4V-08

トピックモデルにおける Latent Dirichlet Allocation を用いた学術論文情報の可視化

宇佐美宏樹†1　　高岡詠子†2
上智大学大学院†1　上智大学†2

1. 研究背景

現在，日本の医師を取り巻く環境は問題を抱えており，特に論文執筆に関してはより重視しながら環境改善に取り組まなければならない．日本と米国の年代別における論文執筆数を見ると，各年代で米国よりも執筆数は劣っているが，特に「研修医」と呼ばれる20年代に大きな差が見られる[1]. この様な現状の原因として，研修医が 1 日の中で研究や自己研磨に当てることができる時間が非常に少ないこと[2]が挙げられる．そこで，現在の環境の中で，短時間で効率的に論文執筆の際に行う先行研究の調査や情報収集ができる様なソリューションが求められている．

2. 研究目的

上記の背景を受けて本研究の目的を，「医学論文の作成を行う日本の若い医師達が，短時間で効率よく資料を検索・収集できる様な環境を提供すること」とする．

3. 研究内容

多くの医師は，PubMed と呼ばれる無料論文検索エンジンを使用し，先行研究や実験結果の調査を行っている．研究や自己研磨の時間を短くするにはどうすれば良いかという点で，着目したのがこの資料検索に当てる時間である．

特に，学術論文の探索的検索において初期段階で行う文献レビューは，膨大な論文数の中から，希望する論文にたどり着くまで，何度もレビューを繰り返さなければならず，時間的コストが高い．特に若い医師や研究者，未知の研究分野に取り組む研究者にはこの文献レビューはより負担となる．

3.1 関連研究

Claus Boye Asmussen 氏らは,機械学習の手法の進歩によって，初期段階の文献レビューの時間的コストを抑えながら，多くの問題に対処することができるとした[3]. 取り入れた機械学習の手法は，教師なし手法の分野であるトピックモデリングの１つ LDA (Latent Dirichlet Allocation)であり，R を用いて探索的な文献レビューのあらゆる文脈に適用でき，完全な文献レビューにも使用できる汎用的なフレームワークを作成した[4].

また Zineb Sabri 氏らは，LDA と Python を用いた教師なしトピックモデリング法により，複数の科学データベースから 400 以上の研究論文を長年にわたって収集し､処理し 3 つの主要トピックを特定した[4].先行研究を参考にから，Pubmed における abstract 情報をデータセットとし，LDA と python パッケージである pyLDAvis を使用して，論文情報の可視化に取り組むこととした.

3.2 研究手法

今回の論文情報の可視化においては以下の手順で行った.

① NCBI (National Center for Biotechnology Information) のデータベースに対するユーザ向けの取得システム Entrez と Biopython を使用し，Pubmed 上にある論文の abstract 情報の取得

② 取得した abstract 情報から，ストップワード及び，記号，数字の除去

③ 全文章において，指定した出現回数に満たない単語の除去

④ LDA モデリングに必要な辞書とコーパスの準備.辞書に関して，Bag-of-Words を用いてベクトル表現に変換し，さらに，特定の文書にしか出現しない単語と文書全体で使用頻度が低い単語の除去を行った.

⑤ トピックス数の検討とモデリング
モデルの評価指標として一般的に用いられる Perplex と Coherence を使用しトピック数の検討を行い，Perplexity の値と Coherence の値の差が大きくなるところをモデルのトピック数とした.

⑥ 選択したトピック数で LDA モデルを行い，pyLDAvis を使用して可視化を行った．今回の可視化を行うにあたり，Metric Multi-dimensional Scaling (MMDS)の尺度を利用して可視化を行った.

今回 1 検索クエリあたりの論文情報の取得件数を 10000 件とした．これは Entrez の EFetch におけるレコード取得数の最大値が 10000 件であるからである．図１は，一例として検索クエリを ICI (Immune Checkpoint Inhibitor) とし 10000 件の論文データからトピックの可視化を行ったものである.

3.3 実証実験

作成した論文情報の可視化が，探索的検索の先行レビューにおいてどの程度軽減できるのかを確かめるために，以下の実証実験を行った.

†1 HIROKI USAMI, Sophia University Graduate School of Science and Technology / Information Science .
†2 EIKO TAKAOKA, Sophia University .

