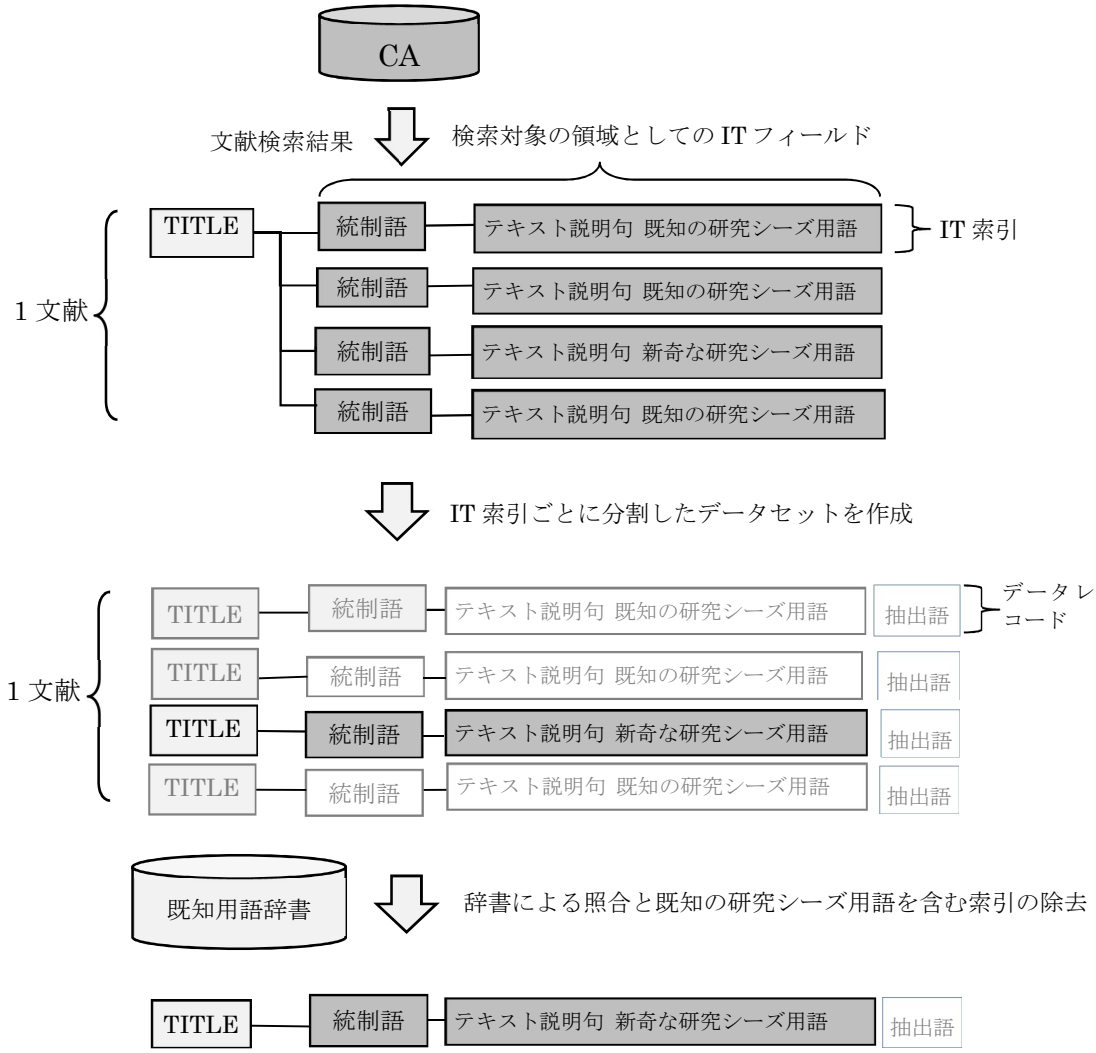

図 10. データセットと既知用語辞書との照合

breast cancer の実験では、複合語の薬理メカニズム用語を構成する単語が照合対象の IT 索引にすべて含まれていることを調べる、論理積を用いた照合方法を使用した。この方法は、既知用語辞書を使用した照合において、データセットの IT 索引中の薬理メカニズム用語を構成する単語の間に形容詞などのその他の単語が複数入った場合にも照合可能なこと、また、作用を示す単語が倒置して薬理メカニズム用語の先頭に使用されているような場合にも照合可能なことなどを考えて使用した。

データセットの既知用語辞書との照合において、照合対象とする個々のデータレコードの情報要素について、対象が異なる 3 パターンの照合をおこないヒットした IT 索引数を比較した（図 11）。データセットの照合対象に使用できるデータレコードの情報要素としては、

文献タイトル、統制語、テキスト説明句、テキスト説明句からの切り出し語の 4 種があるが、文献タイトルを除いた 3 種の情報要素について比較した。

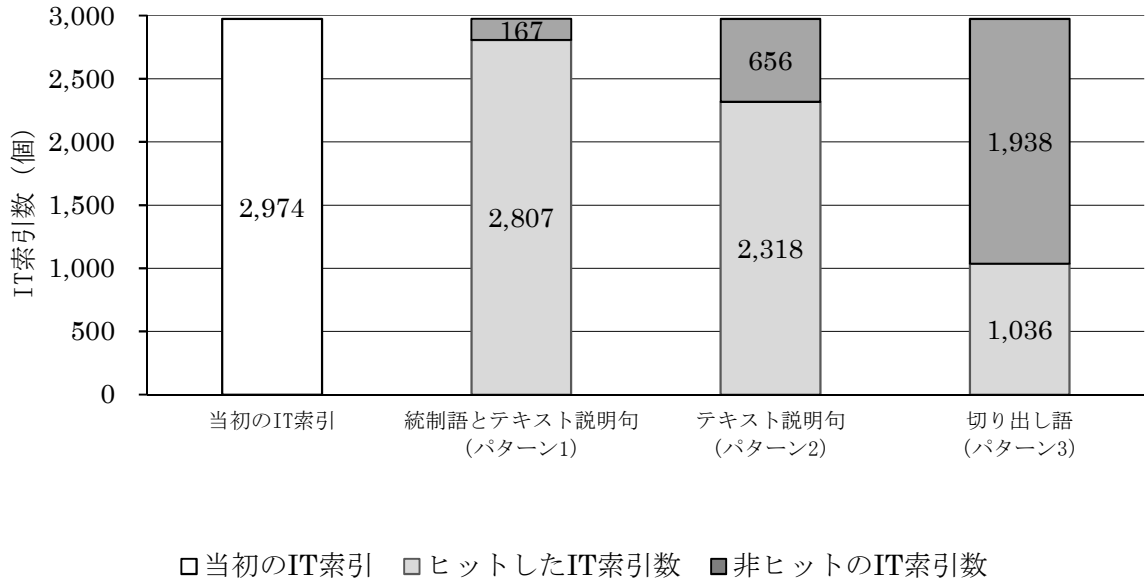

図 11. 情報フィルタリング結果の比較

パターン 1 では、統制語とテキスト説明句の両方の情報要素を対象に既知用語辞書と照合し、2,807 個がヒットし、当初の IT 索引 2,974 個との差である 167 個（5.6%）の IT 索引を新奇な研究シーズ用語が含まれる候補として入手した。

パターン 2 では、テキスト説明句の情報要素のみを対象に既知用語辞書と照合し、2,318 個がヒットし、当初の IT 索引数との差 656 個（22.1%）の IT 索引を新奇な研究シーズ用語が含まれる候補として入手した。

パターン 3 では、テキスト説明句からの切り出し語を対象とした照合により 1,036 個がヒットし、当初の IT 索引数との差 1,938 個（65.2%）の IT 索引を新奇な研究シーズ用語が含まれる候補として入手した。

照合対象とするデータレコードの情報要素により大きく結果が異なり、探索の目的や結果の精度などを考慮して照合対象を選択する必要があることがわかった。

パターン 1 では、統制語とテキスト説明句の両方を対象に照合をおこなうため、薬理メカニズム用語に含まれる単語が統制語の部分に記載されている場合もヒットする。その結果、新奇シーズ用語を含む IT 索引数がもっとも絞り込まれた。

パターン 2 では、自然言語で記載されたテキスト説明句の部分を対象に照合をおこなったため、薬理メカニズム用語に含まれる単語がテキスト説明句に含まれる場合にのみヒットした。このため、フィルタリング結果の新奇なシーズ用語を含む IT 索引数がパターン 1 より大きくなった。

パターン 3 では切り出した薬理メカニズム用語のみを対象に照合をおこなったため、照合の精度はある程度期待できるが、ヒット数は最も少なくなった。情報要素である薬理メカニズムの抽出自体が難しく、データが少ないためヒット数も少なくなり、既知情報として除かれる IT 索引数が少なくなった。

パターンーン 1 から 3 の特徴を検討した結果、本実験では、情報フィルタリングによる情報量の絞込み効果を最も重視し、パターン 1 の統制語とテキスト説明句を照合対象とする方法を選択した。すなわち、167 個のデータレコードを研究シーズ用語の選別に使用することにした。

3.2.4 研究シーズ用語の選別

breast cancer の実験における抽出段階(3)の薬理メカニズム用語の選別について述べる。既知用語辞書による照合により得られた IT 索引 167 個から研究シーズ用語を選別した。IT 索引が 167 個と大きく絞り込まれたことから、選別ルールを用いて手動により薬理メカニズム用語を選別した。選別には、作用語を含むルールと標的分子関連語を含むルールを用いた。

(1) 作用語を含むルール

検索に使用した薬理メカニズムに共起する作用語を含むルールを適用した。薬理メカニズムに共起する作用語は、薬理メカニズムの可能性が最も高いことを想定し、inhibitor を含む用語 17 個（のべ 80 個）、antagonist を含む用語 3 個（のべ 8 個）、agonist を含む用語 5 個（のべ 18 個）の合計 25 個の薬理メカニズム用語を選別した。

(2) 標的分子関連語を含むルール

breast cancer の実験では、情報フィルタリングによる情報量の絞込み効果を重視したため、標的分子関連語についての選別は主なものに絞った。すなわち、作用語を含まない標的分子関連語に関しては、受容体を示す receptor を含む用語 2 個（のべ 4 個）、標的に含まれる factor を含む用語 1 個の合計 3 個の用語のみを選別した。

(1)(2)のルールで選別した結果、28 個の新奇な研究シーズ用語の候補が得られた(表 13)。表 13 の 3 列目には、167 個の IT 索引中に含まれる各用語の頻度を示し、右端列には、薬理メカニズム用語の作用を示す種類を示した。

表 13. 研究シーズ用語の選別結果

No.	新奇な研究シーズ用語の候補	IT 索引中の頻度	作用の種類
1	neurokinin 1 inhibitor	1	inhibitor
2	Eg5 inhibitor	1	inhibitor
3	melatonin inhibitor	1	inhibitor
4	Muscarinic antagonist	1	antagonist
5	Vasopressin receptor agonist	1	agonist
6	bone sialoprotein inhibitor	1	inhibitor
7	PIK3CA inhibitor	1	inhibitor
8	small mol kinase inhibitor	1	inhibitor
9	Toll like receptors 9	1	target molecule
10	phosphoinositide 3 kinase inhibitor	1	inhibitor
11	differentiation 4 inhibitor	1	inhibitor
12	HDAC inhibitor	1	inhibitor
13	GnRH agonist	1	agonist
14	Transcription factor I kappa B	1	target molecule
15	bisphenol agonist	2	agonist
16	Wnt antagonist	2	antagonist
17	pARP inhibitor	2	inhibitor
18	FGF receptor	3	target molecule
19	heat shock protein 90 inhibitor	3	inhibitor
20	matrix metalloprotease inhibitor	4	inhibitor
21	CXCR4 inhibitor	4	inhibitor
22	poly ADP ribose polymerase inhibitor	4	inhibitor
23	adiponectin receptor 1 agonist	5	agonist
24	Wnt beta catenin signaling antagonist	5	antagonist
25	cannabinoid 3 receptor agonist	9	agonist
26	gamma secretase inhibitor	13	inhibitor
27	mTOR inhibitor or rapamycin inhibitor	15	inhibitor
28	PI3K inhibitor	26	inhibitor

薬理メカニズム用語として、inhibitor（阻害剤）関連 17 個、agonist（作動薬）関連 5 個、anatagonist（拮抗薬）関連 3 個、標的分子関連 3 個が得られた。IT 索引中の頻度に関しては、頻度 1 の薬理メカニズム用語が 14 個（50%）と最も多かった。そのなかでも、作用の種類では inhibitor が 9 個と最も多かった。

3.2.5 breast cancer の実験のまとめ

breast cancer の実験における抽出段階(3)で得られた用語が少なかったため、抽出段階(4)はおこなわなかった。breast cancer の治療薬開発に関わる研究シーズ用語の抽出実験の結

果をまとめ、表 14 に示す。

表 14. breast cancer の実験結果のまとめ

段階		過程の内容	実験の内容	実験の結果
抽出段階	(1)	既知用語辞書の作成	医薬品の研究開発情報データベースの Pharmaprojects、Integrity から、全疾病に関する薬理メカニズム用語とその同義語を収集して作成した。	6,613 語からなる既知用語辞書を作成した。
	(2)	データセットと既知用語辞書との照合	データセット(IT 索引の統制語およびテキスト説明句)と既知用語辞書の全データとを、論理積による方法で自動照合した。照合された IT 索引をデータセットから自動除去し、研究シーズ用語が含まれる IT 索引を得た。	照合対象を統制語とテキスト説明句にして、2,974 個のデータレコードのうち、2,807 個が既知用語辞書との照合によりヒットした。ヒットしなかった 167 個の IT 索引に研究シーズ用語が含まれているとして入手した。
	(3)	索引からの研究シーズ用語の選別	得られた IT 索引から、選別ルールを用いて手動により研究シーズ用語を選別した。得られた研究シーズ用語の IT 索引中の頻度を含む、リストを作成した。	28 個の新奇な研究シーズ用語の候補を得た。
	(4)	薬理メカニズム用語の順位付け		実施しなかった。

3.3 lung cancer の治療薬開発に関わる研究シーズ用語の抽出実験

日本人の死亡原因として最も多い疾病である lung cancer の治療薬の研究シーズ探索実験をおこなった。

3.3.1 実験の準備

CA（STN システムの ZCAPLUS ファイル）を情報源とし、疾病名の lung cancer と薬理メカニズムに含まれる作用を示す用語が同じ IT 索引にあることを指定して文献検索をおこなった。薬理メカニズムに含まれる作用を示す用語は、lung cancer を適応症とする治療薬の薬理メカニズムに使用されている作用語を確認して選択することにした。そのため、医薬品の研究開発情報データベースである Pharmapojects[68]に収録されている肺がんの適応症を有する研究開発中の薬剤の情報から 2,642 個（重複を含む）の薬理メカニズム用語を収集した。得られた薬理メカニズム用語から作用語の部分のみの頻度リスト（表 15）を作成した。

表 15. lung cancer 開発薬の作用語頻度

作用語	頻度 （該当薬剤数）
inhibitor	1,440
stimulant	422
antagonist	369
agonist	107
immunotherapy	61
immunosuppressant	11
binder	5
modulator	1
enhancer	1

この結果、頻度上位の inhibitor、antagonist、agonist を薬理メカニズムの作用語として使用することにした。頻度リスト中２番目の stimulant も多く使用されていたが、文献中において薬理メカニズムの作用語以外にも使用されることが多いことから、本実験においては使用しなかった。また、6 番目の immunotherapy は、単独で使用されることが多いため、使用しなかった。

テキスト説明句の部分には、当該疾病名に関係した新奇な薬理メカニズム用語が含まれることが期待される。そのため、選択した三つの作用語が疾病名の lung cancer と同一の IT 索引にあることを指定して検索した。疾病名には、lung cancer の同義語も追加した。近接演算子として“(L)”を用いた検索式 5 を使用し、CA（STN システムの ZCAPLUS ファイル）の検索をおこなった（2016 年 5 月 26 日検索）。

S (lung cancer or lung neoplasm or lung carcinoma)(L)(inhibitor or antagonist or agonist)/IT　・・・（検索式 5）

STN システムにおいて、複数形も自動検索する設定を使用して肺がんに関する 3 種の検索語で検索すると、検索結果の集合 L1 として 134,964 件の文献レコードが得られた（図 12）。

```
* * * * * * * * * * * * * * STN TOKYO * * * * * * * * * * * * * * * * * * * * *
=> FILE ZCAPLUS
FILE 'ZCAPLUS' ENTERED AT 11:22:16 ON 26 MAY 2016
FILE COVERS 1907 - 25 May 2016 VOL 164 ISS 23
FILE LAST UPDATED: 24 May 2016 (20160524/ED)
REVISED CLASS FIELDS (/NCL) LAST RELOADED:  Dec 2015
USPTO MANUAL OF CLASSIFICATIONS THESAURUS ISSUE DATE:  Dec 2015

=> S (LUNG CANCER or LUNG NEOPLASM or LUNG CARCINOMA)          ←肺がんの検索
          395258 LUNG
           66891 LUNGS
          416664 LUNG
                 (LUNG OR LUNGS)
          858403 CANCER
          123152 CANCERS
          886736 CANCER
                 (CANCER OR CANCERS)
           71288 LUNG CANCER
                 (LUNG(W)CANCER)
          395258 LUNG
           66891 LUNGS
          416664 LUNG
                 (LUNG OR LUNGS)
          865001 NEOPLASM
           45226 NEOPLASMS
          885109 NEOPLASM
                 (NEOPLASM OR NEOPLASMS)
           80725 LUNG NEOPLASM
                 (LUNG(W)NEOPLASM)
          395258 LUNG
           66891 LUNGS
          416664 LUNG
                 (LUNG OR LUNGS)
          390586 CARCINOMA
           52278 CARCINOMAS
             204 CARCINOMATA
               1 CARCINOMATAS
          402340 CARCINOMA
                 (CARCINOMA OR CARCINOMAS OR CARCINOMATA OR CARCINOMATAS)
           51620 LUNG CARCINOMA
                 (LUNG(W)CARCINOMA)
L1        134964 (LUNG CANCER OR LUNG NEOPLASM OR LUNG CARCINOMA)   ←ヒット件数
```

図 12. lung cancer の検索過程-1

図 12 の L1 と先に選択した 3 作用語とを“(L)”近接演算子を使用してリンク検索すると、検索結果の集合 L2 として 16,403 件の文献レコードが得られた（図 13)。つぎに、その結果から特許を除くと、L3 として 9,671 件の文献レコードが得られた。さらに、2015 年発行の資料に限定し、L4 として 630 件の文献レコードが得られた。

```
=> S L1(L)(inhibitor or antagonist or agonist)/it          ←作用語と肺がんの検索
          307823 INHIBITOR/IT
          544315 INHIBITORS/IT
          753438 INHIBITOR/IT
                 ((INHIBITOR OR INHIBITORS)/IT)
           57254 ANTAGONIST/IT
          109243 ANTAGONISTS/IT
          145779 ANTAGONIST/IT
                 ((ANTAGONIST OR ANTAGONISTS)/IT)
           38088 AGONIST/IT
           58835 AGONISTS/IT
           83723 AGONIST/IT
                 ((AGONIST OR AGONISTS)/IT)
L2         16403 L1(L)(INHIBITOR OR ANTAGONIST OR AGONIST)/IT   ←ヒット件数

=> S L2 NOT P/DT                                          ←特許を除く
        11256403 P/DT
L3          9671 L2 NOT P/DT

=> S L3 AND 2015/PY                                       ←2015 年発行に限定
         2459821 2015/PY
L4           630 L3 AND 2015/PY

=> D L4 SCAN TI HITIND        ←IT 索引部分はヒットした索引のみ出力

L4     630 ANSWERS     ZCAPLUS    COPYRIGHT 2016 ACS on STN
TI     Afatinib for the treatment of non-small cell lung cancer
IT     80449-02-1, Protein tyrosine kinase
       RL: BSU (Biological study, unclassified); BIOL (Biological study)
          (inhibitor; discussion on afatinib for treatment of non-small
          cell lung cancer)

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1):630        ←結果の出力数を指定
```

図 13. lung cancer の検索過程-2

L4 として得られた 630 件のレコードを対象として、レコード毎にタイトル、ヒットした IT 索引を出力した。得られた検索結果 630 件のうち 2 件のレコードを出力した例を図 14 に示す。

L4 630 ANSWERS ZCAPLUS COPYRIGHT 2016 ACS on STN
TI Upregulation of PD-L1 by EGFR Activation Mediates the Immune Escape in EGFR-Driven NSCLC: Implication for Optional Immune Targeted Therapy for NSCLC Patients with EGFR Mutation
ST non small cell lung cancer EGFR PDL1
IT 80449-02-1, Protein tyrosine kinase
(inhibitor; upregulation of PD-L1 by epidermal growth factor receptor activation mediates immune escape in patients with EGFR-driven non-small-cell lung cancer)

L4 630 ANSWERS ZCAPLUS COPYRIGHT 2016 ACS on STN
TI SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer
ST SMARCE EGFR expression MET ALK inhibitor human lung cancer
IT Gene, animal
RL: BSU (Biological study, unclassified); PRP (Properties); BIOL (Biological study)
(CBX2; study on SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer)
IT Epidermal growth factor receptors
RL: BSU (Biological study, unclassified); PRP (Properties); BIOL (Biological study)
(EGFR; study on SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer)
IT 137632-03-2, MET 142243-02-5, ERK 142805-58-1, MEK 148640-14-6, AKT 166433-56-3, ALK kinase
RL: BSU (Biological study, unclassified); BIOL (Biological study)
(study on SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer)

図 14. lung cancer の検索過程-3

各文献レコードのタイトル（TI）と索引情報（IT）を出力したが、同じ IT フィールドに、疾病名の lung cancer および作用語の inhibitor が付与されていることが確認できる。

630 件の文献レコードに付与されている、検索でヒットした IT 索引のみ 1,375 個を出力した。1 文献で平均 2.2 個の IT 索引が検索によりヒットしたことになる。図 14 の形式で出力したテキストファイルを処理スクリプトにより連続処理して、データセットを作成した（表 16）。

表 16. 作成した lung cancer のデータセット例

文献 ID	索引 ID	文献タイトル	IT 索引		切り出し語
			統制語	テキスト説明句	
567	1279	Improving cancer immunotherapy with DNA methyltransferase inhibitors	CD8-positive T cell Homo sapiens Human Immune cell Immunotherapy Kidney neoplasm Lung neoplasm Natural killer cell Ovarian carcinoma Solid neoplasm	discussion on improving cancer immunotherapy with DNA methyltransferase inhibitors	DNA methyltransferase inhibitor
568	1280	Phase II Study of the AKT Inhibitor MK-2206 plus Erlotinib in Patients with Advanced Non-Small Cell Lung Cancer Who Previously Progressed on Erlotinib	Antitumor agents Combination chemotherapy Homo sapiens Human Non-small-cell lung carcinoma Prognosis Signal transduction	phase II study of AKT inhibitor MK-2206 plus erlotinib in patients with advanced non-small cell lung cancer who previously progressed on erlotinib	AKT inhibitor
568	1281	Phase II Study of the AKT Inhibitor MK-2206 plus Erlotinib in Patients with Advanced Non-Small Cell Lung Cancer Who Previously Progressed on Erlotinib	183321-74-6, Erlotinib 1032349-93-1, MK-2206	phase II study of AKT inhibitor MK-2206 plus erlotinib in patients with advanced non-small cell lung cancer who previously progressed on erlotinib	AKT inhibitor
569	1282	Design and synthesis of dithiocarbamate linked .beta.-carboline derivatives: DNA topoisomerase II inhibition with DNA binding and apoptosis inducing ability	Antitumor agents Apoptosis Homo sapiens Human Lung neoplasm Mammary gland neoplasm Neoplasm Prostate gland neoplasm Structure-activity relationship Uterine cervical carcinoma	dithiocarbamate-linked .beta.-carboline derivs. prepn. as topo II inhibitors and apoptosis inducers	topo II inhibitor
570	1283	Silver I, Gold I and Gold III -N-Heterocyclic carbene complexes of naphthyl substituted annelated ligand: Synthesis, structure and cytotoxicity	Colorectal carcinoma Human cell line A549 Human cell line HCT116 Human cell line MCF-7 Lung carcinoma Mammary gland adenocarcinoma	inhibitors; prepn., structure, DFT, and antitumor activity of silver and gold N-heterocyclic carbene complexes of a naphthyl substituted annelated ligand	prepn. inhibitor

表 16 に示したとおり、一つの文献には複数の統制語が付与されている場合が多い。これらの統制語は文献中において共通する主題に付与されている場合は、一つの IT 索引とし

てまとめられ、共通するテキスト説明句が付与されている。

研究ニーズ、情報源、情報検索、データセットの作成の結果についてまとめて、表 17 に示す。

表 17. lung cancer の実験の準備

準備段階	内容	方法	結果
(1)	研究ニーズの選定と情報源の選択	lung cancer CA データベース	
(2)	情報検索	CA データベース／STN システム 検索式：S (lung cancer or lung neoplasm or lung carcinoma)(L)(inhibitor or antagonist or agonist)/IT 複数形も自動検索する設定 特許を除き、資料の発行年を 2015 年に限定	630 件の文献レコード
(3)	データセットの作成	SCAN TI HITIND 形式を使用して検索結果を出力し、テキストデータとして保存 テキストデータは、perl 言語で作成したデータセット作成スクリプトによって自動作成 IT 索引単位に分割したものをデータレコードとし、その集合体をデータセットとして自動作成 データレコードの情報要素は、①文献 ID、②索引 ID、③文献タイトル、④統制語、⑤テキスト説明句、⑥切り出し語	1,375 個のデータレコードからなるデータセット

