氏　名	小河　邦雄
学位の種類	博　士（　情報学　）
学位記番号	博甲第　８８８５　号
学位授与年月日	平成　３１年　１月　３１日
学位授与の要件	学位規則第４条第１項該当
審査研究科	図書館情報メディア研究科
学位論文題目	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

主	査	筑波大学　教授　博士（図書館情報学）	緑川信之
副	査	筑波大学　教授　博士（教育学）	芳鐘冬樹
副	査	筑波大学　教授　博士（学術）	中山伸一
副	査	筑波大学　教授　文学修士	逸村裕
副	査	長崎大学　教授　博士（理学）	本多正幸

論　文　の　要　旨（2,000 字程度）

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している。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である。この研究目的を達成するために，次の 2 つの研究課題を設定している。

研究課題 1：新奇な研究シーズ用語の候補を抽出する方法の提案

研究課題 2：新奇な研究シーズ用語の抽出実験による，提案した方法の妥当性の確認

第 1 章では，研究背景，研究目的，研究課題を述べた後，先行研究を整理して本研究で用いる方法との関連を示している。

第 2 章では，研究課題 1 に取り組んで，新奇な研究シーズ用語の抽出方法として，文献データベースを検索して作成したデータレコードの集合である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削除するという情報フィルタリングの手法を用いることとし，4 段階に分けて検討している。

抽出段階（1）は，既知用語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方法の検討である。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を決め，そのために，薬理メカニズム用語を医薬品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としている。

抽出段階（2）は，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の検討である。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を絞り込む。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としている。また，この段階の処理を自動化することを検討し

ている。

抽出段階（3）は，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研究シーズ用語を選別する方法の検討である。絞り込まれたデータレコードのテキスト情報を用語に自動分割した後で，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手動で選別するための選別ルールを設定している。ただし，データレコードが多く，手動での選別が困難な場合は，作用語を含むもしくは大文字を含むなどの選別ルールを設けて自動化もできるようにしている。最後に，選別した用語のデータ・クリーニングを行っている。

抽出段階（4）は，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の検討である。選別された研究シーズ用語を用いて文献データベースを検索し，得られた文献レコード数を使用して順位付けをすることによって研究シーズ用語を削減している。ここでは自動化をおこなっている。

第 3 章では，研究課題 2 に取り組んで，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，文献数が少ない新奇な研究シーズ用語の候補を抽出する実験をおこなっている。実験は，準備段階と抽出段階からなる。

準備段階では，まず，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，情報源として文献データベースの Chemical Abstracts(CA)を選択している。そして，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，検索結果の文献レコードを IT 索引単位に分割して，タイトル，統制語，テキスト説明句等からなるデータレコードおよびその集合であるデータセットを準備している。

抽出段階では，breast cancer と lung cancer に関して，第 2 章で検討した，(1) 既知用語辞書の作成，(2) データセットと既知用語辞書との照合によるデータレコードの絞り込み，(3) 絞り込まれたデータレコードからの研究シーズ用語の選別，(4) 選別された研究シーズ用語が多数の場合の順位付けによる削減，という 4 段階に従って作業を進めている。breast cancer では選別された研究シーズ用語が 28 個と適度な数になったため（4）は行わず，lung cancer では選別された研究シーズ用語が 499 個と多数のため，(4) の順位付けを行って 176 個に削減している。

第 4 章では，第 3 章の抽出実験の結果に基づいて，第 2 章で提案した抽出方法の妥当性について考察している。まず，実験により得られた研究シーズ用語を疾病名と共に CA を検索し，得られた文献レコード数が少ないことを確認している。また，開発が進んだ薬理メカニズム用語（既知の研究シーズ用語）での検索結果と比較しても，本研究で得られた研究シーズ用語での検索文献レコード数が少ないことを確認している。さらに，検索文献の発行年代別にみると比較的最近の文献が多く，研究段階の初期の情報が得られることを明らかにしている。以上から，提案した方法により新奇な研究シーズ用語が抽出され，提案方法には妥当性があることを確認したとしている。

第 5 章では，本研究の結論を述べている。

審　査　の　要　旨（2,000 字以上）

【批評】

第 1 章では，本研究の目的と研究課題について述べ，先行研究のレビューを行っている。研究ニーズを解決するための技術である研究シーズを探索することは，企業や研究所において重要な課題と

なっている。特に，研究報告の少ない研究シーズ（新奇な研究シーズ）を見つけることは研究の先端を切り開く上でも不可欠である。新奇な研究シーズ用語の抽出方法の研究は，こうした社会的要請に適っている。