① 情報検索の専門家 3 人から 1 単語ずつ検索クエリを取得
② 取得した検索クエリを元に，各検索クエリにおける可視化情報を作成
③ 作成した可視化情報を①のユーザに提示し、自身が気になるトピックを選択してもらう
④ 選択したトピックにおいて，上位 3 個及び 5 個の単語を選択し，Pubmed の検索クエリとして検索する
⑤ 検索クエリが 1 つの場合と複数で検索したときの出力ページ 1 ページ目を見比べてもらい，ユーザが読みたい論文の出現数の個数を比較

取得した検索クエリは以下の通りである.

Clinical trial	empagliflozion	Medical device	NAFLD
Nanobody	epdemiology	orthopedics	SGLT 2inhibitor

3.4 検証結果

各検索クエリにおける論文比較は表 1 の通りとなった.リスト①検索クエリのみの論文リスト，リスト②検索クエリと「トピックの上位 3 件の単語」での検索した場合，リスト③検索クエリと「トピックの上位 5 件の単語」での検索した場合となっており,各論文数は PubMed で 1 ページ目に表示される 10 件をランダムに並べ換えたものとした.(epdemiology と Nanobody の③はヒット件数が 0 件であり,SGLT 2inhibitor は 1 件のみであった.)

表 1:各検索クエリに対する比較数

各検索クエリ	リスト① 内件数	リスト② 内件数	リスト③ 内件数
Clinical trial	0	3	10
empagliflozion	1	4	5
Medical device	0	2	2
NAFLD	0	5	7
epdemiology	0	1	無し
orthopedics	0	1	5
Nanobody	4	2	無し
SGLT 2inhibitor	0	5	1

4. 結論と展望

今回探索的検索における先行レビューにおいて、PubMed の abstract から，概念的な検索クエリを可視化し，ユーザの時間的コストの軽減を図った.文書からのクラスタリングが鋭敏に区別できるテーマにおいては、検索クエリを増やすとユーザの「読んでみたいと思える論文」が多く含まれる一方で，クラスタリングの分離度が大きくない可視化情報では，自身に有益な文書は 1 ページ目に含まれていなかった.そこで，今後の展望としては以下の 2 点が挙げられる.

① データセットの確保と単語数の適切な除去
今回取得したデータは最新論文 10000 件とした為、論文内容に偏りが生じてしまったと考えられる.可視化への時間的コストはかかってしまうが，全論文からのトピック抽出を行うことで，クラスタリングがより区別的になると考えられる.また，単語数の除去においても，デフォルト値での削除ではなく，専門家による監修の元適切な除去を行うことで，クラスタリングに大きな差が出るのではと考えられる.

② 追加検証実験
今回は，提示した論文リストから論文の評価を行なってもらったが，作成した可視化情報が先行レビューの時間的コストの軽減につながるのかをさらに確かめる為に，可視化情報の有無で分けたグループで，論文検索を行ってもらい，論文に辿り着くまでの時間を検証することで，可視化情報の有効性がより確認できると考えられる.

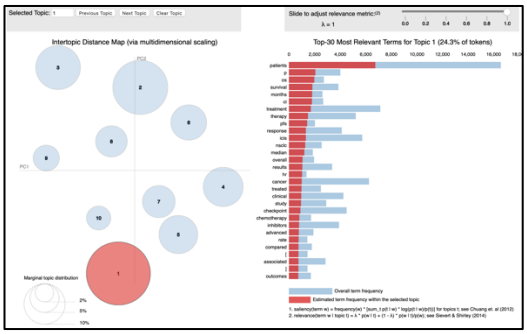

図 1: pyLDAvis によるトピックの可視化

参考文献

[1]20 代医師で大差,「過去 1 年間に、論文 1 本以上」https://www.m3.com/open/iryoIshin/article/569597 （参照 2020/12/26）
[2] 医師の労働時間の実態は？時間外労働を減らす方法もチェック https://doctor.mynavi.jp/column/workingtime (参照 2022/12/22）
[3] Asmussen, C.B., Møller, C. “Smart literature review: a practical topic modelling approach to exploratory literature review”, J Big Data 6, 93 (2019)
[4] Z. Sabri, Y. Fakhri and A. Moumen, “Serious Games in Higher Education: A Literature Survey using Topic Modeling,” 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, (2022)