3.3.2 既知用語辞書の作成

lung cancer の実験における抽出段階(1)の既知用語辞書の作成について述べる。肺がんは日本人の死亡原因として最も多い疾病であり、新たな治療薬の開発が望まれている。肺がんのなかでも腺がんなどは、他の臓器のがんからの転移も多く、他の臓器における抗がん剤が有効な場合がある。そのため、他の疾病に対しては既知の薬理メカニズムであっても、肺がんに対しては新奇な可能性がある薬理メカニズムを候補とするために、既知の範囲を肺がんに定めて既知用語辞書を作成することにした。既知用語辞書の作成については、調査の領域特有の判断が必要であり自動化は難しいこと、また自動化によるメリットも大きくない

ことから自動化はおこなわなかった。

lung cancer の実験では、肺がん関連の適応を有して研究開発中の約 700 薬剤から得られた薬理メカニズム 376 種とその同義語を収集して既知用語辞書を作成した。薬理メカニズムとその同義語の収集には、医薬品の研究開発情報データベースの Pharmaprojects[68]、Integirty[69]、さらに日本で研究開発されている薬剤の情報に詳しい明日の新薬[84]を使用した。はじめにこれらの研究開発情報データベースを疾病の肺がんを示す統制語を使用して検索して得られた薬剤情報から薬理メカニズム用語を切り出し、それに対応する同義語情報を各データベースの用語リストから収集した。薬理メカニズムごとに基本とする薬理メカニズムの名称を指定して、薬理メカニズム 376 種に関する全 3,269 個の同義語を含めた薬理メカニズム用語からなる CSV ファイル形式の既知用語辞書を作成した（図 15)。

```
bradykinin receptor B1 antagonist,B1BKR antagonist
bradykinin receptor B1 antagonist,B1R antagonist
bradykinin receptor B1 antagonist,BDKRB1 antagonist
bradykinin receptor B1 antagonist,BK-1 receptor antagonist
bradykinin receptor B1 antagonist,BKB1R antagonist
bradykinin receptor B1 antagonist,BKR1 antagonist
bradykinin receptor B1 antagonist,BRADYB1 antagonist
bradykinin receptor B1 antagonist,bradykinin B1 receptor antagonist
bradykinin receptor B1 antagonist,bradykinin receptor 1 antagonist
bradykinin receptor B1 antagonist,bradykinin receptor B1 antagonist
bradykinin receptor B2 antagonist,B2 bradykinin receptor antagonist
bradykinin receptor B2 antagonist,B2R antagonist
bradykinin receptor B2 antagonist,BDKRB2 antagonist
bradykinin receptor B2 antagonist,BK2 antagonist
bradykinin receptor B2 antagonist,BK-2 antagonist
bradykinin receptor B2 antagonist,BK-2 receptor antagonist
bradykinin receptor B2 antagonist,BKR2 antagonist
bradykinin receptor B2 antagonist,bradykinin receptor B2 antagonist
bradykinin receptor B2 antagonist,BRB2 antagonist
```

図 15. lung cancer 既知用語辞書の例

3.3.3 データセットと既知用語辞書との照合

lung cancer の実験における抽出段階(2)の情報フィルタリングについて述べる。データセットに含まれる薬理メカニズム用語と既知用語辞書との照合をおこない、さらに、既知の用語を含むデータレコードをデータセットから除いた。これらの処理については、定型的な処

理が可能であり処理プログラムによる自動化のメリットが高いため、自動化した。はじめに、近接演算子の機能を有するスクリプトを作成して、データセットと既知用語辞書との照合をおこなった。近接演算子の機能としては、近接演算子の前後に入力した単語が、順不同で、単語間に 2 個以下の単語を含み近接していることを指定できるようにした。これにより、薬理メカニズムを構成する単語の並び順には倒置などを含めて多様性があり、また単語間に前置詞や冠詞等の記号や単語が不規則に入ることにある程度対応できるようにした。

データセットの 1,375 個の IT 索引のテキスト説明句を対象に、既知用語辞書との照合をおこなった。その結果、636 個が既知の薬理メカニズム用語を含む IT 索引としてヒットし、ヒットしなかった 739 個（53.7%）の IT 索引に新奇な研究シーズ用語が含まれるとして分離した。

3.3.4 研究シーズ用語の選別

lung cancer の実験における抽出段階(3)の薬理メカニズム用語の選別について述べる。IT 索引からの研究シーズ用語選別は手動では処理負担が大きいこと、さらにある程度の定型化が可能であることから自動化のメリットが大きいと判断し、部分的に処理プログラムによる自動化をおこなった。得られた IT 索引から研究シーズ用語の複合名詞などを選別するために、ストップワードによるテキスト説明句の分割処理をした。分割に必要なストップワードのリストは、SMART システム[85]において標準的に使用されている用語を基本に、CA の機能語[86]などを追加した 574 語を使用した（付録 2）。分割により全部で 8,058 個（延べ語数）の単語または複合語からなる研究シーズ用語を含む用語リストを得た。

8,058 個の単語と複合語からなる用語リストの中から、薬理メカニズム用語や関連する標的分子用語を選別するために、選別ルールを用いた関数および手動によるデータ・クリーニングをおこない、さらに検討した特徴量の関数による一般用語の除去をおこなった。

(1) 薬理メカニズム用語を構成する作用語を含むルール

薬理メカニズム用語に含まれる薬剤の標的分子への作用を示す作用語を含むルールを適用し、作用語を含む用語は薬理メカニズムの可能性が最も高いことから、inhibitor を含む用語 577 個、antagonist または agonist を含む用語 46 個を Excel の関数を用いて選別し、合計 623 個の用語を選別した。

(2) 標的分子関連語を含むルール

作用語を含まない標的分子やその略語などがストップワードにより選別された場合を想定し、これらを含めて選別するために薬理メカニズムに含まれる標的分子に関連する用語として target を含む用語 80 個を Excel の関数を用いて選別した。つぎに、標的分子名が大文字を含む略語で記載されることが多いとの研究[71]から、大文字を含む用語 1,064 個を選別した。

作用語を含むルールにより選別した用語と合計すると、合計 1,767 個の用語を薬理メカニズム用語、標的分子用語の候補として選別した。

(3) データ・クリーニング

得られた研究シーズ用語の表記のゆれを正規化するためにデータ・クリーニングをおこなった。ピリオド、ハイフンなどの記号、動詞、形容詞を除去した後にソートし、同じ語幹を有する語で名寄して頻度集計をおこない 499 個（のべ 1,421 個）の選別用語リストを得た。そのうち、IT 索引中の頻度が少ない研究シーズ用語の一部を表 18 に示す。作用語が含まれていた研究シーズ用語に関しては、その種類を示した。

表 18. 選別した研究シーズ用語の例

研究シーズ用語	IT 索引中の頻度	作用の種類
A 893	1	
AZD9150	1	
CSC 3436	1	
ETNPD5	1	
Kir2 1 KCNJ2	1	
MIF rpS3	1	
MIP 1232	1	
Mps1 kinase inhibitor	1	inhibitor
NBM T BBX OS01	1	
SKLB 677	1	
InsP3 Kinase	2	
FGFR3 BAIAP2L1	2	
H1650GR	2	
p21c Ki ras	2	

表 18 に示すように IT 索引中の頻度が少ない研究シーズ用語には、作用語を含む用語は少なかった。

3.3.5 研究シーズ用語の順位付け

lung cancer の実験における抽出段階(4)の順位付けについて述べる。抽出段階(3)では、選別ルールによる関数を使用した選別および手動によるデータ・クリーニングをおこない研究シーズ用語として 499 個を得た。この 499 個の用語には、データ・クリーニングでは除去できなかった出現頻度が高い薬理メカニズム用語、薬理メカニズムとは異なる専門用語、一般用語などが含まれている可能性がある[87]。そのため、用語自体の出現頻度による順位付けをおこなった。

出現頻度による順位付けについては、データベースの自動検索による変数値の入手および順位付けの計算について処理プログラムによる自動化をおこなう。本研究では、使用頻度の少ない用語の抽出を目的としているため、これらに関係する頻度を基にした重み付けとして、関数式 3 の STF-IDF を提案した。STF-IDF の計算に用いる df 値として、文献データベース全体における当該用語の文献レコード数を用いるため、499 個の研究シーズ用語を使用して、Web-API 機能を利用したプログラムによって PubMED[88]を自動検索し、文献レコード数を得た（付録-3)。PubMED は、CA に収録されない医薬品に関する開発が進んだ臨床試験の文献が多く収載されているため、文献レコード数の多少を示す df 値として適していると考えた。検索は、PubMED の“All Fields”を対象に、複合名詞に関してはダブルクォーテションで囲んだ完全一致のフレーズ検索とした。検索によって得られた文献レコード数を df 値とした。また、df 値の取得に使用する API の検索頻度は API の利用規約[89]の遵守を考慮して、1 秒 1 回に制限した。はじめに、得られた df 値を参考にして、研究シーズ用語 499 個に含まれる薬理メカニズム用語と関係のないと思われる用語 64 個を除去して 435 個を使用とした。除去した用語を表 19 に示す。

表 19. 除去した用語の例

除去した理由	除去した用語
df 値が高い PubMED で多く使用されている一般的な専門用語	lung、iressa、PPAR、potential biomarker、RNA、stage III、SNP、Lewis、management、development など
df 値が 0 の PubMED でまったく使用されていない用語	chem inhibitor、plant derived protease、Fuzheng Kang、small mol inhibitor、factor 88 inhibitor、Covalent Irreversible Kinase など
異なる概念の略語が結合したもの	IL 6/JAK1/STAT3、Src/FAK ERK beta catenin、ERKs/RSK2、microRNA 212/132、mTOR complex 1/2、CADM1/TSLC1、A549/taxol など

本実験では、出現頻度の低い用語の抽出を優先しているため、出現頻度である df 値が小さいことを優先とする必要がある。そのため、df 値が数百以上などある程度大きいものは、研究シーズ用語であっても、既知の用語として除去することにした。そこで、得られた 435 個の用語から、PubMED の文献レコード数が多かった用語を除くために閾値として df 値を 200 とし、それ以上の 259 個の用語を除去し、200 未満のもの 176 個に削減した。この閾値は、探索主題や目的によって変化すると考えられるが、再現性を重視する場合は、数値をより高くし、ノイズを少なくしたい場合は、数値を低くすることができる。すなわち、他の疾病などの治療薬として実績のある薬理メカニズムを含めて調査する場合は、df 値を高く設定する必要がある。

176 個の研究シーズ用語の順位付けの計算には CA および PubMED から得られた情報を用いるため、関数式 4 に示すとおり変数を具体的に指定した。

$$\text{STF-IDF} = (\text{stf} + \text{a}) \times \left(\log\left(\frac{\text{n}}{\text{df}}\right) + 1\right) \quad \cdots \text{（関数式 4）}$$

stf：CA の検索によりヒットした全 IT 索引中に含まれる用語の頻度

a：stf 値を調整するための値

df：PubMED における全収録年代の当該用語を含む文献レコード数

n：全文献レコード数（PubMED の 1 年の収録数として 100 万を使用（2015 年収録 120 万件））

関数式 4 において、stf 値には、CA を検索して得られた IT 索引に含まれる用語数を用いた。a 値は stf 値を調整するための値とした。df 値には PubMED における全収録年代の当

該用語を含む文献レコード数を使用した。n 値は全文献レコード数（PubMED の 1 年の収録数として 100 万）を使用した。

本実験では、情報フィルタリングにより得られた全 IT 索引中にある研究シーズ用語の頻度を stf 値とするため、stf 値は研究ニーズとの関係の強さを反映した数値と考える。実験での stf 値は 10 以下の比較的低頻度のものが多かった。また、df 値を使用した計算には対数が使用されているため、実際の数値の差が軽減される。しかし、stf 値を単純に積として計算した場合は、stf 値が小さな値どうしの比較においては計算値が大きな値となるため、STF-IDF 値による順位に大きな影響を与えることがわかった。そこで、a の数値を少しずつ変化させて計算した結果、a の値として 30 を使用することにより、stf 値の影響を残しつつ、出現頻度の低い用語の順位が上位になる傾向にあることがわかった。この 30 に関しては、実際に数字を段階的に変化させて用語の順位と stf 値、df を観察できるように可視化した結果を付録-4 に示す。この数値は、探索主題や CA の索引の傾向により stf 値やそのばらつきに大きな差が生じることが予測されることから、設定が難しいが、可視化することによりある程度のシミュレーションは可能であると考えられる。

df 値が 200 未満による限定で得られた 176 個の用語を使用して、関数式 4 を用いた STF-IDF 値を計算し、降順にソートした。結果のリストを表 20 に示す。このリストの上位の用語は出現頻度が低いことが期待でき、一般用語や既によく知られた専門用語は下位となるため、探索目的に応じて個々の用語を調査することに使用できると考える。

表 20. 疾病を限定した既知用語辞書による研究シーズ用語の選別結果

No.	研究シーズ用語	stf CA 2015 年発行の 該当索引内の数	df PubMED 全収録年代の 文献レコード数	STF-IDF （stf+30 使用）
1	InsP3Kinase	8	2	254.6
2	KLF17	15	31	247.9
3	AZD9150	4	1	238.0
4	CREB CBP	13	58	225.2
5	CSC 3436	2	1	224.0
6	NBM T BBX OS01	2	1	224.0
7	BIP 4	8	17	219.2
8	A 893	1	1	217.0
9	AKT1 PDPK1	1	1	217.0
10	ETNPD5	1	1	217.0
11	MIF rpS3	1	1	217.0
12	MIP 1232	1	1	217.0
13	SKLB 677	1	1	217.0
14	Kir2 1 KCNJ2	1	1	217.0
15	Mps1 kinase inhibitor	1	1	217.0
16	topoisomerase 1 inhibitor	12	70	216.5
17	Stat1 HDAC4	2	2	214.4
18	FGFR3 BAIAP2L1	2	2	214.4
19	transforming growth factor beta inhibitor	6	14	210.7
20	Thymidylate Synthase RNA	2	3	208.7
21	RegIIA	2	3	208.7
22	H1650GR	1	2	207.7
23	microRNA 19	4	8	207.3
24	Mdig	5	15	203.8
25	Betti reaction	1	3	202.2
26	LFC131	1	3	202.2
27	SMYD3 inhibitor	1	3	202.2
28	miRNA 506	2	5	201.6
29	phosphoinositide 3 kinase alpha	7	37	201.0
30	HIV smoker	3	9	199.5
31	CLK kinase	1	4	198.3
32	alpha 3 beta 2 nAChR	1	4	198.3
33	DDR1 inhibitor	1	4	198.3
34	CC 223	1	4	198.3
35	EGFRT790M	1	4	198.3
36	covalent reversible inhibitor	2	7	197.0
37	ST2825	3	12	195.4
38	EGFR C797S	1	5	195.3
39	LY2090314	1	5	195.3
40	BAY 87 2243	1	5	195.3
41	gold N heterocyclic carbene	1	5	195.3
42	SET oncogene	2	8	195.1
43	MEK162	5	28	194.3
44	LW6	3	13	194.2
45	protein kinase D inhibitor	2	9	193.5
46	BCL2 alpha	1	6	192.9
47	EBUS NA	1	6	192.9
48	desmosdumotin B	1	7	190.8
49	DDX3	10	174	190.4
50	2 yr OS	1	8	189.0
51	MET dependency	1	8	189.0
52	SSR128129E	1	8	189.0
53	AMG 232	1	8	189.0
54	FGFR1b	4	28	188.8
55	ARRY 142886	7	81	188.4
56	MHMD	2	13	188.4
57	factor 1 inhibitor	3	20	188.1
58	CQN	1	9	187.4
59	PF 04449913	1	9	187.4

60	ZD 6474	1	10	186.0
61	SHH antagonist	2	16	185.5
62	common DNA methylation	1	11	184.7
63	18F FAZA PET	1	11	184.7
64	G1202R	1	11	184.7
65	checkpoint kinase inhibitor	3	26	184.3
66	FAP inhibitor	1	12	183.5
67	included c Myc	1	12	183.5
68	MET TKIs	1	12	183.5
69	NCI H 460	1	12	183.5
70	angiokinase inhibitor	5	61	182.5
71	miR 27a inhibitor	1	13	182.5
72	TAK 733	1	13	182.5
73	LUX Lung 6	1	14	181.5
74	NF BETA	1	14	181.5
75	U2OS cell line	5	67	181.1
76	1 O acetylbritannilactone	1	15	180.5
77	PNAS 4	1	15	180.5
78	ABLIM	2	23	180.4
79	DYRK1A kinase	2	25	179.3
80	type 1 5 alpha reductase	1	17	178.9
81	microRNA 192	2	26	178.7
82	CX 4945	4	58	178.0
83	Wnt/ beta catenin signaling inhibitor	1	20	176.7
84	TopoII inhibitor	1	21	176.0
85	PF 03084014	1	21	176.0
86	TrkB inhibitor	2	34	175.0
87	beta4 nicotinic acetylcholine	1	23	174.8
88	JMJD5	1	23	174.8
89	LUX Lung 3	1	23	174.8
90	casein kinase 1 alpha	1	23	174.8
91	YM155	6	144	174.3
92	2 oxoglutarate oxygenase	1	24	174.2
93	CXCR2 inhibitor	1	27	172.6
94	CCR9 CCL25	1	28	172.1
95	MAPK11	1	28	172.1
96	macrophage inhibitor	2	44	171.4
97	HDAC1 inhibition	1	31	170.8
98	LDH inhibitor	1	33	169.9
99	BRD4 inhibitor	1	33	169.9
100	Wnt 7B	1	34	169.5
101	acetyl CoA carboxylase inhibitor	1	36	168.8
102	RHAMM receptor	1	39	167.7
103	HDAC6 inhibition	2	58	167.6
104	CARP 1	1	41	167.0
105	geranylgeranyltransferase I inhibitor	1	41	167.0
106	Symptomatic Radiation Pneumonitis	2	61	166.9
107	SJSA 1	1	42	166.7
108	smoothened inhibitor	2	63	166.4
109	FoxO6	1	43	166.4
110	KIF5B RET	1	43	166.4
111	proteins XIAP	1	43	166.4
112	aurora B inhibitor	1	44	166.1
113	TUSC2	1	44	166.1
114	OSI 906	1	47	165.2
115	cIAP 2	5	196	164.8
116	BIR2 domain	1	49	164.6
117	CNI 1493	2	72	164.6
118	RhoGDI2	2	72	164.6
119	adenosylhomocysteine hydrolase inhibitor	1	50	164.3
120	JMJD2	1	52	163.8
121	TIPE2	1	53	163.5
122	NR expression	1	55	163.0
123	AZD9291	1	55	163.0
124	anaplastic lymphoma kinase inhibitor	1	55	163.0
125	Elephantopus scaber	1	56	162.8

126	HOPX	1	57	162.6
127	TPO receptor agonist	1	59	162.1
128	SMYD2	1	59	162.1
129	JARID1B	2	87	161.9
130	Wnt 2	1	61	161.7
131	miR 340	1	64	161.0
132	RANKL inhibitor	2	95	160.7
133	cucurbitacin B	3	136	160.6
134	TMPRSS4	1	67	160.4
135	class III receptor	1	68	160.2
136	chronic immune thrombocytopenia ITP	1	71	159.6
137	HOXD3	1	71	159.6
138	IKB alpha	2	106	159.2
139	DKK4	1	74	159.1
140	Wnt 7A	1	77	158.5
141	NCI H446	2	113	158.3
142	HIF 1 inhibitor	2	116	157.9
143	early DNA damage	1	85	157.2
144	G1 growth arrest	2	127	156.7
145	BIM EL	1	89	156.6
146	multitargeted kinase inhibitor	1	90	156.4
147	AUY922	2	130	156.4
148	microRNA 101	1	91	156.3
149	peroxiredoxin II	1	93	156.0
150	glycogen synthase kinase 3 inhibitor	1	98	155.3
151	ARHGDIB	1	100	155.0
152	metadherin	1	101	154.9
153	sU11274	1	102	154.7
154	SMYD3	1	106	154.2
155	EZH2 inhibitor	1	112	153.5
156	GAS5	1	113	153.4
157	DOCK1	1	120	152.5
158	TSLC1	1	121	152.4
159	Notch 1 signaling	1	123	152.2
160	FAP 1	1	124	152.1
161	ABT 263	2	180	151.8
162	MSH2 expression	1	127	151.8
163	ANRIL	1	132	151.3
164	Hedgehog inhibitor	1	134	151.1
165	cyclin E2	1	139	150.6
166	PC9	2	199	150.4
167	HS 4	1	141	150.4
168	DKK2	1	142	150.3
169	CDKN2D	1	144	150.1
170	neplanocin A	1	157	148.9
171	HAI 1	1	159	148.8
172	ATP synthase inhibitor	1	159	148.8
173	TFPI 2	1	179	147.2
174	mitochondrial complex I inhibitor	1	180	147.1
175	MAPK14	1	192	146.2
176	Src protein tyrosine kinase	1	194	146.1
合計	-	333	-	-

3.3.6 lung cancer の実験のまとめ

lung cancer の治療薬開発に関わる研究シーズ用語の抽出実験の結果をまとめ、表 21 に示す。

表 21. lung cancer の実験のまとめ

段階		過程の内容	実験の内容	実験の結果
抽出段階	(1)	既知用語辞書の作成	医薬品の研究開発情報データベースの Pharmaprojects、Integrity、明日の新薬を参照して、lung cancer に関する薬理メカニズム用語とその同義語を収集して手動により既知用語辞書を作成した。	特定の疾病を対象として、3,269 語からなる既知用語辞書を作成した。
	(2)	データセットと既知用語辞書との照合	データレコードのテキスト説明句を対象に近接演算子による方法で、既知用語辞書の全データと自動照合した。照合された IT 索引をデータセットから自動除去し、研究シーズ用語が含まれる IT 索引を得た。	1,375 個のデータレコードのうち、636 個が既知用語辞書との照合によりヒットした。ヒットしなかった 739 個の IT 索引に研究シーズ用語が含まれているとして入手した。
	(3)	索引からの研究シーズ用語の選別	得られた IT 索引のテキスト説明句をストップワードにより用語に自動分解して得たデータから、選別ルール関数による自動選別および手動により研究シーズ用語を選別した。	499 個の新奇な研究シーズ用語の候補を得た。
	(4)	研究シーズ用語の順位付け	lung cancer の実験では、多くの研究シーズ用語が選別されたため、STF-IDF の自動計算による順位付けをおこなった。	出現頻度による制限と順位付けにより 176 個の新奇な研究シーズ用語の候補を得た。

3.4 第 3 章のまとめ

準備段階(1)として、研究ニーズとしては breast cancer と lung cancer を選定し、情報源については CA を選択した。準備段階(2)の情報検索の方法については、CA は STN システム上で検索することにし、検索語として研究ニーズである疾病名（breast cancert または lung cancer）と、薬理メカニズムに含まれる作用語（inhibitor、antagonist、agonist）とを使用した。これらの検索語が同じ IT 索引に含まれることを“(L)”近接演算子と“/IT”により指定して検索した。

疾病名(L)作用語/IT

特許を除き、指定した 1 年間に発行された文献に限定した。準備段階(3)として検索された文献レコードを SCAN 形式により出力してテキスト形式で保存し、IT 索引単位に分割して、文献 ID、索引 ID、タイトル、統制語、テキスト説明句、切り出し語の情報要素から構成されるデータレコードを作成し、これらをまとめてデータセットとした。このデータセットを対象として、本研究で提案した情報フィルタリングの方法を用いて新奇な研究シーズ用語の抽出実験をおこなった。

抽出段階(1)として、医薬品の研究開発情報データベースから薬理メカニズム用語を収集することにより既知用語辞書を手動により作成した。breast cancer の抽出実験では、医薬品の研究開発情報データベースの Pharmaprojects と Integrity から全疾病の治療薬の薬理メカニズム用語とその同義語を収集して既知用語辞書を作成した。lung cancer の実験では、医薬品の研究開発情報データベースの Pharmaprojects、Integrity に明日の新薬を追加して既知用語辞書を作成した。対象疾病は lung cancer に限定した。

抽出段階(2)として、データセットの全データの先頭から順に既知用語辞書に対して照合を繰り返す自動処理をおこなった。breast cancer の実験では、データセットの IT 索引（統制語とテキスト説明句）と既知用語辞書とを、論理積により自動照合した。lung cancer の実験では、データセットの IT 索引のテキスト説明句と既知用語辞書とを、近接演算子の機能を有するスクリプトを使用して自動照合した。breast cancer の実験と lung cancer の実験とも、照合された IT 索引をデータセットから自動除去し、新奇な研究シーズ用語が含まれる IT 索引を得た。

抽出段階(3)として、IT 索引からの研究シーズ用語の選別については、得られた IT 索引数により選別方法を変更した。breast cancer の実験では、得られた IT 索引が少なかったため、選別ルールを用いて手動により研究シーズ用語を選別した。

lung cancer の実験では選別した研究シーズ用語が多かったため、抽出段階(4)として、研究シーズ用語の順位付けをおこなった。選別した研究シーズ用語の IT 索引中の頻度および PubMED を自動検索した文献レコード数などを使用した STF-IDF 値の自動計算により出現頻度による順位付けをし、研究シーズ用語を削減した。

第 3 章では、研究課題 2 として設定した抽出実験により、提案した方法が実行可能であるかを確認した。実験により確認した研究シーズ用語の抽出方法をまとめ、表 22 に示す。