しかし，大量の情報の中から研究報告の少ない研究シーズを見つけることは困難である。著者は，創薬研究における研究シーズ探索の方法を，文献調査，研究シーズ保有者による公開情報の入手，研究シーズの公募，テクノロジー・スカウティングサービス（専門調査機関）の利用，の 4 つに分類し，それぞれの利点と欠点をあげている。この中から文献調査の方法を採用し，大量の文献情報を効率的に処理するためにテキストマイニングの手法を検討して，情報フィルタリングを本研究の方法として採用している。情報フィルタリングを用いた研究はすでに多く行われているが，第 2 章でみるように段階を踏んで体系的に行っていること，自動化による処理と手動による処理を組み合わせていることに本研究の特徴がある。特に，後者の手動処理の部分では著者の業務上の経験が活かされていて，自動処理だけでは得られない成果を導いている。また，今後，人工知能等によって手動処理部分の自動処理化が可能になるかもしれないが，そのためにも自動処理と手動処理の境界を示したことは意義がある。ただし，この手動処理の記述が十分ではなく，いくらか具体性に欠ける点が惜しまれる。

第 2 章では，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の提案を行っている。第 1 章のところでも述べたが，抽出方法を（1）既知用語辞書の作成，（2）データセットと既知用語辞書との照合によるデータレコードの絞り込み，（3）絞り込まれたデータレコードからの研究シーズ用語の選別，（4）選別された研究シーズ用語が多数の場合の順位付けによる削減，という段階を踏んで体系的に行っているところに，本研究の特徴がある。関連研究の多くは，研究シーズ用語を抽出すること自体を目的として行われているため，結果を得ることが優先されているのに対し，本研究は方法を検討することを目的としている。そのため，それぞれの段階において選択肢を設けることで方法の適用範囲を検討している。たとえば，段階（1）の既知用語辞書の作成においては，抽出したい研究シーズ用語が対象としている領域（本研究では特定疾病）の既知用語を収集する場合と，より広い領域（本研究では全疾病）の既知用語を収集する場合を提案している。

また，段階（3）で既知用語辞書を用いた情報フィルタリングによって候補を絞り込んだとしても大量の研究シーズ用語が残される場合があり，この中から新奇な研究シーズ用語を抽出するのは依然として困難である。その場合のために，本研究では研究シーズ用語を順位付けることによって削減する段階（4）を設けている。研究シーズ用語が出現する文献数を基にして順位を求め，文献数が少ない研究シーズ用語を残すという方法である。しかし，第 4 章でみるように，段階（3）または段階（4）で得られた研究シーズ用語は，その後，その研究シーズ用語が出現する文献数によって新奇性が確認されている。この確認方法は段階（4）での順位付けの方法と同じではないが，文献数を基にするという点では重複しているため，方法の整合性に疑問が残る。どちらか 1 つの方法で十分かどうかも含めて，検討する必要がある。

第 3 章では，研究課題 2 として，新奇な研究シーズ用語の抽出実験を行い，実際に新奇な研究シーズ用語を抽出している。第 2 章のところでみたように，方法の適用範囲を検討するために選択肢を設けているが，この実験では breast cancer と lung cancer を研究シーズの対象疾病として，それぞれに異なる方法を適用している。たとえば，breast cancer では全疾病を対象とした既知用語辞書を作成し，lung cancer では肺がん関連の既知用語辞書を作成している。その結果，breast cancer では乳がん関連の研究シーズ用語が抽出されたが，lung cancer では肺がん以外の疾病に関する研究シーズ用語も得られた。これは，lung cancer では肺がん関連の既知用語しか除去していないためであるが，それがあらたな新薬開発につながる可能性もある。このように，抽出方法の適用範囲を変えるこ

とによって，研究・開発目的に合わせた使い分けができることを実証的に明らかにしたことは意義がある。ただし，breast cancer と lung cancer という 2 つの対象だけで複数の選択肢を設けたため，どの選択肢による効果なのか必ずしも明確に判断できない場合がある。選択肢の数に見合うだけの数の対象を用意するべきであったが，作業量が膨大であることを考慮すればやむを得なかったとも言える。

第 4 章では，第 3 章の抽出実験の結果に基づいて，第 2 章で提案した抽出方法の妥当性について考察を行っている。先に述べたように，抽出された研究シーズ用語が新奇であるかどうかを，その研究シーズ用語を含む文献数で判断し，いずれも文献数が少ないことから新奇な研究シーズ用語が抽出されたと判断し，方法の妥当性が確認されたとしている。ここでも，単に抽出された研究シーズ用語を含む文献数だけでなく，既知の研究シーズ用語を含む文献数と比較したり，文献の発行年分布を確認するなど，慎重に確認作業を行っていることは評価できる。

第 5 章では，結論を述べてから研究の限界について言及している。ここで言及されている以外にも，上で述べてきた課題もあり，本研究の成果は引き続き検証していく必要がある。しかし，抽出方法を体系的に提示していることや貴重な結果を得ていることなど，研究の意義は十分にある。

以上を総合的に判断すると，本論文は情報学の学位論文として十分な内容を有すると認められる。

【最終試験結果】

平成 30 年 12 月 18 日，図書館情報メディア研究科学位論文審査委員会において，審査委員全員出席のもと，本論文について著者に説明を求めた後，関連事項について質疑応答を行った。引き続き，「図書館情報メディア研究科博士後期課程（課程博士）の学位論文審査に関する内規」第 23 項第 3 号に基づく最終試験を行い，審議の結果，審査委員全員一致で合格と判定された。

【結論】

よって，本学位論文の著者は博士（情報学）の学位を受けるに十分な資格を有するものと認める。