表 22. 実験のまとめ

段階		過程の内容	breast cancer	lung cancer
準備段階	(1)	探索主題の研究ニーズと情報源の選択	探索主題の研究ニーズとして疾病の breast cancer と lung cancer を選択した。 情報源として CA を選択した。	
	(2)	情報検索の方法	CA を検索するシステムには、STN システムを選択した。 検索語としては、研究ニーズの疾病名（breast cancer または lung cancer）と、作用語（inhibitor、antagonist、agonist）とを使用した。 これらの検索語が IT 索引フィールドにあることを“(L)”近接演算子と“/IT”により指定し、手動により検索した。 特許を除き、指定した 1 年間に発行された文献に限定した。	
	(3)	データセットの作成	検索結果の文献レコードを出力し、IT 索引単位に分解して、文献 ID、索引 ID、文献タイトル、統制語、テキスト説明句、切り出し語の情報要素から構成されるデータレコードを自動作成し、これらをまとめてデータセットとした。	
抽出段階	(1)	既知用語辞書の作成	医薬品の研究開発情報データベースの Pharmaprojects と Integrity を参照して、薬理メカニズム用語とその同義語を収集して手動により作成した。 対象疾病は全疾病とした。	医薬品の研究開発情報データベースの Pharmaprojects、Integrity、明日の新薬を参照して、薬理メカニズム用語とその同義語を収集して手動により作成した。 対象疾病は lung cancer に限定した。
	(2)	データセットと既知用語辞書との照合	データセットの IT 索引と既知用語辞書とを、breast cancer は論理積、lung cancer は近接演算子により自動照合し、照合された IT 索引をデータセットから自動除去し、研究シーズ用語が含まれる IT 索引を得た。	
	(3)	IT 索引からの研究シーズ用語の選別	得られた IT 索引から、作用語を含むなどの選別ルールを用いて手動により研究シーズ用語を選別した。	得られた IT 索引から、選別ルール（作用語を含む、大文字を含む）を用いて、関数による自動選別および手動により研究シーズ用語を選別した。
	(4)	研究シーズ用語の順位付け	—	選別件数が多かったため、研究シーズ用語を出現頻度で順位付けするスコアとして STF-IDF を定義し、自動計算した STF-IDF 値で降順にソートして、新奇な研究シーズ用語の候補リストを得た。

第 4 章 考察

第 2 章で研究課題 1 の新奇な研究シーズ用語の候補を抽出する方法を提案し、第 3 章で研究課題 2 の抽出実験により、提案した方法の実行可能性を確認した。本章では、抽出実験結果に基づいて提案方法の妥当性について考察する。

一つ目に、抽出した研究シーズ用語を文献レコード数により評価する（4.1 節）。二つ目に、情報フィルタリングの各抽出段階について考察する（4.2 節）。三つ目に、本抽出研究の意義と展望ついて述べる（4.3 節）。

4.1 文献レコード数に基づく新奇性の確認

第 3 章の実験では、疾病による文献検索で得られた最近 1 年間の文献レコードから新奇な研究シーズ用語を抽出した。2.6 節で示した考え方に基づき、抽出した研究シーズ用語の CA データベース収録全期間における文献レコード数を確認することにより、提案方法の妥当性について考察する。

breast cancer の実験では 28 個、lung cancer の実験では 176 個の新奇な研究シーズ用語の候補が得られた。得られた用語が、研究ニーズとして設定した疾病に対して新奇な研究シーズ用語であることについては、最終的に CA を用いた収録全期間を対象とした文献検索により確認する。すなわち、情報フィルタリングによって得られた研究シーズ用語と疾病名の二つの検索語を関連付けた検索式を用いて CA を文献検索し、文献レコード数が多い場合には使用した検索語間の関係性が高いとは限らないが、文献レコード数が少ない場合は調査時点までに発行された文献においては関係性が低かったことは推測できると考える。STN システムを使用して CA を検索するために使用した検索式を、検索式 6 として示す。

S (研究ニーズ用語の疾病名)(L)(抽出した研究シーズ用語)/IT　・・・（検索式 6）

近接演算子“(L)”を使用することにより、この演算子の前後の検索語が、同一情報単位内に存在する文献レコードを検索する指定ができる。検索式 6 に示すとおり、検索式に “ /IT ” を使用することにより、検索対象を IT フィールドに指定して検索することができる。同一の IT 索引に研究シーズ用語と疾病名とが存在する文献レコードに限定されるため、研究ニ

ーズ用語と研究シーズ用語が高い関連性を有する文献を検索できる。

第 3 章の情報フィルタリングの実験に使用したデータセットは最近の 1 年間の文献データを使用したが、確認実験ではデータベース収録の全期間を対象として検索する。また、検索に使用した研究シーズ用語が複合語の場合は、STN システムの近接演算子“(2A)”を用いて検索する。これにより、検索に使用した複合語を構成する単語の語順に関係なく、単語間に 2 語までの他の単語が含まれてもヒットする条件とし、研究シーズ用語の表記の多様性にも対応する。

ここで得られた検索結果の文献レコード数については、その数値が小さい場合は関連する研究の報告が少ないと考えられるため、過去にさかのぼって低頻度の情報であることを確認できる。

4.1.1 breast cancer の実験により得られた研究シーズ用語の新奇性

4.1.1.1 CA の文献レコード数による研究シーズ用語の新奇性

breast cancer の実験において抽出した 28 個の研究シーズ用語について、データベース収録の全期間における文献を検索し、新奇性を確認した。検索式 7 を使用して、CA 収録の全期間の文献を対象に、breast cancer と研究シーズ用語とが一つの IT 索引に記載された文献を近接演算子の“(L)”を使用して検索した。研究シーズ用語が複合語の場合には、研究シーズ用語を構成する単語間に近接演算子“(2A)”を使用して検索した。CA の IT 索引は、異なる主題ごとに独立して作成されているため、そこに含まれる用語間の関係は深いと考えられる（2015 年 4 月 16 日検索）。

S (breast cancer?)(L)(研究シーズ用語)/IT　・・・（検索式 7）

検索結果は、表 23 の右端列の文献レコード数欄に示した。文献レコード数は、10 件以下が 11 用語（39%）、11-50 件が 12 用語（43%）、50 件以上が 5 用語（18%）であった。

表 23. breast cancer に対する研究シーズ用語の新奇性の確認

No.	新奇な研究シーズ用語の候補	IT 索引数	文献レコード数 （2015/4/16 検索）
1	neurokinin 1 inhibitor	1	1
2	Eg5 inhibitor	1	2
3	melatonin inhibitor	1	22
4	muscarinic antagonist	1	2
5	vasopressin receptor agonist	1	2
6	bone sialoprotein inhibitor	1	5
7	PIK3CA inhibitor	1	11
8	small mol kinase inhibitor	1	22
9	toll like receptors 9	1	19
10	phosphoinositide 3 kinase inhibitor	1	25
11	differentiation 4 inhibitor	1	24
12	HDAC inhibitor	1	35
13	GnRH agonist	1	52
14	transcription factor I kappa B	1	240
15	bisphenol agonist	2	1
16	Wnt antagonist	2	8
17	pARP inhibitor	2	49
18	FGF receptor	3	22
19	heat shock protein 90 inhibitor	3	24
20	matrix metalloprotease inhibitor	4	5
21	CXCR4 inhibitor	4	12
22	poly ADP ribose polymerase inhibitor	4	97
23	adiponectin receptor 1 agonist	5	1
24	Wnt beta catenin signaling antagonist	5	3
25	cannabinoid 3 receptor agonist	9	2
26	gamma secretase inhibitor	13	26
27	mTOR inhibitor or rapamycin inhibitor	15	143
28	PI3K inhibitor	26	73

つぎに、breast cancer の実験で得られた研究シーズ用語の新奇性を評価するために、表 23 の文献レコード数を使用して、文献レコード数による研究シーズ用語の分布状況を確認した。結果を図 16 に示す。

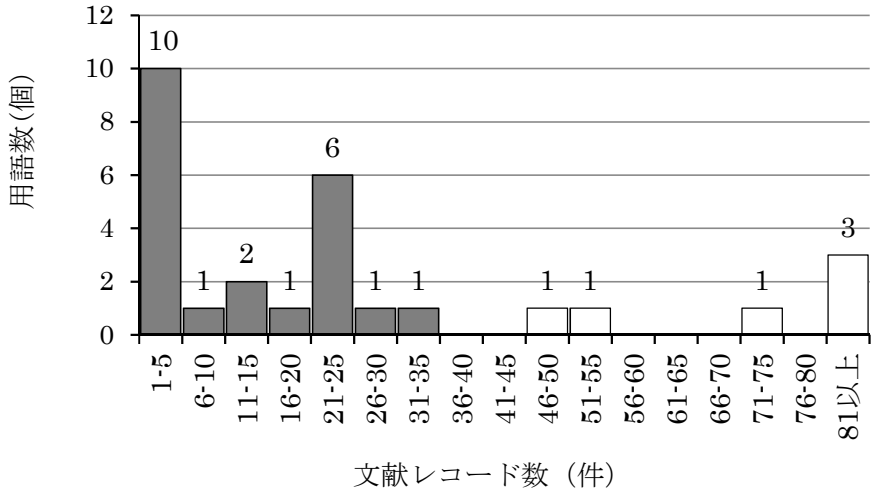

図 16. 文献レコード数による研究シーズ用語の分布

全 28 個の研究シーズ用語のうち、文献レコード数 20 件以下の区間には 14 個（50%）、文献レコード数が 35 件以下の区間の合計では 22 個（79%）が認められた（黒色棒で示す）。文献レコード数による新奇性の判断は、探索目的や競合性により値の評価が異なると考えられる。創薬研究において、新奇性が低いと判断される開発が進んだ薬理メカニズムと疾病に関する文献レコード数は数百件と多くなる傾向がある。多くの研究開発がおこなわれている breast cancer の治療薬の例として、市販薬の標的分子である“HER2”ついて検索式 7 を用いた検索をおこなうと、文献レコード数は 3,403 件となった。これらと比較すると本実験により得られた文献レコード数は少なく、抽出した研究シーズ用語には新奇性があると考えることができる。このように、本研究で提案する情報フィルタリングの方法を用いた breast cancer に関する実験において、新奇な研究シーズ用語を抽出できることがわかった。

4.1.1.2 研究シーズ用語全体と比較した新奇性

抽出した研究シーズ用語の文献レコード数をその他の研究シーズ用語全体の文献レコード数と比較するために、実験により得られた研究シーズ用語が使用されている文献発行年別の文献レコード数を調べた。

表 23 に示した breast cancer に関する研究シーズ用語を検索式 8 を使用して、CA の全収録年代の文献を対象に検索をおこない、文献レコード数を調べた。研究シーズ用語が複合

語の場合には、研究シーズ用語を構成する単語間に近接演算子“(2A)”を使用して検索した。

S (breast cancer?)(L)(研究シーズ用語 22 個)/IT　・・・（検索式 8）

抽出した研究シーズ用語リストのなかで、文献レコード数が 35 件以下であった 22 個の研究シーズ用語（図 16 の黒色棒）について、検討をおこなった。これらの 22 個の研究シーズ用語を含む文献レコード数は 2013 年発行までに限定すると合計で 270 件であり、これらの発行年を集計した。文献発行年別の文献レコード数を図 17 の黒色棒に示す。

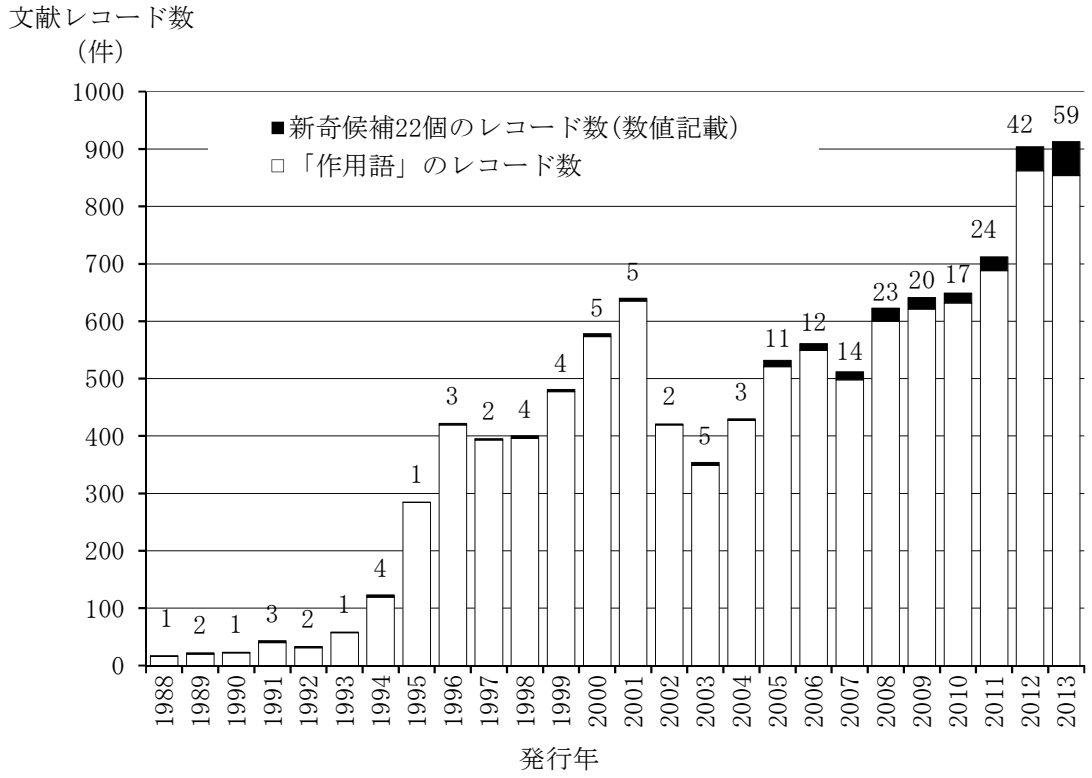

図 17. breast cancer と薬理メカニズムの作用語が記載された文献レコード数

これらの文献が低頻度の情報であることを確認するために、breast cancer と薬理メカニズムの作用として使用されることが多い antagonist、agonist、inhibitor とを関連付けて検索する近接演算子の“(L)”を用いた検索式 9 による検索結果を図 17 の白色棒に示す。

S (breast cancer?)(L)(antagonist or agonist or inhibitor)/IT　・・・(検索式 9)

breast cancer と作用語のリンク検索により得られた文献は、1995 年以降に急速に増加している。黒棒で示した実験で得られた研究シーズ用語は全年代で認められたが、2005 年以降に増加していた。白棒の総計に相当する breast cancer と作用語が同じ IT 索引にある文献は、3.2.1 項の検索において全収録年代で 9,993 件あるため、本研究で提案した方法により新奇な研究シーズ用語を抽出できたと考えられる。

4.1.2 lung cancer の実験により得られた研究シーズ用語の新奇性

lung cancer の実験において抽出した 176 個について、データベース収録の全期間における文献を検索し、新奇性を確認した。検索式 10 を使用して、CA 収録の全期間の文献を対象に、lung cancer と 176 個の研究シーズ用語とが一つの IT 索引に記載された文献を近接演算子の“(L)”を使用して検索した。CA の IT 索引は、異なる主題ごとに独立して作成されているため、そこに含まれる用語間の関係は深いと考えられる（2016 年 5 月 31 日検索)。

S (lung cancer or lung neoplasm or lung carcinoma)(L)(研究シーズ用語)/IT

･･･（検索式10）

実際にCAを検索した際の検索過程を図18、図19に示す。検索では複合語が多い研究シーズ用語のリストに対して、単語が指定した順番で出現し、単語間に2個以内の単語の存在を許容する近接演算子“(2W)”の使用や複数形を含める指定をすることにより、表記の揺れをある程度の許容できるようにした。また、全収録年代を対象として検索した。

```
* * * * * * * * * * * * * * STN TOKYO * * * * * * * * * * * * * * * * * * * * *
FILE 'HOME' ENTERED AT 15:31:26 ON 31 MAY 2016

=> file hcaplus
=> s (lung cancer or lung neoplasm or lung carcinoma)/IT
           292369 LUNG/IT
            13914 LUNGS/IT
           298981 LUNG/IT
                  ((LUNG OR LUNGS)/IT)
           440035 CANCER/IT
            13940 CANCERS/IT
           450603 CANCER/IT
                  ((CANCER OR CANCERS)/IT)
            40280 LUNG CANCER/IT
                  ((LUNG(W)CANCER)/IT)
           292369 LUNG/IT
            13914 LUNGS/IT
           298981 LUNG/IT
                  ((LUNG OR LUNGS)/IT)
           855916 NEOPLASM/IT
            28916 NEOPLASMS/IT
           871031 NEOPLASM/IT
                  ((NEOPLASM OR NEOPLASMS)/IT)
            80419 LUNG NEOPLASM/IT
                  ((LUNG(W)NEOPLASM)/IT)
           292369 LUNG/IT
            13914 LUNGS/IT
           298981 LUNG/IT
                  ((LUNG OR LUNGS)/IT)
           329923 CARCINOMA/IT
             9674 CARCINOMAS/IT
               22 CARCINOMATA/IT
                1 CARCINOMATAS/IT
           330806 CARCINOMA/IT
                  ((CARCINOMA OR CARCINOMAS OR CARCINOMATA OR
CARCINOMATAS)/IT)
            44890 LUNG CARCINOMA/IT
                  ((LUNG(W)CARCINOMA)/IT)
L1         118944 (LUNG CANCER OR LUNG NEOPLASM OR LUNG CARCINOMA)/IT

=> s (A(2w)893)(L)L1
         33699999 A
             4269 893
L2                1 (A(2W)893)(L)L1

=> s (AKT1(2w)PDPK1)(L)L1

             5876 AKT1
              238 PDPK1
L3                1 (AKT1(2W)PDPK1)(L)L1
```

図 18. lung cancer の CA の検索過程-1

```
=> s (AZD9150)(L)L1

                3 AZD9150
L4              1 (AZD9150)(L)L1

=> s (CSC(2w)3436)(L)L1

             6217 CSC
             3121 CSCS
             7619 CSC
                  (CSC OR CSCS)
              272 3436
L5              1 (CSC(2W)3436)(L)L1

=> s (ETNPD5)(L)L1

                1 ETNPD5
L6              1 (ETNPD5)(L)L1

=> s (IL(2w)6(2w)JAK1(2w)STAT3)(L)L1

           282415 IL
             9437 ILS
           287312 IL
                  (IL OR ILS)
          6060762 6
             3221 JAK1
            22414 STAT3
L7              1 (IL(2W)6(2W)JAK1(2W)STAT3)(L)L1

=> s (Kir2(2w)1(2w)KCNJ2)(L)L1

             1270 KIR2
         14450199 1
              321 KCNJ2
L8              2 (KIR2(2W)1(2W)KCNJ2)(L)L1

=> s (MIF(2w)rpS3)(L)L1

             4943 MIF
              146 MIFS
             5007 MIF
                  (MIF OR MIFS)
              845 RPS3
L9              1 (MIF(2W)RPS3)(L)L1

=> s (MIP(2w)1232)(L)L1

            15474 MIP
             4202 MIPS
            18248 MIP
                  (MIP OR MIPS)
             3194 1232
```

図 19. lung cancer の CA の検索過程-2

検索結果の文献レコード数を表 24 の右端列に示す。

表 24. lung cancer に関わる研究シーズ用語を含む文献レコード数

No.	研究シーズ用語	STF-IDF（stf+30 使用）	lung cancer との検索による CA の全収録年代の文献レコード数(2016/5/31)
1	InsP3Kinase	254.6	1
2	KLF17	247.9	2
3	AZD9150	238.0	1
4	CREB CBP	225.2	1
5	CSC 3436	224.0	1
6	NBM T BBX OS01	224.0	1
7	BIP 4	219.2	1
8	A 893	217.0	1
9	AKT1 PDPK1	217.0	1
10	ETNPD5	217.0	1
11	MIF rpS3	217.0	1
12	MIP 1232	217.0	1
13	SKLB 677	217.0	1
14	Kir2 1 KCNJ2	217.0	2
15	Mps1 kinase inhibitor	217.0	3
16	topoisomerase 1 inhibitor	216.5	7
17	Stat1 HDAC4	214.4	1
18	FGFR3 BAIAP2L1	214.4	2
19	transforming growth factor beta Inhibitor	210.7	3
20	Thymidylate Synthase RNA	208.7	1
21	RegIIA	208.7	2
22	H1650GR	207.7	1
23	microRNA 19	207.3	2
24	Mdig	203.8	5
25	Betti reaction	202.2	1
26	LFC131	202.2	1
27	SMYD3 inhibitor	202.2	1
28	miRNA 506	201.6	1
29	phosphoinositide 3 kinase alpha	201.0	8
30	HIV smoker	199.5	1
31	CLK kinase	198.3	1
32	alpha 3 beta 2 nAChR	198.3	1
33	DDR1 inhibitor	198.3	2
34	CC 223	198.3	2
35	EGFRT790M	198.3	2
36	covalent reversible inhibitor	197.0	2
37	ST2825	195.4	1
38	EGFR C797S	195.3	1
39	LY2090314	195.3	1
40	BAY 87 2243	195.3	2
41	gold N heterocyclic carbene	195.3	4
42	SET oncogene	195.1	2
43	MEK162	194.3	1
44	LW6	194.2	1
45	protein kinase D inhibitor	193.5	1
46	BCL2 alpha	192.9	1
47	EBUS NA	192.9	1
48	desmosdumotin B	190.8	3
49	DDX3	190.4	8
50	2 yr OS	189.0	1
51	MET dependency	189.0	1
52	SSR128129E	189.0	1
53	AMG 232	189.0	2
54	FGFR1b	188.8	2
55	ARRY 142886	188.4	4
56	MHMD	188.4	1
57	factor 1 inhibitor	188.1	28
58	CQN	187.4	1
59	PF 04449913	187.4	1
60	ZD 6474	186.0	1

61	SHH antagonist	185.5	1
62	common DNA methylation	184.7	1
63	18F FAZA PET	184.7	2
64	G1202R	184.7	2
65	checkpoint kinase inhibitor	184.3	21
66	FAP inhibitor	183.5	1
67	included c Myc	183.5	1
68	MET TKIs	183.5	1
69	NCI H 460	183.5	1
70	angiokinase inhibitor	182.5	8
71	miR 27a inhibitor	182.5	1
72	TAK 733	182.5	1
73	LUX Lung 6	181.5	2
74	NF BETA	181.5	11
75	U2OS cell line	181.1	3
76	1 O acetylbritannilactone	180.5	2
77	PNAS 4	180.5	4
78	ABLIM	180.4	2
79	DYRK1A kinase	179.3	3
80	type 1 5 alpha reductase	178.9	1
81	microRNA 192	178.7	15
82	CX 4945	178.0	4
83	Wnt/ beta catenin signaling inhibitor	176.7	2
84	TopoII inhibitor	176.0	1
85	PF 03084014	176.0	2
86	TrkB inhibitor	175.0	1
87	beta4 nicotinic acetylcholine	174.8	1
88	JMJD5	174.8	2
89	LUX Lung 3	174.8	2
90	casein kinase 1 alpha	174.8	4
91	YM155	174.3	16
92	2 oxoglutarate oxygenase	174.2	1
93	CXCR2 inhibitor	172.6	1
94	CCR9 CCL25	172.1	3
95	MAPK11	172.1	4
96	macrophage inhibitor	171.4	2
97	HDAC1 inhibition	170.8	2
98	LDH inhibitor	169.9	2
99	BRD4 inhibitor	169.9	6
100	Wnt 7B	169.5	4
101	acetyl CoA carboxylase inhibitor	168.8	3
102	RHAMM receptor	167.7	3
103	HDAC6 inhibition	167.6	1
104	CARP 1	167.0	1
105	geranylgeranyltransferase I inhibitor	167.0	4
106	Symptomatic Radiation Pneumonitis	166.9	4
107	SJSA 1	166.7	1
108	smoothened inhibitor	166.4	2
109	FoxO6	166.4	1
110	KIF5B RET	166.4	25
111	proteins XIAP	166.4	132
112	aurora B inhibitor	166.1	8
113	TUSC2	166.1	9
114	OSI 906	165.2	14
115	cIAP 2	164.8	46
116	BIR2 domain	164.6	1
117	CNI 1493	164.6	1
118	RhoGDI2	164.6	16
119	adenosylhomocysteine hydrolase inhibitor	164.3	1
120	JMJD2	163.8	2
121	TIPE2	163.5	2
122	NR expression	163.0	1
123	AZD9291	163.0	17
124	anaplastic lymphoma kinase inhibitor	163.0	67
125	Elephantopus scaber	162.8	2
126	HOPX	162.6	6

127	TPO receptor agonist	162.1	2
128	SMYD2	162.1	3
129	JARID1B	161.9	4
130	Wnt 2	161.7	11
131	miR 340	161.0	7
132	RANKL inhibitor	160.7	2
133	cucurbitacin B	160.6	11
134	TMPRSS4	160.4	17
135	class III receptor	160.2	1
136	chronic immune thrombocytopenia ITP	159.6	1
137	HOXD3	159.6	10
138	IKB alpha	159.2	2
139	DKK4	159.1	6
140	Wnt 7A	158.5	16
141	NCI H446	158.3	74
142	HIF 1 inhibitor	157.9	9
143	early DNA damage	157.2	3
144	G1 growth arrest	156.7	2
145	BIM EL	156.6	4
146	multitargeted kinase inhibitor	156.4	17
147	AUY922	156.4	14
148	microRNA 101	156.3	22
149	peroxiredoxin II	156.0	7
150	glycogen synthase kinase 3 inhibitor	155.3	7
151	ARHGDIB	155.0	5
152	metadherin	154.9	11
153	sU11274	154.7	20
154	SMYD3	154.2	2
155	EZH2 inhibitor	153.5	15
156	GAS5	153.4	6
157	DOCK1	152.5	6
158	TSLC1	152.4	407
159	Notch 1 signaling	152.2	5
160	FAP 1	152.1	6
161	ABT 263	151.8	23
162	MSH2 expression	151.8	11
163	ANRIL	151.3	9
164	Hedgehog inhibitor	151.1	40
165	cyclin E2	150.6	28
166	PC9	150.4	27
167	HS 4	150.4	1
168	DKK2	150.3	3
169	CDKN2D	150.1	12
170	neplanocin A	148.9	2
171	HAI 1	148.8	2
172	ATP synthase inhibitor	148.8	3
173	TFPI 2	147.2	15
174	mitochondrial complex I inhibitor	147.1	1
175	MAPK14	146.2	10
176	Src protein tyrosine kinase	146.1	3
合計	-	-	1,575

つぎに、表 24 に示した文献レコード数を使用して、文献レコード数による研究シーズ用語の分布状況を図 20 に示し、lung cancer に対して研究シーズ用語が低頻度の情報であることを確認する。文献レコード数が 10 件以下の研究シーズ用語が 145 用語（82.3%）、文献レコード数が 30 件以下では研究シーズ用語は 170 用語（96.6%）となり、研究ニーズの lung cancer に対して新奇な研究シーズ用語の候補が得られたと考える。

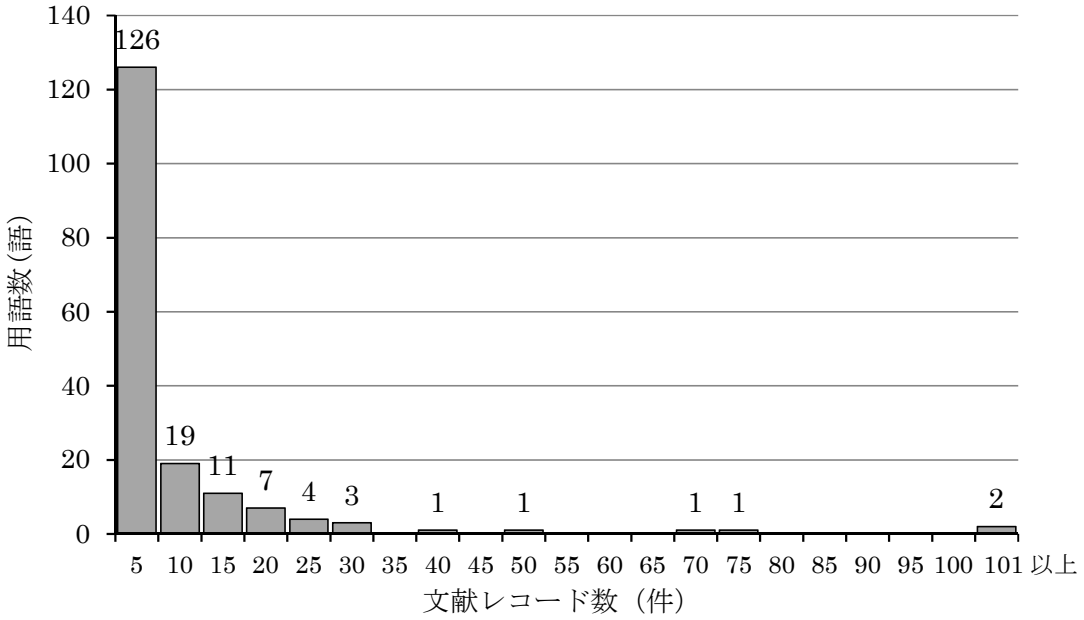

図 20. 文献レコード数による研究シーズ用語の分布

縦軸に示した用語数については、探索主題の文献数の規模と比較した相対的な数値となる。lung cancer に関しては、CA に収載されている全文献は 10 万件以上あり、lung cancer の代表的な標的分子の例である epidermal growth factor receptor の文献レコード数は lung cancer と検索すると 2,000 件以上となる。これらと比較して抽出した研究シーズ用語を用いて lung cancer と検索すると文献レコード数は 30 件以下に絞り込むことができていることから、これらのなかに新奇な研究シーズ用語が含まれる可能性が高いと考えられる。このように、本研究で提案する情報フィルタリングの方法を用いた lung cancer に関する実験において、新奇な研究シーズ用語の候補を抽出できることがわかった。とくに、大量の研究シーズ用語が得られた場合も、用語の出現頻度の観点で順位付けすることにより、必要な範囲の用語の調査を検討することができるようになった。このように、従来は大量の文献検索の結果から手動により研究シーズ用語を確認して集計する必要があったが、それらの負担を軽減する方法として期待できることがわかった。

また、検索結果の文献レコード数と STF-IDF との関係を図 21 に示す。STF-IDF の値が高い研究シーズ用語は、実際に CA の収録の全期間における文献レコード数についても少なくなる傾向があることがわかった。

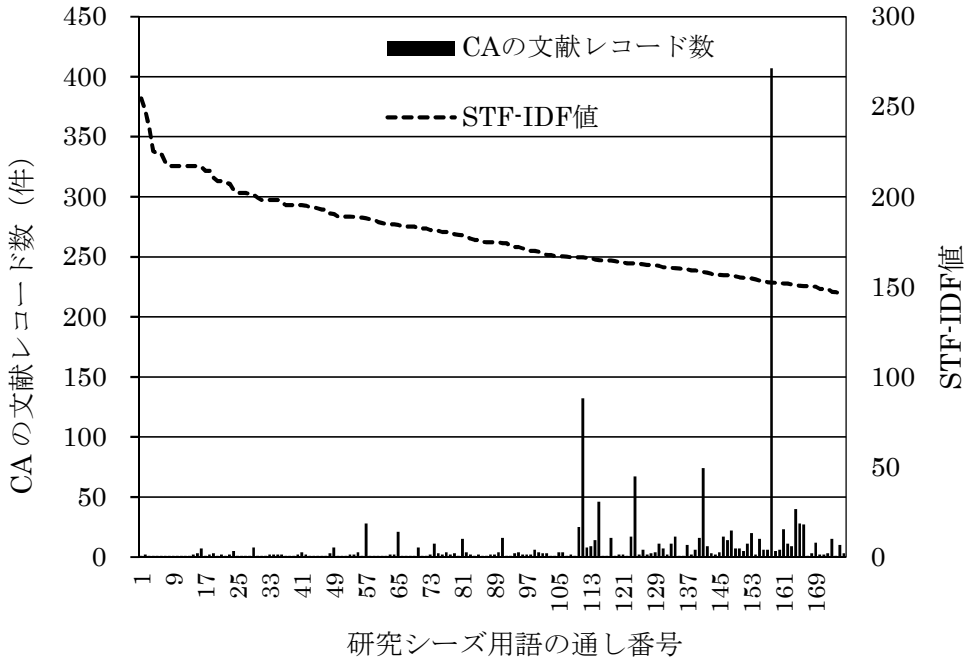

図 21. lung cancer に関わる研究シーズ用語を含む文献レコード数

4.1.2 項において、lung cancer の実験により得られた研究シーズ用語 176 個について、データベース収録の全期間における新奇性を確認した。しかし、文献レコード数は少なく、初期の研究段階である可能性が高い場合でも、最初の文献発行年については古い場合もある。そのため、文献の発行年を区切った検索をすることにより、これらを確認する。

情報フィルタリングのためにデータセットを作成する際に検索対象とした2015年時点から2年以上前の、2013年以前に発行された文献数を調べるために、検索式11を用いて検索した（2017年4月18日検索）。

S (lung cancer or lung neoplasm or lung carcinoma)(L)(新奇な研究シーズ用語)/IT NOT 2014-/PY　・・・（検索式11）

この文献レコード数が0件となった場合は、上記の2016年5月に全収録年代を対象として検索された文献が、2014年以降に発行された文献であることが確定できる。これにより、当該用語に関係する最初の文献発行が過去2年半以内となるものが明らかになり、時期的に研究初期である可能性が高い研究シーズ用語であることがわかる。検索の結果、研究シーズ

用語176個に関する927件の文献レコードが得られた。検索結果の詳細は、付録-5の表の右端列に示した。2014年以降に発行された文献は、研究シーズ用語176個中、89個（51%）であったことが付録-5の右端列の結果が0件である研究シーズ用語を数えた結果から確認できる（図22)。このように、抽出で得られた新奇な研究シーズ用語を文献の発行年代を過去に指定して検索した文献レコード数により、本研究の情報フィルタリングにより、時期的な意味における初期の研究段階の情報が得られることがわかった。

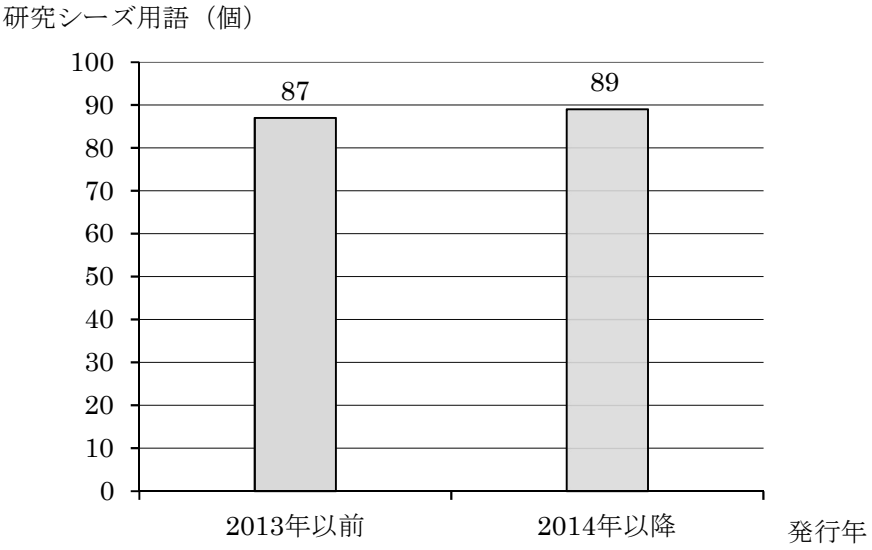

図 22. 最初の文献発行年による研究シーズ用語の研究開始年代の確認

4.2 新奇な研究シーズ用語の抽出における情報フィルタリングの役割

本研究で新奇な研究シーズ用語の抽出に用いた情報フィルタリングの役割について考察する。

4.2.1 既知用語辞書

研究シーズ用語抽出の抽出段階(1)の既知用語辞書の作成について、既知の範囲が異なる辞書による実験について考察する。

既知用語辞書を作成するためには、探索主題および目的に応じて「既知」の範囲を定義することが必要である。「既知」には、「すでに知られていること、知っていること」などの意味があるが、既知用語辞書を作成するためには、どのような範囲の人がどの程度知っている用語であるかについて定義する必要がある。さらに、

・その用語について単独で知られているのか、もしくは、他の概念の知識との関係性も含めて知られているのか、

・得られた新奇な研究シーズを使用して実施しようとする研究テーマの場合は、実際に競合する可能性のある研究の内容や研究者、研究機関の範囲をどのように想定しているのか、

・どのような目的を達成するために得られた研究シーズを使用するか、

などにも関係してくる。そのため、既知用語辞書を作成するために、探索目的の達成に必要な適切な既知の範囲を十分に検討し、その内容や目的に則した範囲を設定する必要がある。

本研究においては、探索目的が情報フィルタリングを使用して新奇な研究シーズ用語を抽出することにある。抽出した薬理メカニズム用語を使用した薬剤のスクリーニング試験をおこない、創薬に必要な薬剤を発見することが抽出した薬理メカニズム用語の使用目的である。そのため、どのような範囲の人が知っている用語を既知とするかに関しては、医薬品として既に発売されている薬剤に関する薬理メカニズム情報を既知とする方法や、創薬に係わる多くの研究者が知る可能性が高い整理された形で公開されている医薬品の研究開発情報データベースに収録された薬理メカニズム情報を既知と定義する方法が考えられる。市販されている薬剤や臨床研究が進められている薬剤は明確な情報が得られるため、既知の基準として設定することは可能である。しかし、非臨床段階の実験情報については、研究

の明確な進行度合いについての定義がなく、医薬品の研究開発情報データベースの作成会社により新規収載の時期についての方針も異なる。そのため、本研究ではとくに研究開発の進行度についての制限を設定することなく、開発の中止、中断を含めて代表的な医薬品の研究開発情報データベースである Pharmaprojects、Integrity、明日の新薬に収載された情報を使用することにした。

また、既知の範囲として研究ニーズである疾病の種類にはかかわらず、その存在が知られているすべての薬理メカニズムを既知とする場合と、設定した特定の疾病や疾病領域に直接関係する薬理メカニズムだけを既知とする場合と、大きく二通りが考えられる。そのため、既知用語辞書を作成する方法として、すべての疾病を対象として薬理メカニズム用語を収集して作成した既知用語辞書と特定の疾病領域に限定した薬理メカニズム用語を収集して作成した既知用語辞書を用いた二つの実験をおこなった。

既知の範囲に全疾病領域の疾病を含める場合は、情報フィルタリングによる絞込み効果は大きくなるが、調査目的の疾病に対してはまだ報告が少ない薬理メカニズムの場合は、既知用語辞書を使用した照合によって除去される可能性はある。しかし、実際の医療では、既知の薬理メカニズムにより作用するとして知られている薬剤でも、他の疾病において効果のあることが発見され、医薬品となる事例もある[90]。そのため、この全疾病領域を対象とした既知用語辞書を使用する場合は、探索目的の疾病に特異的な薬理メカニズムや新しく発見された薬理メカニズムを探索したい場合に適していると考えられる。breast cancer に関係する標的分子についてこの疾病に特異的な研究シーズを探索する目的でこの全疾病領域を対象とした既知用語辞書を使用して実験をおこなった。その結果、既知情報として除去される情報は多くなり、2,974 個の索引から 167 個（5.6%）の研究シーズ用語を含む索引が得られ、絞込みの効果は大きくなった。そのため、情報フィルタリングの結果から研究シーズ用語を選別するための情報負荷を大きく軽減することになり、選別ルールを用いた手動による方法によって 28 個の研究シーズ用語を抽出することができた。

一方、既知の範囲を特定の疾病に限定して作成した場合は、調査目的以外の疾病においては既知の薬理メカニズムであっても、既知用語辞書を使用した照合により除去されないため、新奇な研究シーズ情報として残り、調査目的の疾病に対して適応する可能性を検討できることになる。疾病領域を限定した既知用語辞書を使用する方法は、より広く研究シーズ用語を抽出するためには効果的であると考えられる。たとえば、Carperitide は心不全の治療薬として 1995 年の発売以降、多く使用されているが、近年、Carperitide が有する E-selectin

の発現抑制作用が手術後の肺癌の再発を有意に軽減することが発見され、ドラッグリポジショニング（既存薬再開発）として臨床試験が進められている。このように、他の疾病で知られている既知の薬理メカニズムであっても、探索目的の疾病に使用例がなければ広く収集することを目的とする探索が必要な場合もある。しかし、これにより既知用語辞書とデータセットにおいて照合される情報は限定され、その結果として既知用語辞書とデータセットによる照合により除去されずに新奇な研究シーズ用語として残る情報が多くなることが考えられる。そのため、この疾病領域を限定した既知用語辞書を使用する場合は、広く薬理メカニズムを探索したい場合に適していると考えられる。lung cancer は、他の部位のがんが転移する場合も多いため、広く研究シーズを探索する目的で特定の疾病に限定した既知用語辞書を使用した実験をおこなった。その結果、1,375 個の索引から研究シーズ用語を含む索引が 739 個（53.7%）得られ、ストップワードによる分割と選別ルールにより 8,058 個の研究シーズ用語が選別され、絞り込みの効果は少なかった。

そのため、得られた研究シーズ用語の PubMED における文献レコード数の df 値を用いた STF-IDF 値を用いて出現頻度により順位付けする方法を検討し、抽出する範囲の設定については、探索者が裁量をもつことができるようにした。この順位付けの方法により、既知の範囲を限定した辞書を用いた場合にも、情報フィルタリングで得られたある程度の大量の情報を処理することが可能になった。また、この方法では、既によく知られている医薬品の薬理メカニズムなどの文献レコード数が多い研究シーズ用語は、df 値が高くなるために STF-IDF 値が小さくなることが考えられる。これらを探索する場合は疾病名と研究シーズ用語を使用して PubMED の文献レコード数である df 値を求める方法も考えられるが、今後の検討課題としたい。

4.2.2 照合方法

抽出段階(2)のデータセットと既知用語辞書との照合について、breast cancer と lung cancer では異なる照合方法を用いて実験をおこなった。

breast cancer の実験では、照合方法として薬理メカニズム用語の複合名詞を構成する単語が対象範囲にすべて含まれている場合にヒットする方法として、論理積による照合方法を使用した。IT 索引は異なる主題ごとに比較的短くまとめられているため、その範囲内においては論理積を使用した照合が再現性においては有利となる。また、照合対象については、

比較実験をした結果、IT 索引を構成する統制語とテキスト説明句の部分とを合わせた範囲を使用した場合に最も絞り込みの効果が高く、2,974 個の IT 索引中、807 個（94%）と多くの IT 索引がヒットした。そのため、この範囲を対象にした結果を使用した。

一方、CA の IT 索引は主題内容ごとに独立して作成されている[83]ことが特徴であるが、テキスト説明句は、自然言語で記載されるため、薬理メカニズム以外の単語にヒットして誤検知となる可能性も無視できないと考えられる。そのため、lung cancer の実験ではこれらの点を改善する方法として近接演算子の機能を使用する照合により、細かい照合条件の調整をした精度を重視した照合処理をおこなった。その結果、IT 索引 1,375 個のテキスト説明句の部分のみを対象に近接演算子を使用する照合をおこない、636 個（46%）の IT 索引がヒットした。

近接演算子の機能は、照合する複合名詞の構成単語間に 2 個以内の単語を許容する設定とした。近接演算子を使用すると、前後の検索語は一つの IT 索引内の一定の範囲内に使用されている場合にヒットする。それに対して、論理積を使用すると、一つの IT 索引内であれば位置とは関係なく二つの検索語が使用されていればヒットする。このため、近接演算子を使用すると、複合名詞が精度よくヒットすることになるがヒットする割合は低くなる。この近接演算子を使用した照合方法は、照合されたパターンとしてヒットした用語の情報を保存できるため、これらの情報を照合後に確認することにより、新奇な情報が含まれていないことが確認できる。

本研究では、CA の IT 索引を構成するテキスト説明句の特徴を使用して情報の収集をおこなった。テキスト説明句には文献中に記載されたセンテンスとして文献タイトルなどが使用される場合もある。そのため、CA 以外の文献データベースとして MEDLINE を使用し、文献タイトルやセンテンス単位に限定した情報フィルタリングがおこなえる可能性もある。この場合は、データセットを作成するための文献検索の際に、検索対象をセンテンス中に限定する近接演算子である“(S)”を使用する必要がある。また、出力の際に KWIC（keyword in context）形式を指定するもしくは抄録を出力後にセンテンス単位に分解して、ヒットしたセンテンスを収集するなどの検討が必要になる。

自然言語を対象としたテキストマイニング処理は、不確定要素が多く最適な条件を設定することは難しいが、個々の処理結果を確認することにより、全体としてのプロセスを最適化して、よりよい結果を得ていくことが可能である。また、ライフサイエンス分野の文献のテキストマイニングの研究は抄録を対象とした研究が多いが、抄録から得られる情報だけ

では目的とする情報が得られないと主張する研究もある[91][92]。本研究においては、多くの主題に関係する重要な情報を含んでいる IT 索引から情報を抽出する方法を検討し、再現性を重視した論理積による照合方法および精度を重視した近接演算子による照合方法を提案した。とくに、近接演算子による方法は、用語辞書として CSV ファイルを使用できることや照合に使用する複合名詞間に含まれる単語数を自由に設定できるなど、汎用性の面で有利であると考える。

4.2.3 研究シーズ用語の抽出方法と選別ルールの妥当性

本項では、提案した新奇な研究シーズ用語の候補を抽出する方法や選別ルールの妥当性について考察する。

4.2.3.1 既知情報として除去された情報の確認

データセットと既知用語辞書との照合において、negative profile 方式により除去された既知情報に関しても、内容を確認することが情報フィルタリングの方法を評価するために必要である。すなわち、情報フィルタリングでは検知漏れが多いことが問題になると考えられるため、照合でヒットした情報、すなわち除去する情報について確認した。既知の研究シーズ用語が正しく含まれていることを確認できれば、新奇な研究シーズ用語の検知漏れが少ないと考えられる。疾病の範囲を限定した既知用語辞書を使用した lung cancer の実験では、既知用語辞書によるデータセットとの照合により、1,375 個の IT 索引のうち 636 個（46.3%）を既知の薬理メカニズム関連語が含まれる IT 索引として分離した（図 23）。

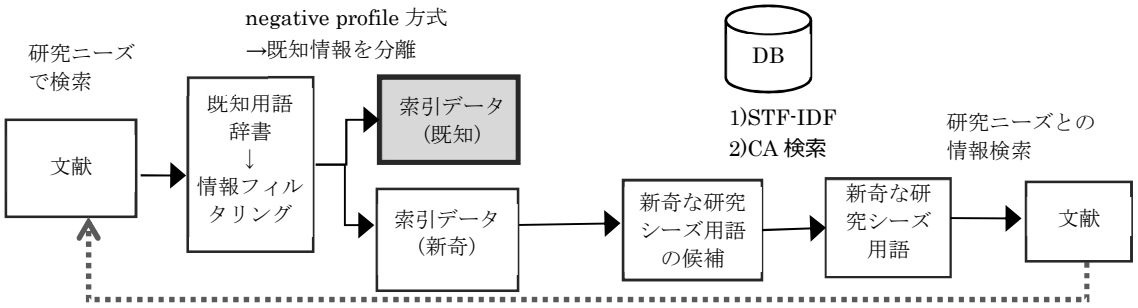

図 23. 情報フィルタリングの過程

データセットの中の一つのデータレコードの要素であるテキスト説明句に複数の既知用語辞書の用語（同義語など）がヒットする場合があり、それらを含めるとヒットした延べの索引は 1,353 個となり、1 件のデータレコードに対して平均 2.1 個ヒットしていた。各データレコードのテキスト説明句において、複数の薬理メカニズム用語がヒットしている場合は、既知用語辞書により付与された薬理メカニズム用語のなかで最も適切と思われるものを選択した。このデータを薬理メカニズム用語とテキスト説明句中におけるヒットパターンで集計した。既知の薬理メカニズム用語として、49 種類が得られた（付録-6 の表）。一部を表 25 に示す。

表 25. 既知用語辞書でヒットした薬理メカニズム用語の例

既知用語辞書の薬理メカニズム用語	既知用語辞書とマッチングしたパターン	用語数
AKT inhibitor	AKT inhibitor	3
	AKT/mTOR inhibitors	3
	AKT-mTOR inhibitors	1
anaplastic lymphoma receptor tyrosine kinase inhibitor	ALK inhibitor	19
	ALK- inhibitor	1
	ALK inhibitors	39
angiogenesis inhibitor	angiogenesis inhibitor	1
	angiogenesis inhibitors	7
	inhibitors of angiogenesis	1
apoptosis inducer	apoptosis inducers	1
apoptosis regulator Bcl-2 inhibitor	BCL-2 family inhibitors	1

データセットと既知用語辞書との照合によりヒットした辞書の中の表記を確認すると、同じ薬理メカニズム用語（aurora kinase A inhibitor）を含む用語に関して 2 パターンの 5 個の表記で誤ヒットが認められた。これは、薬理メカニズムの辞書パターンとして 1 字の “a”の表記に関して、単語境界を示す正規表現の指定（¥b）を前後に入れなかったことが原

因であった。これにより、複数の“a”を含む用語にヒットすることになったが、これに関しては、¥ba¥b と正規表現を修正することにより対応可能である。

また、既知用語辞書の 4 種の用語においてはヒットした 21 個の用語に関して、ヒットパターンの中に標的分子の異なるサブタイプ（CDK4/6 inhibitors）や異なる標的分子がスラッシュ“/”などを挟んで併記されているもの（telomerase/topoisomerase inhibitors）が認められた。これらは、用語を構成する数字や単語の前後にスペースがない場合が多く、正規表現により区別することは困難であり、今後の課題とする。また、1 個の薬理メカニズム用語（proto-oncogene tyrosine-protein kinase receptor Ret inhibitor）に関する 3 個のパターンでヒットした 13 個の索引に関しては、既知用語辞書の同義語のパターンに汎用性のある単純な表記“receptor tyrosine kinase inhibitor”が含まれていたことが原因であり、既知用語辞書からこれを削除することにより修正可能である。

誤検知の合計は、既知の薬理メカニズム用語が含まれた IT 索引 636 個の中において 39 個（6.1%）が認められ、597 個（93.9%）が適切にヒットしていた。このため、新奇な研究シーズ用語が失われる検知漏れは少ないと考えられた。

情報検索における適合性フィードバックの役割に類似して、既知用語辞書とデータセットとの照合によりヒットした用語のヒットパターンを確認することにより、既知用語辞書の修正が可能となり、より精度の高い情報フィルタリングを実施することができる。また、636 個のヒットした IT 索引が、49 種の既知の薬理メカニズム用語に整理されたことから、探索分野で既知の研究シーズとしてどのような薬理メカニズムの研究開発がどの程度実施されているかなどを調査するトレンド情報調査も可能になる。

4.2.3.2 使用頻度が低い同義語を新奇な研究シーズ用語とした可能性の確認

CA の検索において文献レコード数が少なく低頻度であった原因が、検索に使用した研究シーズ用語が文献の記述に使用されることの少ない同義語であった場合は、新奇な研究シーズ用語であるとは限らない。そのため、新奇性の確認段階における検知漏れを確認する方法として、抽出した研究シーズ用語を用いて、PubMED のシソーラスである MeSH データベースを検索し、統制語である MeSH との関係性を確認した。

PubMED を検索する際には、検索語にフィールドタグの[mesh]を指定した検索によって QueryTranslation がおこなわれる。そのため、検索に使用した用語が統制語の MeSH 用語

の同義語の場合は、MeSH 用語と PubMED における文献数が表示される。MeSH 用語に変換された場合は、検索に使用した研究シーズ用語が同義語である可能性がある。その場合は、他の同義語を含めた CA の検索によって肺がんとの関係を再調査し、正確な文献レコード数を知ることができる。

これらの同義語の確認は、探索目的の疾病領域に関してある程度の専門知識がある人がおこなう場合は、既知の薬理メカニズム用語関連の同義語であることの判断が可能であることも多いと考えられる。しかし、perl 言語のプログラムによって Web-API 検索を利用して MeSH データベースを使用することによる同義語の確認をおこなうことができれば、検知漏れに関する調査負担の軽減に有効な方法となる。本実験では、lung cancer の処理の流れと結果において最終的に 3.3.5 項で絞り込まれた 176 個の標的分子関連の用語には該当するものはなく、同義語は含まれていなかった。その他の新奇ではないとして除去した用語には表 26 に示す変換例が認められ、この方法により同義語が確認できることがわかった。

表 26. MeSH データベースを使用した同義語変換例

研究シーズ用語	文献レコード数		MeSH 用語への変換例
	変換前	変換後	
CDKN1B	2,239	5,272	cyclin-dependent kinase inhibitor p27[MeSH Terms]
NF kappa B p65	215	4,803	transcription factor rela[MeSH Terms]
HIF 1 alpha	723	10,001	hypoxia-inducible factor 1, alpha subunit[MeSH Terms]
IGF1R	1,351	5,400	Receptor, igf type 1[MeSH Terms]
MAPK14	192	582	mitogen-activated protein kinase 14[MeSH Terms]

4.2.3.3 新奇な研究シーズ用語の属性の適切性

情報フィルタリングの実験により抽出された用語を確認し、情報フィルタリングの方法や選別ルールの妥当性を評価するために、個々の用語の属性調査を実施し、薬理メカニズム以外の用語がどの程度抽出されているかについて確認した。

疾病の範囲を限定した既知用語辞書を使用した lung cancer の治療薬の薬理メカニズムを探索する実験において得られた 176 個の新奇な研究シーズ用語を使用して CA を検索し、得られたレコードの索引と抄録中の記述を確認することにより、薬理メカニズム関連語の属性を確認した（付録-7 の表）。その結果、176 個の用語の中において 117 個（66.5%）が

標的分子を含む薬理メカニズム関連の用語（リストの内容の項目に薬理メカニズムと記載）であることがわかった（図 24）。

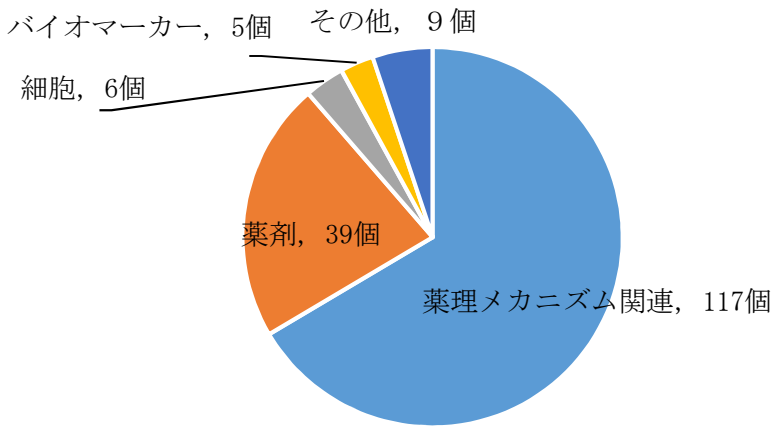

図 24．新奇な研究シーズ用語の属性

また、薬剤の名称が 39 個（22.2%）、細胞関連用語 6 個（3.4%）、バイオマーカー関連用語 5 個（2.8%）、その他の用語 9 個（5.1%）が含まれていた。薬剤の名称には大文字だけで構成される略語も多いため、標的分子の選別ルール（大文字を含む）により薬剤の名称が選別されたと考えられる。また、大文字を含むことが多い細胞関連用語やバイオマーカー関連用語も、11 個と量的には少なかったが選別されていた。薬剤などの用語がある程度含まれていたが、実際に研究シーズの候補を選択する段階で確認できるため、再現性を重視する研究シーズ用語の抽出方法としては適切であると考える。

4.2.3.4 研究シーズ用語の順位付け

研究シーズ用語抽出の抽出段階(4)では、情報フィルタリングにより得られた大量の研究シーズ用語を出現頻度により順位付けした。情報フィルタリングの条件によっては大量の情報が得られるため、得られた用語を順位付けすることができれば用語を抽出する目安として効果的であると考えた。本研究では lung cancer の実験により得られた多くの情報の中から、文献数に基づいて用語を区別するために、用語の重みづけに使用される関数である TF-IDF を応用した STF-IDF による順位付けをおこなった。TF-IDF は、本来は特定の文章中における当該用語の重み付けを計算する経験則であり、TF は特定の文書中の用語頻度を使用して、その文書における当該用語の量的要約として見なされる。しかし、本実験においては、TF を STF として、文献検索によって得られた 1 年分の全 IT 索引中における頻度を使用した。そのため STF 値として比較的大きいものが含まれる傾向があった。さらに、CA の索引語の特徴として、1 文献レコードにおいて、異なる主題内容の統制語（Subject Index）や異なる物質の統制語（Substance Index）ごとに多数の IT 索引が付与され、統制語の文献中における役割を説明するために同じテキスト説明句が何回も使用される場合も多い。そのため、テキスト説明句中の用語の頻度を使用して、単純に量的な評価をすることは適切ではない可能性がある。実際に STF-IDF 値を計算すると STF-IDF 値は高いが、PubMED の検索レコード数である df 値が大きく、新奇性の面で順位が適切ではないと思われる用語が確認された。すなわち、PubMED において数十から数百など大きい文献レコード数が認められるような、df 値が高くよく知られている用語の場合にも、STF 値の高さが原因で STF-IDF 値が高くなり、上位にランクされる場合があることが確認された。

今回おこなった lung cancer の実験では、PubMED における文献レコード数である df 値が比較的小さくあまり知られていない用語の抽出を優先するために、df 値の閾値として 200 を設定し、df 値が 200 未満の用語に限定して順位付けの処理をおこなった。また、stf 値は補正項の使用を検討し、lung cancer の実験に関しては 30 を stf 値に加算した値を計算に使用することにした。このことにより、stf 値の影響を残しつつ、文献データベースで出現頻度が低い用語の順位付けを上位にすることができた。

このように、情報フィルタリングによる情報の絞り込みが十分ではない場合は、STF-IDF 値を順序尺度として使用し、順位付けする方法を使用できることが確認された。また、実際にこれらの研究シーズ用語と疾病名で CA を検索した確認実験では、図 21 に示したとおり、

STF-IDF の数値が高い研究シーズ用語は、CA の収録の全期間における文献レコード数についても少なくなる傾向があることがわかった。

4.3 新奇な研究シーズ用語抽出の意義と展望

4.3.1 情報フィルタリングの役割

Web 上の情報の増加により、利用者が必要とする情報を効率よく収集できる仕組みが求められ、開発研究がおこなわれている。たとえば、検索対象を統計的に処理して絞り込みをおこなう情報フィルタリングシステムが PubMED で実装されている[89]。しかし、これらは既知情報による絞り込みには適しているが、未知の概念には適用できない。新たな研究をはじめるには、新奇な情報が必要になる場合が多い。とくに製品開発では、他で実施されていない研究が重要であり、特許出願により初めて事業化される。いまだ顕在化していない潜在ニーズを明らかにし、新しくニーズを創造することがイノベーションにつながり、さらにそれらを新奇な研究シーズによって実現することにより、価値ある成果が期待できる（図 25）。

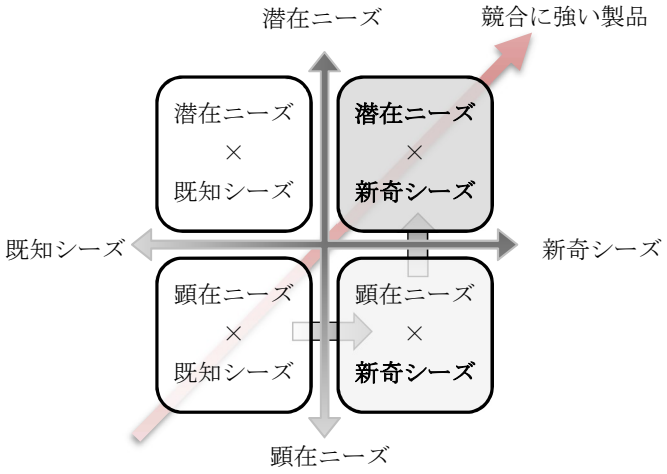

図 25. 探索目的とする情報

図 25 では、横軸の上が潜在ニーズで明らかになっていないニーズ、下が顕在ニーズで既に明らかなニーズを示す。縦軸の左が既知シーズでよく知られている技術を示し、右が新奇シーズでよく知られていない技術を示す。左下の第三象限のよく知られている顕在ニーズから情報フィルタリングにより得られた新しい新奇シーズにより右下の第四象限に移動して差別化を図ることができる。さらに新奇シーズを検索に使用してその結果の共起語や類推により右上の第一象限に該当する潜在ニーズや新奇なシーズの組み合わせを発見するこ

とができれば、競合に強いイノベーションが可能になる。このように、当初の情報要求が変化し、新たな知識発見が可能になることが探索行動の特徴でもあり[93]、情報フィルタリングの最終的な目標と考える。

小さな変化がある時点に達すると一気に広まる大きな変化を生む場合があり[94]、この小さく稀な変化を予兆として捉えて「チャンス発見」を可能にするために、適切に情報を収集し分析することが重要と考える。本研究における特定のニーズと共起する新奇な研究シーズ用語のリストは、このような意味から実際の研究開発に貢献できる可能性を有していると考える。少なくとも、セレンディピティの役割でもある「研究者がすでに抱えている問題の概念化または解決策の強化」[95][96]のための多様な解決策を提示する方法の一つである。本方法により得られた新奇な情報のリストが研究者の発想を刺激し、発想的推論によってイノベーティブなアイデアが生まれることを期待する[67]。とくに、創薬の場合は、疾病の原因に関係する標的となるタンパク質を決めることが、研究テーマの基本であり[97][98]、本研究で作成したリストが役立つと考える。

医薬品の研究では、研究者の様々な背景を理解し、膨大な情報の中から研究シーズ情報を探索し、有望で新奇性、独創性の高い情報を入手することが求められている[99]。情報支援の対象となる組織が有する独自のコンテキストを理解し、必要な情報の取捨選択をおこなう「ゲートキーパー」の機能が重要であり[100]、本研究で提案する情報フィルタリングシステムが有効となる。

4.3.2 処理の部分的自動化による人の判断への支援

本研究の情報フィルタリングでは、処理の各過程において部分的な自動処理をおこなった。大量の情報を処理する必要がある場合でも、調査主題に関して専門知識がある探索者であれば新奇な情報をある程度、的確に選別することが可能である。しかし、新しく研究を始める場合など、その領域の知識が十分ではない場合は、的確な情報の選別は難しい。そのため、人工知能の情報源[101]としても使用されている文献データベースの情報と連動させることにより、判断を支援できると考えた。医師を含む専門家でも、大量のデータの前では簡単に知識のパターンを見つけることは難しいとの指摘もある[44]。本研究が提案した部分的に自動化された分析方法によって、文献データベースから専門家による知識発見が比較的低いコストで促進できると期待する。

情報探索の重要なプロセスとして、人が判断して選択するための複数の解決につながる選択肢を適切に作成することがある。しかし、探索的な調査では網羅的な調査が難しい場合もあり、よいアイデアが見つかるとそれ以上の調査はおこなわない傾向がある[102]。情報フィルタリングにより大量の情報を半自動的に処理することで人が判断すべき重要な部分に集中して、適切な「キュレーション」[103]が可能になることが本方法の重要な利用目的でもあり、一つの方法を提案できたと考える。

4.3.3 探索的情報検索の評価方法による本方法の評価

本研究で提案した情報フィルタリングは、準備段階と抽出段階の各処理過程において探索目的に応じてそれぞれの条件を設定し、研究シーズ用語抽出の部分的自動化による検知漏れを少なくするなど処理全体として最も良い結果が得られることを目指している。このように、探索的に調査をするためには各種の処理条件の調整が必要であり、ここでは、White らによる探索的情報検索の 8 個の必要条件[104]を本方法の各処理過程にあてはめ、本研究が提案する方法の評価を試みた。探索的情報検索の 8 個の必要条件を本研究の方法に当てはめて評価した結果を表 27 に示す。

表 27. 探索的情報検索の必要条件と判定結果

必要条件	概要	判定結果
1 : クエリ修正と再検索の支援	検索式を検索結果の確認により改善し、再検索をおこなうための支援機能	考慮している
2 : 情報フィルタリング機能	必要な情報のプロフィルを定義して、半自動的に情報を絞り込む機能	考慮している
3 : コンテキストの利用	探索の背景や問題点などの情報を収集して検索に活かす機能	考慮している
4 : 結果の可視化による発見支援	情報の可視化により、必要な情報の人による発見を支援する機能	考慮している
5 : 学習支援	調査で得られた知識を学習して必要な情報を明確にする機能	考慮している
6 : 協調検索作業	探索者と調査担当者が協調して効率的な調査をおこなう機能	考慮している
7 : 履歴提示	検索履歴を管理し、検索式の確認や未調査の部分を明らかにする機能	考慮されていない点がある
8 : タスク管理	検索結果をタスク単位に保存して、検索内容の確認や再検索を支援する機能	考慮されていない

本方法では情報要求者と情報担当者とが協働して調査する必要があることから、クエリ修正や再検索の実施に有利である。また、情報担当者が同じ組織に所属する場合は、調査目的の領域に関するある程度の知識をもって調査するため、より有効な支援が可能になり、必要条件 1 を考慮している。

情報フィルタリング機能は、探索調査に有効な本研究の核となる機能である。文献量が多いライフサイエンス分野においては、情報フィルタリングによる不要な情報の除去によって、内容確認による情報過負荷の低減が期待できる。そのため、必要条件 2 は考慮している。

情報要求者と背景を共有する情報担当者が協働して探索調査することにより、コンテキストを意識した調査戦略が可能となる。本研究では、処理の部分的自動化と人の分析により、コンテキストを意識した処理が可能になるため、必要条件 3 は考慮している。

本研究では新奇な情報に限定したリストの作成が可能であるため、結果の可視化に有効である。また、既知の情報についても、既知用語辞書でヒットした情報に整理し、集計することにより、グラフ化が可能である。このため、結果の可視化に関する必要条件 4 は考慮している。

本研究で作成した新奇な研究シーズ用語リストから、情報要求者が探索主題に関する新しい概念を学習し、文献検索をおこなうことにより、より発展的な調査が期待できる。そのため、学習支援に関する必要条件 5 は考慮している。

本研究では、情報処理結果を情報要求者が見ることにより、適合性フィードバックとして、既知用語辞書に用語を追加できるため、協調的な処理が可能である。そのため、必要条件 6 は考慮している。

本研究では、個々の処理単位で結果のファイルを保存する設計になっている。そのため、結果の確認が必用な場合にも、途中の処理結果を確認して分析方法を修正することが可能であるが、全体の処理履歴の自動記録機能はない。そのため、必要条件 7 は部分的に考慮している。

本研究では、必要条件 8 のタスク管理の自動化の機能は考慮されていない。

以上のように、本研究の方法は必要条件 1 から 6 の条件を考慮しているため、探索的な調査に有効な方法であると考える。一方、必要条件 7 の「履歴提示」、必要条件 8 の「タスク管理」については条件を満たしていない。また、条件を考慮している項目についても「結果の自動視覚化」や「データ処理の自動化」などが今後の課題であり、「タスク管理機能」

「既知用語辞書へのフィードバック」も必要であると考える。情報フィルタリングの機能を取り入れた探索方法の必要性は高いと考えるが、実際に使用するには処理全体の過程を統合化し、利用者が使いやすくする必要がある[105]。本研究で提案する方法にこれらの機能を付加していくことにより、より改善された信頼性の高い探索方法にすることが期待できる。

4.3.4 研究シーズ用語のその他の疾病に対する新奇性

lung cancer の実験で得られた新奇な研究シーズ用語について、肺がん以外の疾病に対する新奇性を確認した。3.3.1 項において、肺がん関連の検索語（lung cancer、lung neoplasm、lung carcinoma）を用いて検索された 630 件の文献には、肺がん以外の疾病名も統制語として付与されているものがあった。CA ではテキスト説明句が同じ場合は、統制語の部分に関連する疾病名がまとめて記載されているため、統制語からがん関連の疾病名を切り出した。得られたがん関連の疾病名と抽出で得られた新奇な研究シーズ用語とのすべての組み合わせについて、論理積による CA 検索式を作成し、STN システムを使用して CA を検索した（2016 年 7 月 30 日検索）。その結果の一部を表 28 に示す。

表 28. がん関連の疾病名と研究シーズ用語との検索結果の文献レコード数

疾病名に含まれるがん発症の部位 ＼ 研究シーズ用語	Lung	Non-small-cell lung	Mammary	Colorectal	Colon	Pancreatic	Prostate	Ovary	Hepatocellular
SMYD3 inhibitor	1	0	2	2	0	1	2	1	1
AKT1 PDPK1	1	1	0	2	0	0	1	0	0
AZD9150	1	1	1	0	1	1	1	1	0
SKLB 677	1	1	1	0	1	0	0	0	0
CSC 3436	1	1	0	0	0	0	0	0	0
MIF rpS3	1	1	0	0	0	0	0	0	0
InsP3 Kinase	1	0	1	0	0	0	0	0	0
H1650GR	1	1	0	0	0	0	0	0	0
Betti reaction	1	0	1	0	0	0	0	0	0
alpha 3 beta 2 nAChR	1	0	0	0	0	1	0	0	0
A 893	1	0	0	0	0	0	0	0	0

肺以外の部位のがんに対しても文献レコード数が少ない傾向が認められた。このように、抽出で得られた新奇な研究シーズ用語と他の疾病名とを検索に使用することにより、当該研究シーズ用語がどのような疾病と関係しているかを推測できる。そのため、新たな研究ニーズのヒントを提供できる可能性があり、検討していきたい。

4.4 第 4 章のまとめ

実験の結果に基づいて、本研究で提案した抽出方法の文献レコード数に基づく新奇性の確認、新奇な研究シーズ用語の抽出における情報フィルタリングの役割、新奇な研究シーズ用語抽出の意義と展望の三つの点について考察した。

4.1 節では、文献レコード数に基づく新奇性の確認について述べた。第一に、breast cancer の実験により得られた研究シーズ用語の新奇性を確認するため、疾病名と共に抽出した研究シーズ用語を用いて CA を検索し、得られた文献レコード数を評価に用いた。その結果、breast cancer の実験で抽出した研究シーズ用語の文献レコード数が少ないことが確認できた。さらに、研究シーズ用語全体と比較した新奇性の評価をおこなった。この結果、提案した抽出方法により新奇な研究シーズ用語が得られていることがわかった。第二に、lung cancer の実験により得られた研究シーズ用語の新奇性を確認するために、疾病名と共に抽出した研究シーズ用語を用いて CA を検索し、得られた文献レコード数を評価に用いた。その結果、lung cancer の実験で抽出した研究シーズ用語の文献レコード数が少ないことが確認できた。さらに、抽出で得られた新奇な研究シーズ用語を文献の発行年代を区切った検索による新奇性の評価をおこなった。その結果、時期的な意味においても初期の研究段階の情報が多く得られることがわかった。

4.2 節では、新奇な研究シーズ用語の抽出における情報フィルタリングの役割について述べた。第一に、既知の範囲が異なる辞書について考察した。探索目的に適した範囲の既知用語辞書を使用して近接演算子の機能により照合することにより大量の情報を処理できることを確認した。第二に、照合方法について考察し、論理積を使用する方法と近接演算子を使用する方法の相違について述べた。第三に、索引からの研究シーズ用語の抽出方法と選別ルールの妥当性について、既知情報として除去された情報の確認をおこない、多くが適切にヒットしているため、新奇な研究シーズ用語の検知漏れは少ないと判断した。第四に、使用頻度が低い同義語を新奇な研究シーズ用語とした可能性の確認をおこない、抽出した研究シーズ用語に既知用語の同義語が含まれていないことがわかった。第五に、新奇な研究シーズ用語の属性を確認し、薬理メカニズム用語が多く得られていることがわかった。第六に、研究シーズ用語の順位付けについて、新奇な研究シーズ用語を選択するために、STF-IDF 値を順序尺度として順位付けする方法が使用できることを述べた。

4.3 節では、新奇な研究シーズ用語抽出の意義と展望について述べた。第一に、情報フィ

ルタリングの役割は知識発見にあることから、新奇な研究シーズ用語の抽出が可能になれば、イノベーションを支援することが期待できることを述べた。第二に、処理の部分的自動化による人の判断への支援について、研究シーズ用語抽出の処理の部分的自動化は、大量の情報から比較的低コストで情報を抽出することが可能であること、さらに人の判断への支援をも担っていることを述べた。第三に、探索的情報検索の評価方法による本方法の評価について、探索的情報検索の評価方法を使用した結果、履歴提示、タスク管理、結果の可視化などの機能を追加することにより、本研究が提案する抽出方法はさらに使い易い方法になることを述べた。第四に、研究シーズ用語のその他の疾病に対する新奇性について、lung cancer の実験で得られた新奇な研究シーズ用語を使用して肺がん以外の疾病に対する新奇性を確認し、当該研究シーズ用語がどのような疾病と関係しているかを推測することにより、新たな研究ニーズを提供できる可能性について述べた。

第 5 章 結論

本研究の目的は、文献データベースを使用して新奇な研究シーズ用語の候補を抽出する方法を提案し、抽出実験によってその妥当性を明らかにすることである。そのために、研究課題 1 として新奇な研究シーズ用語の候補を抽出する方法を提案すること、さらに研究課題 2 として抽出実験により提案した方法の妥当性を確認すること、の二つを設定し、これらを解決することにより、研究の目的を達成することにした。研究課題 2 で設定した妥当性については、研究シーズ用語の候補を抽出し、その新奇性を文献レコード数により確認することで判断することにした。その結果、設定した疾病用語と本研究により提案した方法により抽出した研究シーズ用語とを使用した文献検索により得られた文献レコード数が少なかったことから、本研究が提案する方法により新奇な研究シーズ用語を抽出できることの妥当性が確認された。

本研究では、抽出の方法を 4 抽出段階に分けて検討した。抽出段階(1)では医薬品の研究開発情報データベースから薬理メカニズム用語を参照して既知用語辞書を作成し、抽出段階(2)ではデータセットと既知用語辞書との照合によりヒットした IT 索引をデータセットから除去することにより、研究シーズ用語が含まれる IT 索引を絞り込み、抽出段階(3)では選別ルールを使用して、IT 索引から関数もしくは手動により新奇な薬理メカニズム用語を選別し、抽出段階(4)では選別件数が多い場合の処理について、STF-IDF 値で降順にソートすることにより、研究シーズ用語を削減できることを示した。とくに、抽出段階(2)、抽出段階(3)、抽出段階(4)については処理の自動化が必要な過程があるため、専用プログラムを作成して対応した。

得られた新奇な研究シーズ用語の候補と疾病名とが、一つの IT 索引にある CA の全期間における文献レコード数を調べた。その結果、文献レコード数が少ないことにより疾病との関係において新奇な研究シーズ用語であることが確認できた。さらに、設定した疾病に関して作用語のみで検索したその他の研究シーズ用語の文献レコード数と比較し、低頻度であることを確認した。また文献発行年における文献レコード数を比較し、年代的にも最近の研究であることがわかった。以上の方法により調べた結果、目的とする新奇な用語を含む可能性が高い情報が得られたことを確認した。

また、抽出で得られた新奇な研究シーズ用語について、肺がん以外の疾病に対する新奇性を確認し、新たな研究ニーズを提供できる可能性について示した。

本研究では、抽出の対象であるデータセットの情報源には調査主題に関す情報を含み、他の文献データベースと比較して多様で詳細な索引情報をもつ CA を使用した。しかし、MEDLINE のような他のデータベースを使用しても、文献タイトルやセンテンス単位に限定した情報フィルタリングがおこなえる可能性があり、情報源拡大の検討が必要であると考える。情報源を変更する場合には、データベースの特徴を考慮して、データセットの作成方法の調整が必要となる。

また、本方法は、対象とした疾病が breast cancer と lung cancer に限られているため、他の疾病を対象とする際には細部の調整が必要であると考える。そして、この調整により創薬テーマ以外の研究シーズ用語の探索にも適応できる可能性があると考える。本方法の可能性を見極めるために、新しい主題領域への適応を検討していきたい。

謝辞

本研究を進めるために、研究の計画、実行、論文作成などの多くの過程でたくさんの皆様にご支援頂きました。ここにお名前を記し、深く感謝の意を表します。

緑川信之先生には、指導教員として親身なご指導を賜わりました。とくに、博士論文の完成にいたるまでお忙しい中、多くの時間を割いてくださいまして、厳しいながらも詳細かつ的確なご指摘を頂きましたことは貴重な経験となりました。芳鐘冬樹先生には、副指導教員をお引き受け頂き、異なる視点から貴重なご意見を賜わりました。中山伸一先生には、副指導教員をお引き受け頂き、研究手法についての貴重なご意見を賜わりました。

岩澤まり子先生には、ご在職中には指導教員として、またご退職後も引き続き、本研究全般にわたり常に力強い励ましと熱心なご指導を頂きました。

緑川先生および岩澤先生のゼミに参加されました大学院、社会人の皆様には、私の研究について議論、指摘を頂きました。同時に、他の多くの方々の研究活動について拝見できましたことは、研究を進める上での大きな励みになりました。感謝申し上げます。

文献リスト

[1] 菰田文男ほか. 技術と市場ニーズの探索・融合. 税務経理協会, 2007, 187p.

[2] 石川昭; 辻本篤. 新製品・新事業開発の創造的マーケティング: 開発情報探索のマネジメント. 生産性出版, 2006, 262p.

[3] 出川通. 技術経営の考え方: MOT と開発ベンチャーの現場から. 光文社, 2004, 191p.

[4] グローバルタスクフォース. MOT: テクノロジーマネジメント. 総合法令出版, 2004, 223p.

[5] 星野達也. 研究開発効率化によるメガファーマへの挑戦: 製薬業界におけるオープンイノベーションの紹介. 日本薬理学雑誌. 2013, vol.142, no.4, p.184-189.

[6] 古矢修一. ネットワーク構築と積極的活用から生まれるオールジャパン創薬: オープンイノベーション戦略とネットワーク構築・活用が日本の基礎研究力を活かす「創薬」新パラダイムとなる. 日本薬理学雑誌. 2015, vol.145, no.5, p.243-249.

[7] 尾本巧; 工藤寛長. 製薬業界が生き残るために必要な業態変革: 特集 2030 年のヘルスケア. 知的資産創造. 2016, vol.24, no.3, p.38-51.

[8] 夏目やよいほか. 計算システム生物学による創薬: 分子, 構造からネットワークへ. 日本薬理学雑誌. 2017, vol.149, no.2, p91-95.

[9] 岸田和明. 情報検索の理論と技術. 勁草書房, 1998, 314p.

[10] 船戸奈美子; 五十嵐康子. 解析ツールの比較検討: STN, STN AnaVist, および SciFinder の特許, 文献, 化学物質解析機能の特徴と使い分け. 情報プロフェッショナルシンポジウム予稿集. 2006, vol.3, p.83-86.

[11] “SCIFINDER”. 化学情報協会. https://www.jaici.or.jp/SCIFINDER/index.php, (参照 2017-2-6).

[12] 小池麻子. 情報の価値化・知識化技術の実現へ向けて: テキストマイニングによる潜在的知識の発見支援. 情報処理. 2007, vol.48, no.8, p.824-829.

[13] 岸田和明. 情報検索における評価方法の変遷とその課題. 情報管理. 2011, vol.54, no.8, p.439-448.

[14] Feldman, R. et al.; 辻井潤一訳. テキストマイニングハンドブック. 東京電機大学出版局, 2010, 540p.

[15] 池田美奈子ほか. 感性テーブルを用いた生活者ニーズと研究シーズのマッチングシステム. 日本感性工学会大会予稿集. 2008, vol.10, ROMBUNNO.11G-05.

[16] 奥野弘之ほか. 企業ニーズを研究シーズに関連づける情報探索システムの試作. 人工知能学会人工知能基礎論研究会資料. 1999, no.37, p.169-174.

[17] 菰田文男. 「単語セット」の作成と進化に基づくテキストマイニング手法: MOT（技術経営）のためのテキストデータ解析を事例として. 情報管理. 2011, vol.54, no.9, p.568-578.

[18] 藤原貴典; 松浦啓克. わかりやすい研究シーズ発信方法: 研究シーズ集作成のアプローチ例. 産学連携学会大会講演予稿集. 2005, vol.3, p.23-24.

[19] “東京大学産学連携プロポーザル”. 東京大学.
http://proposal.ducr.u-tokyo.ac.jp/, (参照 2017-11-30).
産学連携を検討する提案テーマのデータベース.

[20] “筑波大学・研究シーズ検索”. 筑波大学.
https://www.seeds.tsukuba.ac.jp/seeds/help/about.html, (参照 2017-7-25).
産学連携を検討する研究シーズ検索サイト.

[21] “創薬におけるオープンイノベーション: 外部連携による研究資源の活用”.
ヒューマンサイエンス振興財団.
http://www.jhsf.or.jp/paper/report/report_no78.pdf, (参照 2017-11-30).
研究資源委員会の調査報告書

[22] Schuhmacher, A. et al. Changing R&D models in research-based pharmaceutical companies. Journal of Translational Medicine. 2016, vol.14, no.1, p.1-11.

[23] 高橋さやかほか. 「オープンイノベーション」と創薬支援コンサルタント. 2015, vol.146, no.4, p.208-214.

[24] 稲垣治. イノベーションをマーケットへ: 企業の立場からアカデミアに望むもの.
臨床評価. 2013, vol.41, no.1, p.146-149.

[25] 秋元浩. 今,我々ができること,そして,今後期待されること: アカデミア創薬の課題と解決策. 日本薬理学雑誌. 2013, vol.142, no.6, p.297-303.

[26] 松本弥生. 公募型医薬品研究シーズ発掘の試み. 構造活性フォーラム講演要旨集.
2014, p.54-75.

[27] 藤田義文. 創薬研究公募 TaNeDS によるイノベーション. 日本薬理学雑誌. 2013, vol.142, no.2, p.89-95.

[28] “Co-Create Knowledge for Pharma Innovation with Takeda (COCKPI-T)”. 武田薬品工業. https://www.takeda.co.jp/research/openi/cockpit/index.html, (参照 2017-11-30).

[29] Francesca, M. et al. Measuring open innovation in the bio-pharmaceutical industry. Creativity and Innovation Management. 2015, vol.24, no.1, p.4-28.

[30] Tralau-Stewart, C. J. et al. Drug discovery: New models for industry-academic partnerships. Drug Discovery Today. 2009, vol.14, no.1, p.95-101.

[31] “Oxford University Innovation”. University of Oxford. https://innovation.ox.ac.uk/, (accessed 2017-11-30).

[32] Allen, T. J. Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the Research and Development Organization. MIT Press, 1984, 320p.

[33] 小河邦雄. PubMed と MEDLINE とその他のデータベースの比較. 薬学図書館. 2006, vol.51, no.4, p.287-298.

[34] Kwon, Y. et al. A novel evaluation measure for identifying drug targets from the biomedical literature. IPSJ Transactions on Bioinformatics. 2014, vol.7, p.16-23.

[35] 広川貴次; 美宅成樹. Web で実践: 生物学情報リテラシー. 中山書店, 2013, 190p.

[36] 小島史照ほか. インフォプロの専門知識を活用した研究テーマの立案支援: CAplus における作用機作の機械抽出と活用および化合物の解析の試み. 第二回情報プロフェッショナルシンポジウム予稿集. 2005, vol.2, p.115-119.

[37] 岡紀子; 田中章夫. Chemical Abstracts 文献を利用した有望中間体の探索方法. 情報管理. 2004, vol.47, no.3, p.175-181.

[38] “Pathway Studio”. Elsevier. http://jp.elsevier.com/online-tools/pathway-studio, (参照 2016-12-15).

[39] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999, 234p.

[40] Belkin, N. J. et al. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM. 1992, vol.35, no.12, p.29-38.

[41] Kitamura, Y. et al. Discovered rule filtering using information retrieval technique. Proceedings of International Workshop on Active Mining. 2002, p.80-84.

[42] 北研二ほか. 情報検索アルゴリズム. 共立出版, 2002, 220p.

[43] Yoon, J. W. et al. Hybrid spam filtering for mobile communication. Computers & Security. 2010, vol.29, no.4, p.446-459.

[44] 津本周作; 高林克日己. 知識発見手法により生成された知識の比較と評価. 人口知能学会誌. 2000, vol.15, no.5, p.790-797.

[45] Krauthammer, M. et al. Using BLAST for identifying gene and protein names in journal articles. Gene. 2000, vol.259, no.1, p.245-252.

[46] Fukuda, K. et al. Toward information extraction: Identifying protein names from biological papers. Proceedings of the Pacific Symposium on Biocomputing'98. 1998, p.707-718.

[47] Hanisch, D. et al. ProMiner: Rule-based protein and gene entity recognition. BMC Bioinformatics. 2005, vol.6, Suppl.1, S14.

[48] Imaichi, O. et al. A comparison of rule-based and machine learning methods for medical information extraction. International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields. 2013, p.38-42.

[49] Collier, N. et al. Extracting the names of genes and gene products with a hidden markov model. Proceedings of the 18th International Conference on Computational Linguistics. 2000, p.201-207.

[50] Kazama, J. et al. Tuning support vector machines for biomedical named entity recognition. Workshop on Natural Language Processing in the Biomedical Domain at the Association for Computational Linguistics. 2002, vol.3, p.1-8.

[51] Settles, B. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005, vol.21, no.14, p.3191-3192.

[52] Leaman, R. et al. BANNER: An executable survey of advances in biomedical named entity recognition. Proceedings of the Pacific Symposium on Biocomputing. 2008, p.652-663.

[53] Shen, D. et al. Multi-criteria-based active learning for named entity recognition. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain. 2004, p.589-596.

[54] Bruijn, B. et al. Machine-learned solutions for three stages of clinical information extraction. Journal of the American Medical Informatics Association. 2011, vol.18, no.5, p.557-562.

[55] Neves, M. et al. A survey on annotation tools for the biomedical literature. Briefings in Bioinformatics. 2012, vol.15, no.2, p.327-340.

[56] Manning, C. D.; 徳永健伸ほか訳. 情報検索の基礎. 共立出版, 2012, 476p.

[57] 辻慶太. 重要な専門用語となる新語の特定・予測. 情報知識学会誌. 2005, vol.15, no.4, p.11-22.

[58] Hisamitsu, T. et al. Extracting terms by a combination of term frequency and a measure of term representativeness. Terminology. 2000, vol.6, no.2, p.211-232.

[59] Nakagawa, H. Automatic term recognition based on statistics of compound nouns. Terminology. 2000, vol.6, no.2, p.195-210.

[60] 池尻恭介ほか. 希少性と一般性に基づいた意外性のある食材の抽出. コンピュータソフトウェア. 2014, vol.31, no.3, p.70-78.

[61] 島田敏明ほか. 単語重要度を用いた N-gram 補完手法が与える音声認識性能の調査. 研究報告音声言語情報処理. 2010, SLP-82, no.19, p.1-6.

[62] 豊田裕貴; 菰田文男. 特許情報のテキストマイニング. ミネルヴァ書房, 2011, 278p.

[63] 菰田文男; 那須川哲哉. ビッグデータを活かす: 技術戦略としてのテキストマイニング. 中央経済社, 2014, 316p.

[64] Koike, A.; Takagi, T. Knowledge discovery based on an implicit and explicit conceptual network. Journal of the American Society for Information Science and Technology. 2007, vol.58, no.1, p.51-65.

[65] 小池麻子; 高木利久. 医学生物学分野におけるシソーラスとテキストマイニング技術の開発. 実験医学. 2008, vol.26, no.7, p.1161-1166.

[66] 那須川哲哉. テキストマイニングを使う技術/作る技術: 基礎技術と適応事例から導く本質と活用法. 東京電機大学出版局, 2006, 236p.

[67] 大澤幸夫. チャンス発見の情報技術. 東京電機大学出版局, 2003, 354p.

[68] “Pharmaprojects”. Informa PLC. http://citeline.com/products/pharmaprojects/, (accessed 2016-4-26). このデータベースは研究開発中の薬剤情報を含む.

[69] “Integirty”. Clarivate Analytics. https://clarivate.jp/products/integrity/, (accessed 2018-10-30).

[70] Cohen, A. et al. A survey of current work in biomedical text mining. Briefings in Bioinformatics. 2005, vol.6, no.1, p.57-71.

[71] 福田賢一郎ほか. 医学生物学文献からの専門用語の抽出に向けて: タンパク質名の自動抽出. 情報処理学会論文誌. 1998, vol.39, no.8, p.2421-2430.

[72] 特集：乳癌治療剤市場と新薬開発の動向. NEW CURRENT. 2014, vol.25, no.18, p.2-17.

[73] 赤松弘朗; 山本信之. 肺がん（第 5 土曜特集 がん標的分子と治療開発: 現状と将来）. 医学のあゆみ. 2016, vol.258, no.5, p549-552.

[74] 土原一哉. 新薬展望 2017 第 I 部 新薬創出に向けたプラットフォーム構築: 1.データベース～がん創薬を中心に～. 医薬ジャーナル. 2017, vol.53, 増刊号, p.217-224.

[75] 岡野栄之ほか. 分子標的薬開発への新たなる挑戦. 実験医学増刊. 2009, vol.27, no.5, p.20-27.

[76] Spasic, I. et al. Text mining of cancer-related information: Review of current status and future directions. International Journal of Medical Informatics. 2014, vol.83, no.9, p.605-623.

[77] Zhu, F. et al. Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics. 2013, vol.46, no.2, p.200-211.

[78] 小河邦雄. ライフサイエンス系データベースシステムの現状と将来展望. ファルマシア. 2010, vol.46, no.12, p.1149-1154.

[79] 中西陽子. 複数のデータベースにおける抗菌物質研究に関する文献の索引語分析. 情報管理. 2003, vol.46, no.1, p.1-8.

[80] “STN システム”. 化学情報協会. http://www.jaici.or.jp/stn/stn_doc.html, (参照 2016-4-26).

[81] “CAplus/CA 基礎編 CA 文献検索(2015.7)”. 化学情報協会. https://www.jaici.or.jp/stn/pdf/text_ca.pdf, (参照 2017-2-6).

[82] 佐々木啓子. 知的情報分析による索引作成とその意義: CA 作成における特許分析を中心に. 情報管理. 2012, vol.55, no.7, p.472-480.

[83] 宮崎佐智子. Chemical Abstracts(CA)の索引方針と検索のポイント. 情報の科学と技術. 2008, vol.58, no.4, p.188-193.

[84] “明日の新薬”. テクノミック.
http://www.technomics.co.jp/database/asunoshinyaku.html, (参照 2016-12-15).
このデータベースは世界で研究開発された薬剤情報含み,とくに日本の情報に強い.

[85] “Stop Word List 2”. Onix Text Retrieval Toolkit API Reference.
http://www.lextek.com/manuals/onix/stopwords2.html, (accessed 2016-12-15).

[86] 川村敬一. Articulated Subject Index の構造的特性と機械化プログラム. 情報管理. 1981, vol.24, no.5, p.447-456.

[87] 相澤彰子. デジタル時代の日本語 デジタル化された学術文献の言語解析について. 情報の科学と技術. 2014, vol.64, no.11, p.469-474.

[88] “PubMed”. National Center for Biotechnology Information.
http://www.ncbi.nlm.nih.gov/pubmed, (accessed 2016-4-26).

[89] “A General Introduction to the E-utilities”. National Center for Biotechnology Information.
https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines_and_Requiremen, (accessed 2016-12-15).

[90] 塩村仁. ドラッグ・リポジショニングの実践. 医薬ジャーナル. 2016, vol.52, no.6, p.141-148.

[91] Dai, H. J. et al. New challenges for biological text-mining in the next decade. Journal of Computer Science and Technology. 2010, vol.25, no.1, p.169-178.

[92] Ananiadou, S. et al. Text mining and its potential applications in systems biology. Trends in Biotechnology. 2006, vol.24, no.12, p.571-579.

[93] Bates, M. J. The design of browsing and berrypicking techniques for the online search interface. Online Information Review. 1989, vol.13, no.5, p.407-424.

[94] Malcolm, G.; 高橋啓訳. ティッピング・ポイント. 飛鳥新社, 2000, 310p.

[95] 三輪眞木子. 情報行動: システム志向から利用者志向へ. 勉誠出版, 2012, 205p.

[96] Foster, A. et al. Serendipity and information seeking: An empirical study. Journal of Documentation. 2003, vol.59, no.3, p.321-340.

[97] Tamas, B. et al.; 神沼二眞訳. 薬づくりの真実. 日経 BP, 2014, 392p.

[98] “「薬学の展望とロードマップ」について”. 日本薬学会. http://www.pharm.or.jp/hotnews/archives/2010/12/post_220.html, (参照 2017-2-6).

[99] 村上知子ほか. 推薦の意外性向上のための手法とその評価. 人口知能学会論文誌. 2009, vol.24, no.5, p.428-436.

[100] 吉川日出行. サーチアーキテクチャ. ソフトバンククリエイティブ, 2007, 280p.

[101] Chen, Y. et al. IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. Clinical Therapeutics. 2016, vol.38, no.4, p.688-701.

[102] Hearst M. A.; 角谷和俊監訳. 情報検索のためのユーザインターフェース. 共立出版, 2011, 415p.

[103] 佐々木俊尚. キュレーションの時代: 「つながり」の情報革命が始まる. 筑摩書房, 2011, 314p.

[104] White, R. W.; Roth, R. A. Exploratory Search: Beyond the Query Response Paradigm. Morgan and Claypool, 2009, p.61-69.

[105] Xu M. et al. Intelligent agent systems for executive information scanning, filtering and interpretation: Perceptions and challenges. Information Processing & Management. 2011, vol.47, no.2, p.186-201.

全研究業績のリスト

査読付論文誌論文

1. 小河邦雄.　文献データベースの検索と辞書マッチングによる探索的フィルタリングの研究. 情報メディア研究. 第 15 巻, 第 1 号, 2016, pp.14-25.
2. 小河邦雄；岩澤まり子. 文献データベースの検索と辞書マッチングによる探索的フィルタリング：第二報 Web-API を利用した半自動化処理. 情報メディア研究. 第 15 巻 第 1 号, 2016, pp.26-37.

付録

付録−1

情報フィルタリングのプログラム

```
$data_file = "kensaku_1.txt";

#①データセットの作成
open(FH,"< $data_file") or die qq{can't open};
@arr=<FH>;close(FH);open(OUT, ">20chu_near_2.txt");
foreach (@arr){
s/¥n//sg;s/CC¥s{3}/¥nCC:/sg;#s/TI¥s{3}/¥tTI:/sg;s/TI¥s{3}/¥t/sg;
s/ST¥s{3}/¥tST:/sg;s/IT¥s{3}/¥nIT:/sg;s/(L¥d{1,2})¥s{3}.*//sg;
push(@text,$_);
}
foreach $_ (@text){
print OUT "$_";   }
close(OUT);@arr= ();@text= ();
open(FH,"<20chu_near_2.txt");@arr=<FH>;close(FH);
open(OUT, ">20chu_near_3.txt");
foreach (@arr){if(/RL:/){ s/IT:(.*)¥s{5}(RL:.*)¥s{7,8}(¥(.*)$/$1¥t$3¥t$3¥t$2/g;}
else{s/IT:(.*)¥s{8}(¥(.*)$/$1¥t$2¥t$2/;}
s/¥s{8}/ /g;#s/¥s{5}/¥s/g;#s/¥s¥s¥s¥s¥s/¥s/g;#s/¥s¥s¥s¥s/¥s/g;
push(@text,$_);}
foreach $_ (@text){
print OUT "$_";
}
close(OUT);@arr= ();@text= ();
open(FH,"<20chu_near_3.txt");
@arr=<FH>;close(FH);
open(OUT, ">20chu_near_3_2.txt");
$x=0;$no=1;
foreach(@arr){
    if($_ eq "¥n"){   ;
    }elsif($_ =~ /^CC:(.*)¥n/sg){
      $title = $1;
      $x++;
    }else{
       print OUT "$x¥t$no¥t$title¥t$_";
       $no++;
       }}
close(OUT);@arr= ();@text= ();

#②テキスト説明句からの薬理メカニズム用語抽出
open(FH,"<20chu_near_3_2.txt");
@arr=<FH>;close(FH);
open(OUT, ">20chu_near_4.txt");
my $modi_1;
foreach (@arr){
chomp ($_);
@data = split(/¥t/, $_);
$modi1 = $data[6];
$modi1 =~ s/(antagonists?)¥sand¥s(agonists?)/$1-$2/sg;
$modi1 =~
s/(inhibitors?)¥sof¥s(.*)(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|¥sand¥s|¥sto¥s|¥sby¥s|¥swith¥s|,)
/$2 $1 /sg;
$modi1 =~
s/¥((inhibitors?|agonist?|antagonist?);¥s(.*?)(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|¥sand¥s|¥sto¥s
```

```
|¥sby¥s|¥swith¥s|,)/($2 $1 /sg;
if($modi1 =~ /(inhibitor|agonist)/g){
$modi1 =~
s/.*(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|^¥(|,¥s|:¥s|¥sand¥s|¥sto¥s|¥sby¥s|¥swith¥s)(.{1,80}(
agonist|inhibitor|stimulant)).*/$2/ig;
$modi1 =~ s/¥s{8}/i /g;
}else{$modi1 = "-";}
$total =
join("¥t",$data[0],$data[1],$data[2],$data[3],$data[4],$data[5],$modi1,$data[7],$data[8],"¥n");
push(@all,$total);
}for $line (@all){
print OUT "$line";}
close(OUT);
@arr= ();@all= ();
open(FH,"<20chu_near_4.txt");
@arr=<FH>;close(FH);
open(OUT, ">20chu_near_5.txt");
foreach(@arr){
s/<s¥d¥d?>//g;
s/<e¥d¥d?>/ /g;
s/¥(/ /g;
s/¥)/ /g;
push(@text,$_);
}
for $line (@text){
print OUT "$line";
}
close(OUT);@arr= ();@text= ();

#③既知用語辞書による照合（近接演算子機能を使用）
open(FH,"<20chu_near_5.txt");
@arr=<FH>;close(FH);
open(FH2,"<cancer_dic_700_7.txt");
@arr2=<FH2>;close(FH2);
open(OUT, ">20chu_near_10.txt");
foreach(@arr){
chomp ($_);
@data = split(/¥t/, $_);
$_ =~ s/[0-9]{2,9}-[0-9]{0,2}-[0-9].?//sg;
$line = $_;
$text1 = $data[7];
$near13 = '(?:¥W+¥w+){0,13}?¥W+' ;
$near12 = '(?:¥W+¥w+){0,12}?¥W+' ;
$near11 = '(?:¥W+¥w+){0,11}?¥W+' ;
$near10 = '(?:¥W+¥w+){0,10}?¥W+' ;
$near9 = '(?:¥W+¥w+){0,9}?¥W+' ;
$near8 = '(?:¥W+¥w+){0,8}?¥W+' ;
$near7 = '(?:¥W+¥w+){0,7}?¥W+' ;
$near6 = '(?:¥W+¥w+){0,6}?¥W+' ;
$near5 = '(?:¥W+¥w+){0,5}?¥W+' ;
$near4 = '(?:¥W+¥w+){0,4}?¥W+' ;
$near3 = '(?:¥W+¥w+){0,3}?¥W+' ;
$near2 = '(?:¥W+¥w+){0,2}?¥W+' ;
$near1 = '(?:¥W+¥w+){0,1}?¥W+' ;
$near0 = '(?:¥W+¥w+){0}?¥W+' ;

foreach(@arr2){
chomp ($_);
@dat = split(/¥t/);
$number = @dat;
if($number == 5 ){
```

```
if($text1 =~ /¥b($dat[3]s?$near1$dat[4]s?)¥b/i){print OUT
$dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}
if($text1 =~ /¥b($dat[4]s?$near1$dat[3]s?)¥b/i){print OUT
$dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}
}
elsif($number == 6 ){
$key = "($dat[3]|$dat[4]|$dat[5])";
if($text1 =~ /¥b(${key}s?$near2${key}s?$near2${key}s?)¥b/i)
        {$chu1 = $1;my @tmp_arr;my %h;
         @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                  for my $t (@tmp_arr) {
                       $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                       if(!$h{$t}) {$h{$t}++;}
                           else {goto LABEL1;}}
              {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$chu1."¥t".$line."¥n";} }
                  LABEL1:;}
elsif($number == 7 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6])";
if($text1 =~ /¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?)¥b/i)

    {$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                  for my $t (@tmp_arr) {
                  $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                       if(!$h{$t}) {$h{$t}++;}
                           else {goto LABEL2;}}
              {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$chu1."¥t".$line."¥n";} }
                  LABEL2:;}
elsif($number == 8 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7])";
if($text1 =~ /¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?)¥b/i)
{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                  for my $t (@tmp_arr) {
                  $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                       if(!$h{$t}) {$h{$t}++;}
                           else {goto LABEL3;}}
              {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$chu1."¥t".$line."¥n";} }
                  LABEL3:;}
elsif($number == 9 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8])";
if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?)¥b/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                  for my $t (@tmp_arr) {
                  $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                       if(!$h{$t}) {$h{$t}++;}
                           else {goto LABEL4;}}
              {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$chu1."¥t".$line."¥n";} }
                  LABEL4:;}
elsif($number == 10 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9])";
if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?)¥b/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                  for my $t (@tmp_arr) {
                  $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
```

```
                        if(!$h{$t}) {$h{$t}++;}
                            else {goto LABEL5;}}
            {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}}
            LABEL5:;}
elsif($number == 11 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9]|$dat[10])";
if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?$near2${key}s?)¥b/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                for my $t (@tmp_arr) {
                $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                    if(!$h{$t}) {$h{$t}++;}
                        else {goto LABEL6;}}
            {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}}
            LABEL6:;}
elsif($number == 12 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9]|$dat[10]|$dat[11])";

if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?$near2${key}s?$near2${key}s?)¥b/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                for my $t (@tmp_arr) {
                $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                    if(!$h{$t}) {$h{$t}++;}
                        else {goto LABEL7;}}

                  {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}}
                   LABEL7:;}

elsif($number == 13 ){
$key = "($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9]|$dat[10]|$dat[11]|$dat[12])";

if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?$near2${key}s?$near2${key}s?$near2${key}s?)/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                for my $t (@tmp_arr) {
                $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                    if(!$h{$t}) {$h{$t}++;}
                        else {goto LABEL8;}}

                  {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";}}
                   LABEL8:;}
elsif($number == 14 ){
$key =
"($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9]|$dat[10]|$dat[11]|$dat[12]|$dat[13
])";

if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?)¥b/i)

 {$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
```

```
                    for my $t (@tmp_arr) {
                    $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                          if(!$h{$t}) {$h{$t}++;}
                              else {goto LABEL9;}}
                    {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";} }
                     LABEL9:;}
elsif($number == 15 ){
$key =
"($dat[3]|$dat[4]|$dat[5]|$dat[6]|$dat[7]|$dat[8]|$dat[9]|$dat[10]|$dat[11]|$dat[12]$dat[13]
|$dat[14])";

if($text1 =~
/¥b(${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${
key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?$near2${key}s?)¥b/i)

{$chu1 = $1;my @tmp_arr;my %h;
        @tmp_arr = split(/[¥s+¥-.;]/,$chu1);
                    for my $t (@tmp_arr) {
                    $t = lc($t);$t =~ s/inhibitors/inhibitor/sg;
                          if(!$h{$t}) {$h{$t}++;}
                              else {goto LABEL10;}}
               {print OUT $dat[0]."¥t".$dat[1]."¥t".$dat[2]."¥t".$1."¥t".$line."¥n";} }
                LABEL10:;}}}
close(OUT);
@arr= ();
my %title_datas = ();
open(IN, '20chu_near_5.txt');
while(my $chu_new = <IN>){
       chomp($chu_new);
       my ($id_j,$id,$title,$keyword,$index,$target,$modifi,$role) = split(/¥t/,$chu_new,8);
       $title_datas{$id} = [$id_j,$id,$title,$keyword,$index,$target,$modifi,$role];
}
close(IN);
open(OUT, ">20chu_near_11.txt");
open(IN, "20chu_near_10.txt");

#④ヒットしたデータレコードの除去
while(my $line =<IN>){
       chomp($line);
       my
($symbole,$yakuzai,$yakuzai2,$match,$id_j,$id,$title,$keyword,$index,$target,$modifi,$role)=sp
lit(/¥t/,$line,10);
        if($title_datas{$id}) {
        print OUT
"$symbole¥t$yakuzai¥t$yakuzai2¥t$match¥t$id_j¥t$id¥t$title¥t$keyword¥t$index¥t$target¥t
$modifi¥t$role¥n";
        delete $title_datas{$id};
        }else{
          print OUT
"$symbole¥t$yakuzai¥t$yakuzai2¥t$match¥t$id_j¥t$id¥t$title¥t$keyword¥t$index¥t$target¥t
$modifi¥t$role¥n";
        }
}
close(IN);
foreach my $key (keys %title_datas){
$data = $title_datas{$key};
print OUT
"¥t¥t¥t¥t@$data[0]¥t@$data[1]¥t@$data[2]¥t@$data[3]¥t@$data[4]¥t@$data[5]¥t@$data[6]¥t
@$data[7]¥t@$data[8]¥t@$data[9]¥t@$data[10]¥n";
}close(OUT);
```

```
$data_file = "lung_9.txt";
open(FH,"< $data_file") or die qq{can't open};
@arr=<FH>;close(FH);open(OUT, ">20chu_near_2.txt");
foreach (@arr){
s/¥n//sg;s/CC¥s{3}/¥nCC:/sg;#s/TI¥s{3}/¥tTI:/sg;s/TI¥s{3}/¥t/sg;
s/ST¥s{3}/¥tST:/sg;s/IT¥s{3}/¥nIT:/sg;s/(L¥d{1,2})¥s{3}.*//sg;
push(@text,$_);
}
foreach $_ (@text){
print OUT "$_";   }
close(OUT);@arr= ();@text= ();
open(FH,"<20chu_near_2.txt");@arr=<FH>;close(FH);
open(OUT, ">20chu_near_3.txt");
foreach (@arr){if(/RL:/){ s/IT:(.*)¥s{5}(RL:.*)¥s{7,8}(¥(.*)$/$1¥t$3¥t$3¥t$2/g;}
else{s/IT:(.*)¥s{8}(¥(.*)$/$1¥t$2¥t$2/;}
s/¥s{8}/ /g;#s/¥s{5}/¥s/g;#s/¥s¥s¥s¥s¥s/¥s/g;#s/¥s¥s¥s¥s/¥s/g;
push(@text,$_);}
foreach $_ (@text){
print OUT "$_";
}
close(OUT);@arr= ();@text= ();
open(FH,"<20chu_near_3.txt");
@arr=<FH>;close(FH);
open(OUT, ">20chu_near_3_2.txt");
$x=0;$no=1;
foreach(@arr){
    if($_ eq "¥n"){   ;
    }elsif($_ =~ /^CC:(.*)¥n/sg){
      $title = $1;
      $x++;
    }else{
       print OUT "$x¥t$no¥t$title¥t$_";
       $no++;
       }}
close(OUT);@arr= ();@text= ();
open(FH,"<20chu_near_3_2.txt");
@arr=<FH>;close(FH);
open(OUT, ">20chu_near_4.txt");
my $modi_1;
foreach (@arr){
chomp ($_);
@data = split(/¥t/, $_);
$modi1 = $data[6];
$modi1 =~ s/(antagonists?)¥sand¥s(agonists?)/$1-$2/sg;
$modi1 =~
s/(inhibitors?)¥sof¥s(.*)(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|¥sand¥s|¥sto¥s|¥sby¥s|¥swith¥s|,)/$
2 $1 /sg;
$modi1 =~
s/¥((inhibitors?|agonist?|antagonist?);¥s(.*?)(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|¥sand¥s|¥sto¥s|
¥sby¥s|¥swith¥s|,)/($2 $1 /sg;
if($modi1 =~ /(inhibitor|agonist)/g){
$modi1 =~
s/.*(¥sas¥s|¥sof¥s|¥son¥s|¥sin¥s|^¥(|,¥s|:¥s|¥sand¥s|¥sto¥s|¥sby¥s|¥swith¥s)(.{1,80}(ag
onist|inhibitor|stimulant)).*/$2/ig;
$modi1 =~ s/¥s{8}/i /g;
}else{$modi1 = "-";}
$total =
join("¥t",$data[0],$data[1],$data[2],$data[3],$data[4],$data[5],$modi1,$data[7],$data[8],"¥n");
push(@all,$total);
```

```
}for $line (@all){
print OUT "$line";}
close(OUT);
```

付録-2

使用したストップワードのリスト

a a's able about above according accordingly across actually after afterwards again against ain't all allow allows almost alone along already also although always am among amongst an and another any anybody anyhow anyone anything anyway anyways anywhere apart appear appreciate appropriate are aren't around as aside ask asking associated at available away awfully b be became because become becomes becoming been before beforehand behind being believe below beside besides best better between beyond both brief but by c c'mon c's came can can't cannot cant cause causes certain certainly changes clearly co com come comes concerning consequently consider considering contain containing contains corresponding could couldn't course currently d definitely described despite did didn't different do does doesn't doing don't done down downwards during e each edu eg eight either else elsewhere enough entirely especially et etc even ever every everybody everyone everything everywhere ex exactly example except f far few fifth first five followed following follows for former formerly forth four from further furthermore g get gets getting given gives go goes going gone got gotten greetings h had hadn't happens hardly has hasn't have haven't having he he's hello help hence her here here's hereafter hereby herein hereupon hers herself hi him himself his hither hopefully how howbeit however i i'd i'll i'm i've ie if ignored immediate in inasmuch inc indeed indicate indicated indicates inner insofar instead into inward is isn't it it'd it'll it's its itself j just k keep keeps kept know knows known l last lately later latter latterly least less lest let let's like liked likely little look looking looks ltd m mainly many may maybe me mean meanwhile merely might more moreover most mostly much must my myself n name namely nd near nearly necessary need needs neither never nevertheless new next nine no nobody non none noone nor normally not nothing novel now nowhere o obviously of off often oh ok okay old on once one ones only onto or other others otherwise ought our ours ourselves out outside over overall own p particular particularly per perhaps placed please plus possible presumably probably provides q que quite qv r rather rd re really reasonably regarding regardless regards relatively respectively right s said same saw say saying says second secondly see seeing seem seemed seeming seems seen self selves sensible sent serious seriously seven several shall she should shouldn't since six so some somebody somehow someone something sometime sometimes somewhat somewhere soon sorry specified specify specifying still sub such sup sure t t's take taken tell tends th than thank thanks thanx that that's thats the their theirs them themselves then thence there there's thereafter thereby therefore therein theres thereupon these they they'd they'll they're they've think third this thorough thoroughly those though three through throughout thru thus to together too took toward towards tried tries truly try trying twice two u un under unfortunately unless unlikely until unto up upon us use used useful uses using usually uucp v value various very via viz vs w want wants was wasn't way we we'd we'll we're we've welcome well went were weren't what what's whatever when whence whenever where where's whereafter whereas whereby wherein whereupon wherever whether which while whither who who's whoever whole whom whose why will willing wish with within without won't wonder would would wouldn't x y yes yet you you'd you'll you're you've your yours yourself yourselves z zero , ; .

付録−3

Web-API による PubMED の自動検索と STF-IDF の計算

```
use warnings;
use utf8;
use LWP::Simple;
my $total = 1000000;
BEGIN {
$ENV{http_proxy}='http://gate.local:8080/';
}
open(FH,"< api_1_list_4-5_0ken.csv") or die("error :$!");
@arr=<FH>;
close(FH);
open(OUT, "> api_2_kekka_11.txt");
my @api;
foreach(@arr){
chomp;
my @data = split(/,/);
my $key    = "¥"$data[1]¥t¥"[All Fields]";
my $key2 = "¥"$data[1]s¥t¥"[All Fields]";
my $data = get("http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?
                db=pubmed&term=($key+or+$key2)&rettype=Count");
my($num)= ($data=~/<Count>(¥d+)<¥/Count>/);
if($num != 0){
$tfidf1 = ($data[0]+30)*(log10($total/($num))+1);
$tfidf2 = sprintf("%.2f", $tfidf1);
}
$tfidf3 = $tfidf2.",".$data[1].",".$num.",".$data[0];
push(@api,$tfidf3);
$tfidf1= 0;
$tfidf2= 0;
$tfidf3= 0;
$num = 0;
sleep 1;
}
foreach(@api){
print OUT "$_¥n";
}
close(OUT);
sub log10 {
   my $x = shift;
   return log($x) / log(10);
}
```

付録-4

STF-IDF の計算における a 値による補正の効果

色を表示する範囲	
stf 値	
10-15	
5-9	
df 値	
101-500	
50-100	
20-49	

no	stf a=0	df	stf a=10	df	stf a=20	df	stf a=30	df	stf a=40	df	stf a=50	df	stf a=60	df
1	15	31	15	31	15	31	8	2	8	2	8	2	8	2
2	13	58	13	58	8	2	15	31	4	1	4	1	4	1
3	12	70	12	70	13	58	4	1	15	31	2	1	2	1
4	8	2	8	2	4	1	13	58	2	1	2	1	2	1
5	10	174	10	174	12	70	2	1	2	1	15	31	1	1
6	8	17	8	17	8	17	2	1	1	1	1	1	1	1
7	7	37	7	37	2	1	8	17	1	1	1	1	1	1
8	7	81	6	14	2	1	1	1	1	1	1	1	1	1
9	6	14	7	81	6	14	1	1	1	1	1	1	1	1
10	5	15	4	1	2	2	1	1	1	1	1	1	1	1
11	6	144	5	15	2	2	1	1	1	1	1	1	1	1
12	4	1	6	144	1	1	1	1	1	1	1	1	1	1
13	5	28	5	28	1	1	1	1	1	1	1	1	2	2
14	5	61	5	61	1	1	1	1	2	2	2	2	2	2
15	5	67	5	67	1	1	1	1	2	2	2	2	15	31
16	4	8	4	8	1	1	12	70	13	58	1	2	1	2
17	5	196	5	196	1	1	2	2	8	17	2	3	2	3
18	4	28	4	28	1	1	2	2	1	2	2	3	2	3
19	4	58	4	58	1	1	6	14	2	3	8	17	1	3
20	3	9	3	9	7	37	2	3	2	3	1	3	1	3
21	3	12	3	12	4	8	2	3	6	14	1	3	1	3
22	3	13	3	13	5	15	1	2	4	8	1	3	8	17
23	3	20	3	20	2	3	4	8	12	70	13	58	2	5
24	3	26	3	26	2	3	5	15	1	3	4	8	1	4
25	3	136	2	1	10	174	1	3	1	3	6	14	1	4
26	2	1	2	1	1	2	1	3	1	3	2	5	1	4
27	2	1	2	2	3	9	1	3	2	5	1	4	1	4
28	2	2	2	2	5	28	2	5	1	4	1	4	1	4
29	2	2	2	3	2	5	7	37	1	4	1	4	4	8
30	2	3	2	3	7	81	3	9	1	4	1	4	6	14
31	2	3	3	136	1	3	1	4	1	4	1	4	1	5
32	2	5	2	5	1	3	1	4	1	4	1	5	1	5
33	2	7	2	7	1	3	1	4	5	15	1	5	1	5
34	2	8	2	8	3	12	1	4	3	9	1	5	1	5
35	2	9	2	9	2	7	1	4	2	7	1	5	13	58
36	2	13	2	13	3	13	2	7	1	5	3	9	2	7
37	2	16	2	16	1	4	3	12	1	5	5	15	3	9
38	2	23	2	23	1	4	1	5	1	5	2	7	1	6
39	2	25	2	25	1	4	1	5	1	5	12	70	1	6
40	2	26	2	26	1	4	1	5	2	8	1	6	5	15
41	2	34	2	34	1	4	1	5	7	37	1	6	2	8
42	2	44	2	44	2	8	2	8	1	6	2	8	1	7
43	2	58	2	58	4	28	5	28	1	6	2	9	2	9
44	2	61	2	61	2	9	3	13	3	12	1	7	3	12
45	2	63	2	63	1	5	2	9	2	9	3	12	1	8
46	2	72	2	72	1	5	1	6	3	13	3	13	1	8

47	2	72	2	72	1	5	1	6	1	7	1	8	1	8
48	2	87	2	87	1	5	1	7	1	8	1	8	1	8
49	2	95	2	95	3	20	10	174	1	8	1	8	12	70
50	2	106	2	106	1	6	1	8	1	8	1	8	3	13
51	2	113	2	113	1	6	1	8	1	8	7	37	1	9
52	2	116	2	116	5	61	1	8	5	28	1	9	1	9
53	2	127	2	127	2	13	1	8	1	9	1	9	1	10
54	2	130	2	130	5	67	4	28	1	9	2	13	2	13
55	2	180	2	180	1	7	7	81	2	13	1	10	7	37
56	2	199	2	199	3	26	2	13	1	10	5	28	1	11
57	1	1	1	1	1	8	3	20	3	20	1	11	1	11
58	1	1	1	1	1	8	1	9	4	28	1	11	1	11
59	1	1	1	1	1	8	1	9	1	11	1	11	1	12
60	1	1	1	1	1	8	1	10	1	11	3	20	1	12
61	1	1	1	1	2	16	2	16	1	11	1	12	1	12
62	1	1	1	1	1	9	1	11	2	16	1	12	1	12
63	1	1	1	1	1	9	1	11	1	12	1	12	5	28
64	1	1	1	1	1	10	1	11	1	12	1	12	2	16
65	1	2	1	2	6	144	3	26	1	12	2	16	1	13
66	1	3	1	3	4	58	1	12	1	12	1	13	1	13
67	1	3	1	3	1	11	1	12	1	13	1	13	3	20
68	1	3	1	3	1	11	1	12	1	13	4	28	1	14
69	1	4	1	4	1	11	1	12	3	26	1	14	1	14
70	1	4	1	4	1	12	5	61	1	14	1	14	4	28
71	1	4	1	4	1	12	1	13	1	14	1	15	1	15
72	1	4	1	4	1	12	1	13	7	81	1	15	1	15
73	1	4	1	4	1	12	1	14	1	15	3	26	1	17
74	1	5	1	5	2	23	1	14	1	15	1	17	3	26
75	1	5	1	5	1	13	5	67	10	174	2	23	2	23
76	1	5	1	5	1	13	1	15	2	23	2	25	1	20
77	1	5	1	5	2	25	1	15	1	17	1	20	2	25
78	1	6	1	6	1	14	2	23	2	25	2	26	1	21
79	1	6	1	6	1	14	2	25	5	61	7	81	1	21
80	1	7	1	7	2	26	1	17	2	26	1	21	2	26
81	1	8	1	8	1	15	2	26	1	20	1	21	1	23
82	1	8	1	8	1	15	4	58	5	67	1	23	1	23
83	1	8	1	8	1	17	1	20	1	21	1	23	1	23
84	1	8	1	8	2	34	1	21	1	21	1	23	1	23
85	1	9	1	9	1	20	1	21	1	23	1	23	1	24
86	1	9	1	9	1	21	2	34	1	23	5	61	7	81
87	1	10	1	10	1	21	1	23	1	23	1	24	1	27
88	1	11	1	11	1	23	1	23	1	23	10	174	2	34
89	1	11	1	11	1	23	1	23	1	24	5	67	5	61
90	1	11	1	11	1	23	1	23	4	58	2	34	1	28
91	1	12	1	12	1	23	6	144	2	34	1	27	1	28
92	1	12	1	12	1	24	1	24	1	27	1	28	5	67
93	1	12	1	12	2	44	1	27	1	28	1	28	1	31
94	1	12	1	12	5	196	1	28	1	28	4	58	4	58
95	1	13	1	13	1	27	1	28	1	31	1	31	1	33
96	1	13	1	13	1	28	2	44	2	44	1	33	1	33
97	1	14	1	14	1	28	1	31	1	33	1	33	1	34
98	1	14	1	14	1	31	1	33	1	33	1	34	10	174
99	1	15	1	15	2	58	1	33	1	34	2	44	2	44
100	1	15	1	15	1	33	1	34	1	36	1	36	1	36
101	1	17	1	17	1	33	1	36	6	144	1	39	1	39
102	1	20	1	20	1	34	1	39	1	39	1	41	1	41
103	1	21	1	21	2	61	2	58	1	41	1	41	1	41
104	1	21	1	21	2	63	1	41	1	41	1	42	1	42
105	1	23	1	23	1	36	1	41	1	42	1	43	1	43
106	1	23	1	23	1	39	2	61	1	43	1	43	1	43
107	1	23	1	23	2	72	1	42	1	43	1	43	1	43
108	1	23	1	23	2	72	2	63	1	43	1	44	1	44
109	1	24	1	24	1	41	1	43	2	58	1	44	1	44
110	1	27	1	27	1	41	1	43	1	44	2	58	1	47
111	1	28	1	28	1	42	1	43	1	44	1	47	2	58
112	1	28	1	28	1	43	1	44	2	61	2	61	1	49

113	1	31	1	31	1	43	1	44	1	47	6	144	1	50
114	1	33	1	33	1	43	1	47	2	63	1	49	2	61
115	1	33	1	33	1	44	5	196	1	49	2	63	2	63
116	1	34	1	34	1	44	1	49	1	50	1	50	1	52
117	1	36	1	36	3	136	2	72	1	52	1	52	1	53
118	1	39	1	39	1	47	2	72	1	53	1	53	1	55
119	1	41	1	41	1	49	1	50	2	72	1	55	1	55
120	1	41	1	41	2	87	1	52	2	72	1	55	1	55
121	1	42	1	42	1	50	1	53	1	55	1	55	1	56
122	1	43	1	43	1	52	1	55	1	55	1	56	1	57
123	1	43	1	43	1	53	1	55	1	55	1	57	6	144
124	1	43	1	43	2	95	1	55	1	56	2	72	1	59
125	1	44	1	44	1	55	1	56	1	57	2	72	1	59
126	1	44	1	44	1	55	1	57	1	59	1	59	2	72
127	1	47	1	47	1	55	1	59	1	59	1	59	2	72
128	1	49	1	49	1	56	1	59	1	61	1	61	1	61
129	1	50	1	50	1	57	2	87	1	64	1	64	1	64
130	1	52	1	52	1	59	1	61	2	87	1	67	1	67
131	1	53	1	53	1	59	1	64	1	67	1	68	1	68
132	1	55	1	55	1	61	2	95	1	68	2	87	1	71
133	1	55	1	55	2	106	3	136	5	196	1	71	1	71
134	1	55	1	55	1	64	1	67	1	71	1	71	2	87
135	1	56	1	56	2	113	1	68	1	71	1	74	1	74
136	1	57	1	57	1	67	1	71	2	95	2	95	1	77
137	1	59	1	59	2	116	1	71	1	74	1	77	2	95
138	1	59	1	59	1	68	2	106	1	77	5	196	1	85
139	1	61	1	61	1	71	1	74	3	136	2	106	2	106
140	1	64	1	64	1	71	1	77	2	106	1	85	1	89
141	1	67	1	67	1	74	2	113	1	85	3	136	1	90
142	1	68	1	68	2	127	2	116	2	113	1	89	1	91
143	1	71	1	71	2	130	1	85	2	116	1	90	1	93
144	1	71	1	71	1	77	2	127	1	89	2	113	2	113
145	1	74	1	74	1	85	1	89	1	90	1	91	3	136
146	1	77	1	77	1	89	1	90	1	91	2	116	2	116
147	1	85	1	85	1	90	2	130	1	93	1	93	5	196
148	1	89	1	89	1	91	1	91	2	127	1	98	1	98
149	1	90	1	90	1	93	1	93	1	98	1	100	1	100
150	1	91	1	91	1	98	1	98	2	130	1	101	1	101
151	1	93	1	93	1	100	1	100	1	100	2	127	1	102
152	1	98	1	98	1	101	1	101	1	101	1	102	2	127
153	1	100	1	100	1	102	1	102	1	102	2	130	1	106
154	1	101	1	101	1	106	1	106	1	106	1	106	2	130
155	1	102	1	102	2	180	1	112	1	112	1	112	1	112
156	1	106	1	106	1	112	1	113	1	113	1	113	1	113
157	1	112	1	112	1	113	1	120	1	120	1	120	1	120
158	1	113	1	113	2	199	1	121	1	121	1	121	1	121
159	1	120	1	120	1	120	1	123	1	123	1	123	1	123
160	1	121	1	121	1	121	1	124	1	124	1	124	1	124
161	1	123	1	123	1	123	2	180	1	127	1	127	1	127
162	1	124	1	124	1	124	1	127	1	132	1	132	1	132
163	1	127	1	127	1	127	1	132	1	134	1	134	1	134
164	1	132	1	132	1	132	1	134	2	180	1	139	1	139
165	1	134	1	134	1	134	1	139	1	139	1	141	1	141
166	1	139	1	139	1	139	2	199	1	141	1	142	1	142
167	1	141	1	141	1	141	1	141	1	142	1	144	1	144
168	1	142	1	142	1	142	1	142	1	144	2	180	2	180
169	1	144	1	144	1	144	1	144	2	199	1	157	1	157
170	1	157	1	157	1	157	1	157	1	157	1	159	1	159
171	1	159	1	159	1	159	1	159	1	159	1	159	1	159
172	1	159	1	159	1	159	1	159	1	159	2	199	2	199
173	1	179	1	179	1	179	1	179	1	179	1	179	1	179
174	1	180	1	180	1	180	1	180	1	180	1	180	1	180
175	1	192	1	192	1	192	1	192	1	192	1	192	1	192
176	1	194	1	194	1	194	1	194	1	194	1	194	1	194

付録-5

研究シーズ用語のリスト

No.	研究シーズ用語	lung cancer との再検索	
		CA/全収録年代の文献レコード数 (2016/5/31)	CA/2013 年以前発行の文献レコード数
1	InsP3Kinase	1	0
2	KLF17	2	0
3	AZD9150	1	0
4	CREB CBP	1	0
5	CSC 3436	1	0
6	NBM T BBX OS01	1	0
7	BIP 4	1	0
8	A 893	1	0
9	AKT1 PDPK1	1	0
10	ETNPD5	1	0
11	MIF rpS3	1	0
12	MIP 1232	1	0
13	SKLB 677	1	0
14	Kir2 1 KCNJ2	2	1
15	Mps1 kinase inhibitor	3	0
16	topoisomerase 1 inhibitor	7	3
17	Stat1 HDAC4	1	0
18	FGFR3 BAIAP2L1	2	0
19	transforming growth factor beta Inhibitor	3	2
20	Thymidylate Synthase RNA	1	0
21	RegIIA	2	0
22	H1650GR	1	0
23	microRNA 19	2	1
24	Mdig	5	2
25	Betti reaction	1	0
26	LFC131	1	0
27	SMYD3 inhibitor	1	0
28	miRNA 506	1	0
29	phosphoinositide 3 kinase alpha	8	4
30	HIV smoker	1	0
31	CLK kinase	1	0
32	alpha 3 beta 2 nAChR	1	0
33	DDR1 inhibitor	2	0
34	CC 223	2	0
35	EGFRT790M	2	1
36	covalent reversible inhibitor	2	0
37	ST2825	1	0
38	EGFR C797S	1	0
39	LY2090314	1	0
40	BAY 87 2243	2	1
41	gold N heterocyclic carbene	4	0
42	SET oncogene	2	0
43	MEK162	1	0
44	LW6	1	0
45	protein kinase D inhibitor	1	0
46	BCL2 alpha	1	0
47	EBUS NA	1	0
48	desmosdumotin B	3	2
49	DDX3	8	1
50	2 yr OS	1	0
51	MET dependency	1	0
52	SSR128129E	1	0
53	AMG 232	2	0
54	FGFR1b	2	0
55	ARRY 142886	4	2
56	MHMD	1	0
57	factor 1 inhibitor	28	22

58	CQN	1	0
59	PF 04449913	1	0
60	ZD 6474	1	0
61	SHH antagonist	1	0
62	common DNA methylation	1	0
63	18F FAZA PET	2	0
64	G1202R	2	0
65	checkpoint kinase inhibitor	21	10
66	FAP inhibitor	1	0
67	included c Myc	1	0
68	MET TKIs	1	0
69	NCI H 460	1	0
70	angiokinase inhibitor	8	6
71	miR 27a inhibitor	1	0
72	TAK 733	1	0
73	LUX Lung 6	2	0
74	NF BETA	11	4
75	U2OS cell line	3	2
76	1 O acetylbritannilactone	2	0
77	PNAS 4	4	3
78	ABLIM	2	0
79	DYRK1A kinase	3	1
80	type 1 5 alpha reductase	1	0
81	microRNA 192	15	6
82	CX 4945	4	2
83	Wnt/ beta catenin signaling inhibitor	2	0
84	TopoII inhibitor	1	0
85	PF 03084014	2	0
86	TrkB inhibitor	1	0
87	beta4 nicotinic acetylcholine	1	0
88	JMJD5	2	1
89	LUX Lung 3	2	0
90	casein kinase 1 alpha	4	2
91	YM155	16	14
92	2 oxoglutarate oxygenase	1	0
93	CXCR2 inhibitor	1	0
94	CCR9 CCL25	3	2
95	MAPK11	4	2
96	macrophage inhibitor	2	0
97	HDAC1 inhibition	2	1
98	LDH inhibitor	2	0
99	BRD4 inhibitor	6	1
100	Wnt 7B	4	1
101	acetyl CoA carboxylase inhibitor	3	2
102	RHAMM receptor	3	2
103	HDAC6 inhibition	1	0
104	CARP 1	1	0
105	geranylgeranyltransferase I inhibitor	4	3
106	Symptomatic Radiation Pneumonitis	4	3
107	SJSA 1	1	0
108	smoothened inhibitor	2	1
109	FoxO6	1	0
110	KIF5B RET	25	13
111	proteins XIAP	132	97
112	aurora B inhibitor	8	6
113	TUSC2	9	6
114	OSI 906	14	8
115	cIAP 2	46	33
116	BIR2 domain	1	0
117	CNI 1493	1	0
118	RhoGDI2	16	12
119	adenosylhomocysteine hydrolase inhibitor	1	0
120	JMJD2	2	1
121	TIPE2	2	0
122	NR expression	1	0
123	AZD9291	17	0

124	anaplastic lymphoma kinase inhibitor	67	31
125	Elephantopus scaber	2	0
126	HOPX	6	4
127	TPO receptor agonist	2	1
128	SMYD2	3	1
129	JARID1B	4	1
130	Wnt 2	11	8
131	miR 340	7	0
132	RANKL inhibitor	2	0
133	cucurbitacin B	11	5
134	TMPRSS4	17	5
135	class III receptor	1	0
136	chronic immune thrombocytopenia ITP	1	0
137	HOXD3	10	8
138	IKB alpha	2	1
139	DKK4	6	2
140	Wnt 7A	16	9
141	NCI H446	74	60
142	HIF 1 inhibitor	9	4
143	early DNA damage	3	2
144	G1 growth arrest	2	1
145	BIM EL	4	3
146	multitargeted kinase inhibitor	17	13
147	AUY922	14	7
148	microRNA 101	22	13
149	peroxiredoxin II	7	6
150	glycogen synthase kinase 3 inhibitor	7	6
151	ARHGDIB	5	3
152	metadherin	11	5
153	sU11274	20	16
154	SMYD3	2	1
155	EZH2 inhibitor	15	1
156	GAS5	6	1
157	DOCK1	6	4
158	TSLC1	407	302
159	Notch 1 signaling	5	4
160	FAP 1	6	4
161	ABT 263	23	12
162	MSH2 expression	11	10
163	ANRIL	9	2
164	Hedgehog inhibitor	40	22
165	cyclin E2	28	15
166	PC9	27	17
167	HS 4	1	0
168	DKK2	3	2
169	CDKN2D	12	7
170	neplanocin A	2	1
171	HAI 1	2	1
172	ATP synthase inhibitor	3	2
173	TFPI 2	15	11
174	mitochondrial complex I inhibitor	1	0
175	MAPK14	10	4
176	Src protein tyrosine kinase	3	2
合計	-	1575	927

付録-6

既知用語辞書でヒットした索引

既知用語辞書の研究シーズ用語	既知用語辞書とマッチングしたパターン	集計
AKT inhibitor	AKT inhibitor	3
	AKT/mTOR inhibitors	3
	AKT-mTOR inhibitors	1
anaplastic lymphoma receptor tyrosine kinase inhibitor	ALK inhibitor	19
	ALK- inhibitor	1
	ALK inhibitors	39
angiogenesis inhibitor	angiogenesis inhibitor	1
	angiogenesis inhibitors	7
	inhibitors of angiogenesis	1
apoptosis inducer	apoptosis inducers	1
apoptosis regulator Bcl-2 inhibitor	BCL-2 family inhibitors	1
aurora kinase A inhibitor	as a kinase inhibitor	4
	aurora kinase A inhibitor	5
	inhibitors; anaplastic Lymphoma Kinase as a	1
breakpoint cluster region/tyrosine-protein kinase ABL1 inhibitor	ABL inhibitors	2
cell cycle inhibitor	cell cycle inhibitors	1
	cell lung cancer cells to mTOR inhibitors	1
	inhibitor, induces cell cycle	1
chemokine (C-X-C motif) receptor 4 antagonist	CXCR4 antagonist	3
cyclin-dependent kinase 2 inhibitor	inhibitors against CDK2	1
cyclin-dependent kinase 4 inhibitor	CDK4/6 inhibitors	14
	inhibitor of cdk4	1
cyclin-dependent kinase 7 inhibitor	CDK7 inhibitor	2
cyclin-dependent kinase inhibitor	inhibitors of CDK	1
dipeptidyl-peptidase 4 inhibitor	CD26/DPP4 inhibitor	2
DNA repair inhibitor	DNA double-strand repair inhibitor	1
DNA topoisomerase 1 inhibitor	topoisomerase I inhibitor	1
	topoisomerase-I inhibitors	1
E3 ubiquitin-protein ligase Mdm2/tumor protein p53 interaction inhibitor	inhibitor blocked MDM2-p53 interaction	1
epidermal growth factor receptor expression inhibitor	inhibitor; epidermal growth factor receptor tyrosine kinase inhibitors	1
epidermal growth factor receptor inhibitor	EGFR inhibitor	11
	EGFR inhibitors	27
	EGFR kinase inhibitor	1
	EGFR-inhibitor	4
	EGFR-T790M inhibitors	1
	EGFR-TK inhibitors	1
	epidermal growth factor receptor blocker tyrosine kinase inhibitor	4
	epidermal growth factor receptor inhibitors	1
	epidermal growth factor receptor tyrosine kinase inhibitor	8
	epidermal growth factor receptor tyrosine kinase inhibitors	38
	epidermal growth factor receptor-tyrosine kinase inhibitor	8
	epidermal growth factor receptor-tyrosine kinase inhibitors	25
	inhibitor EGFR	4
	inhibitor AZD9291 to epidermal growth factor receptor	1
	inhibitor insulin, epidermal growth factor, and androgen receptor	1
	inhibitor of EGFR	4
	inhibitor; epidermal growth factor receptor	1
	inhibitors in EGFR	4
	inhibitors of epidermal growth factor receptor	2
	inhibitors of the epidermal growth-factor receptor	2

	inhibitors, EGFR	15
	inhibitors; acquired EGFR	1
	inhibitors; EGFR	1
	inhibitors; epidermal growth factor receptor	1
	inhibitors; targeting EGFR	1
fibroblast growth factor receptor 1 inhibitor	inhibitors in FGFR1	8
fibroblast growth factor receptor inhibitor	FGFR inhibitor	1
	FGFR kinase inhibitors	6
	fibroblast growth factor receptor inhibitor	1
heat shock protein HSP 90 inhibitor	Hsp90 inhibitor	12
	HSP90 inhibitors	1
	inhibitor of HSP90	1
hepatocyte growth factor receptor inhibitor	c-Met inhibitor	1
	c-Met kinase inhibitors	2
	inhibitor of MET	5
	inhibitors of Met	1
	MET inhibitors	7
histone deacetylase inhibitor	HDAC inhibitor	1
	HDAC inhibitors	2
	HDAC-8 inhibitors	1
	histone deacetylase inhibitor	29
	histone deacetylase inhibitors	2
	histone deacetylase pan inhibitor	1
insulin like growth factor 1 receptor inhibitor	IGF-1R inhibitors	1
	IGF-1R/IR inhibitor	1
	insulin-like growth factor 1 receptor and its inhibitor	5
kinesin family member 11 inhibitor	Eg5 inhibitor	1
lysine-specific histone demethylase 1 inhibitor	LSD1 inhibitors	1
mitogen-activated protein kinase kinase 1 inhibitor	MEK1/2 inhibitor	1
	MEK1/2 inhibitors	1
myeloid cell leukemia 1 inhibitor	MCL-1 inhibitors	1
phosphatidylinositol 3-kinase inhibitor	inhibitors; PI3K	7
	PI3K inhibitor	3
	PI3K/mT0R inhibitor	3
	PI3K/mTOR inhibitor	1
polo-like kinase 1 inhibitor	Polo-like kinase 1 inhibitor	6
	polo-like kinase inhibitor	1
poly [ADP-ribose] polymerase 1 inhibitor	PARP inhibitors	1
programmed cell death 1 antagonist	antagonists of PD-1	1
prostaglandin-endoperoxide synthase 2 inhibitor	cox-2 inhibitor	2
	COX-2 inhibitors	8
	cyclooxygenase-2 inhibitors	2
proteasome inhibitor	proteasome inhibitor	9
	proteasome inhibitors	1
protein kinase C inhibitor	protein kinase C inhibitors	3
proto-oncogene tyrosine-protein kinase receptor Ret inhibitor	receptor tyrosine kinase inhibitor	4
	receptor tyrosine kinase inhibitors	8
	receptor tyrosine kinases inhibitors	1
	RET kinase inhibitors	5
proto-oncogene tyrosine-protein kinase ROS inhibitor	inhibitor in ROS1	3
	ROS1 kinase inhibitors	2
Raf proto-oncogene, serine/threonine kinase inhibitor	raf inhibitor	1
	Raf kinase inhibitor	3
RAS protein inhibitor	ras; selective inhibitor	2
serine/threonine-protein kinase B-raf inhibitor	BRAF inhibitor	3
serine/threonine-protein kinase mTOR inhibitor	inhibitor of mTOR	1
	mTOR dual inhibitor	1
	mTOR inhibitor	8
	mTOR inhibitors	5
substance P antagonist	substance P analog antagonist	4
telomerase inhibitor	telomerase/topoisomerase inhibitors	1
tyrosine kinase inhibitor	tyrosine kinase inhibitor	64
	tyrosine kinase inhibitors	80

	tyrosine-kinase inhibitors	1
vascular endothelial growth factor expression inhibitor	inhibitor PDTC on VEGF and endostatin expression	4
vascular endothelial growth factor inhibitor	vascular endothelial growth factor inhibitors	1
vascular endothelial growth factor receptor 2	VEGFR kinase inhibitor	1
inhibitor	VEGFR-2 inhibitors	1
vascular endothelial growth factor receptor inhibitor	vascular endothelial growth factor receptor tyrosine kinase inhibitor	2
総計		636

付録-7

抽出した研究シーズ用語の内容

No.	研究シーズ用語	内容（薬理メカニズム関連を網掛け）	stf (CA/2015 年発行の該当索引内の数)	df (PubMED/全収録年代の文献レコード数)	STF-IDF (stf+30)	lung cancer との再検索 CA/全収録年代の文献レコード数 (2016/5/31)
1	InsP3Kinase	薬理メカニズム	8	2	254.6	1
2	KLF17	薬理メカニズム	15	31	247.9	2
3	AZD9150	薬剤	4	1	238.0	1
4	CREB CBP	薬理メカニズム	13	58	225.2	1
5	CSC 3436	薬剤	2	1	224.0	1
6	NBM T BBX OS01	薬剤	2	1	224.0	1
7	BIP 4	薬剤	8	17	219.2	1
8	A 893	薬剤	1	1	217.0	1
9	AKT1 PDPK1	薬理メカニズム	1	1	217.0	1
10	ETNPD5	薬理メカニズム	1	1	217.0	1
11	MIF rpS3	薬理メカニズム	1	1	217.0	1
12	MIP 1232	薬剤	1	1	217.0	1
13	SKLB 677	薬剤	1	1	217.0	1
14	Kir2 1 KCNJ2	薬理メカニズム	1	1	217.0	2
15	Mps1 kinase inhibitor	薬理メカニズム	1	1	217.0	3
16	topoisomerase 1 inhibitor	薬理メカニズム	12	70	216.5	7
17	Stat1 HDAC4	薬理メカニズム	2	2	214.4	1
18	FGFR3 BAIAP2L1	薬理メカニズム	2	2	214.4	2
19	transforming growth factor beta Inhibitor	薬理メカニズム	6	14	210.7	3
20	Thymidylate Synthase RNA	バイオマーカー	2	3	208.7	1
21	RegIIA	薬理メカニズム	2	3	208.7	2
22	H1650GR	細胞	1	2	207.7	1
23	microRNA 19	薬理メカニズム	4	8	207.3	2
24	Mdig	薬理メカニズム	5	15	203.8	5
25	Betti reaction	反応	1	3	202.2	1
26	LFC131	リガンド	1	3	202.2	1
27	SMYD3 inhibitor	薬理メカニズム	1	3	202.2	1
28	miRNA 506	薬理メカニズム	2	5	201.6	1
29	phosphoinositide 3 kinase alpha	薬理メカニズム	7	37	201.0	8
30	HIV smoker	その他	3	9	199.5	1
31	CLK kinase	薬理メカニズム	1	4	198.3	1
32	alpha 3 beta 2 nAChR	薬理メカニズム	1	4	198.3	1
33	DDR1 inhibitor	薬理メカニズム	1	4	198.3	2
34	CC 223	薬剤	1	4	198.3	2
35	EGFRT790M	薬剤	1	4	198.3	2
36	covalent reversible inhibitor	薬理メカニズム	2	7	197.0	2
37	ST2825	薬剤	3	12	195.4	1
38	EGFR C797S	その他	1	5	195.3	1
39	LY2090314	薬剤	1	5	195.3	1
40	BAY 87 2243	薬剤	1	5	195.3	2
41	gold N heterocyclic carbene	薬剤	1	5	195.3	4
42	SET oncogene	薬理メカニズム	2	8	195.1	2
43	MEK162	薬剤	5	28	194.3	1
44	LW6	薬剤	3	13	194.2	1
45	protein kinase D inhibitor	薬理メカニズム	2	9	193.5	1
46	BCL2 alpha	薬理メカニズム	1	6	192.9	1
47	EBUS NA	その他	1	6	192.9	1
48	desmosdumotin B	薬剤	1	7	190.8	3
49	DDX3	薬理メカニズム	10	174	190.4	8

50	2 yr OS	その他	1	8	189.0	1
51	MET dependency	薬理メカニズム	1	8	189.0	1
52	SSR128129E	薬剤	1	8	189.0	1
53	AMG 232	薬剤	1	8	189.0	2
54	FGFR1b	薬理メカニズム	4	28	188.8	2
55	ARRY 142886	薬剤	7	81	188.4	4
56	MHMD	薬理メカニズム	2	13	188.4	1
57	factor 1 inhibitor	薬理メカニズム	3	20	188.1	28
58	CQN	薬剤	1	9	187.4	1
59	PF 04449913	薬剤	1	9	187.4	1
60	ZD 6474	バイオマーカー	1	10	186.0	1
61	SHH antagonist	薬理メカニズム	2	16	185.5	1
62	common DNA methylation	バイオマーカー	1	11	184.7	1
63	18F FAZA PET	その他	1	11	184.7	2
64	G1202R	薬剤	1	11	184.7	2
65	checkpoint kinase inhibitor	薬理メカニズム	3	26	184.3	21
66	FAP inhibitor	薬理メカニズム	1	12	183.5	1
67	included c Myc	薬理メカニズム	1	12	183.5	1
68	MET TKIs	薬理メカニズム	1	12	183.5	1
69	NCI H 460	細胞	1	12	183.5	1
70	angiokinase inhibitor	薬理メカニズム	5	61	182.5	8
71	miR 27a inhibitor	薬理メカニズム	1	13	182.5	1
72	TAK 733	薬剤	1	13	182.5	1
73	LUX Lung 6	その他	1	14	181.5	2
74	NF BETA	薬理メカニズム	1	14	181.5	11
75	U2OS cell line	細胞	5	67	181.1	3
76	1 O acetylbritannilactone	薬剤	1	15	180.5	2
77	PNAS 4	薬理メカニズム	1	15	180.5	4
78	ABLIM	薬理メカニズム	2	23	180.4	2
79	DYRK1A kinase	薬理メカニズム	2	25	179.3	3
80	type 1 5 alpha reductase	薬理メカニズム	1	17	178.9	1
81	microRNA 192	薬理メカニズム	2	26	178.7	15
82	CX 4945	薬剤	4	58	178.0	4
83	Wnt/ beta catenin signaling inhibitor	薬理メカニズム	1	20	176.7	2
84	TopoII inhibitor	薬理メカニズム	1	21	176.0	1
85	PF 03084014	薬剤	1	21	176.0	2
86	TrkB inhibitor	薬理メカニズム	2	34	175.0	1
87	beta4 nicotinic acetylcholine	薬理メカニズム	1	23	174.8	1
88	JMJD5	薬理メカニズム	1	23	174.8	2
89	LUX Lung 3	その他	1	23	174.8	2
90	casein kinase 1 alpha	薬理メカニズム	1	23	174.8	4
91	YM155	薬剤	6	144	174.3	16
92	2 oxoglutarate oxygenase	薬理メカニズム	1	24	174.2	1
93	CXCR2 inhibitor	薬理メカニズム	1	27	172.6	1
94	CCR9 CCL25	薬理メカニズム	1	28	172.1	3
95	MAPK11	薬理メカニズム	1	28	172.1	4
96	macrophage inhibitor	薬理メカニズム	2	44	171.4	2
97	HDAC1 inhibition	薬理メカニズム	1	31	170.8	2
98	LDH inhibitor	薬理メカニズム	1	33	169.9	2
99	BRD4 inhibitor	薬理メカニズム	1	33	169.9	6
100	Wnt 7B	薬理メカニズム	1	34	169.5	4
101	acetyl CoA carboxylase inhibitor	薬理メカニズム	1	36	168.8	3
102	RHAMM receptor	薬理メカニズム	1	39	167.7	3
103	HDAC6 inhibition	薬理メカニズム	2	58	167.6	1
104	CARP 1	薬剤	1	41	167.0	1
105	geranylgeranyltransferase I inhibitor	薬理メカニズム	1	41	167.0	4
106	Symptomatic Radiation Pneumonitis	薬理メカニズム	2	61	166.9	4
107	SJSA 1	細胞	1	42	166.7	1

108	smoothened inhibitor	薬理メカニズム	2	63	166.4	2
109	FoxO6	薬理メカニズム	1	43	166.4	1
110	KIF5B RET	薬理メカニズム	1	43	166.4	25
111	proteins XIAP	薬理メカニズム	1	43	166.4	132
112	aurora B inhibitor	薬理メカニズム	1	44	166.1	8
113	TUSC2	薬理メカニズム	1	44	166.1	9
114	OSI 906	薬剤	1	47	165.2	14
115	cIAP 2	薬剤	5	196	164.8	46
116	BIR2 domain	薬理メカニズム	1	49	164.6	1
117	CNI 1493	薬剤	2	72	164.6	1
118	RhoGDI2	薬理メカニズム	2	72	164.6	16
119	adenosylhomocysteine hydrolase inhibitor	薬理メカニズム	1	50	164.3	1
120	JMJD2	薬理メカニズム	1	52	163.8	2
121	TIPE2	薬理メカニズム	1	53	163.5	2
122	NR expression	薬理メカニズム	1	55	163.0	1
123	AZD9291	薬剤	1	55	163.0	17
124	anaplastic lymphoma kinase inhibitor	バイオマーカー	1	55	163.0	67
125	Elephantopus scaber	その他	1	56	162.8	2
126	HOPX	薬理メカニズム	1	57	162.6	6
127	TPO receptor agonist	薬理メカニズム	1	59	162.1	2
128	SMYD2	薬理メカニズム	1	59	162.1	3
129	JARID1B	薬理メカニズム	2	87	161.9	4
130	Wnt 2	薬理メカニズム	1	61	161.7	11
131	miR 340	薬理メカニズム	1	64	161.0	7
132	RANKL inhibitor	薬理メカニズム	2	95	160.7	2
133	cucurbitacin B	薬剤	3	136	160.6	11
134	TMPRSS4	薬理メカニズム	1	67	160.4	17
135	class III receptor	薬理メカニズム	1	68	160.2	1
136	chronic immune thrombocytopenia ITP	薬理メカニズム	1	71	159.6	1
137	HOXD3	薬理メカニズム	1	71	159.6	10
138	IKB alpha	薬理メカニズム	2	106	159.2	2
139	DKK4	薬理メカニズム	1	74	159.1	6
140	Wnt 7A	薬理メカニズム	1	77	158.5	16
141	NCI H446	細胞	2	113	158.3	74
142	HIF 1 inhibitor	薬理メカニズム	2	116	157.9	9
143	early DNA damage	薬理メカニズム	1	85	157.2	3
144	G1 growth arrest	薬理メカニズム	2	127	156.7	2
145	BIM EL	薬理メカニズム	1	89	156.6	4
146	multitargeted kinase inhibitor	薬理メカニズム	1	90	156.4	17
147	AUY922	薬剤	2	130	156.4	14
148	microRNA 101	薬理メカニズム	1	91	156.3	22
149	peroxiredoxin II	薬理メカニズム	1	93	156.0	7
150	glycogen synthase kinase 3 inhibitor	薬理メカニズム	1	98	155.3	7
151	ARHGDIB	薬理メカニズム	1	100	155.0	5
152	metadherin	薬理メカニズム	1	101	154.9	11
153	sU11274	薬剤	1	102	154.7	20
154	SMYD3	薬剤	1	106	154.2	2
155	EZH2 inhibitor	薬理メカニズム	1	112	153.5	15
156	GAS5	薬理メカニズム	1	113	153.4	6
157	DOCK1	薬理メカニズム	1	120	152.5	6
158	TSLC1	薬理メカニズム	1	121	152.4	407
159	Notch 1 signaling	薬理メカニズム	1	123	152.2	5
160	FAP 1	バイオマーカー	1	124	152.1	6
161	ABT 263	薬剤	2	180	151.8	23
162	MSH2 expression	薬理メカニズム	1	127	151.8	11
163	ANRIL	薬理メカニズム	1	132	151.3	9
164	Hedgehog inhibitor	薬理メカニズム	1	134	151.1	40
165	cyclin E2	薬理メカニズム	1	139	150.6	28

166	PC9	細胞	2	199	150.4	27
167	HS 4	薬剤	1	141	150.4	1
168	DKK2	薬理メカニズム	1	142	150.3	3
169	CDKN2D	薬理メカニズム	1	144	150.1	12
170	neplanocin A	薬剤	1	157	148.9	2
171	HAI 1	薬理メカニズム	1	159	148.8	2
172	ATP synthase inhibitor	薬理メカニズム	1	159	148.8	3
173	TFPI 2	薬理メカニズム	1	179	147.2	15
174	mitochondrial complex I inhibitor	薬理メカニズム	1	180	147.1	1
175	MAPK14	薬理メカニズム	1	192	146.2	10
176	Src protein tyrosine kinase	薬理メカニズム	1	194	146.1	3

※網掛けは、薬理メカニズム、標的分子関連語