大規模言語モデルを用いた学術論文検索における ブーリアン型検索クエリ作成の支援

福田悟志
中央大学 理工学部
fukuda.satoshi.3238@kc.chuo-u.ac.jp

概要

本論文では，学術論文検索におけるブーリアン型検索クエリの考案を支援するシステムを提案する．大規模言語モデルを活用し，「AND 結合」「OR 結合」「クエリ置換」「クエリ削除」の 4 つの操作パターンに基づくクエリ推薦と，その推薦理由を提示することで，ユーザが効率的にクエリを修正できる機能を提供する．本システムは少量のアノテーションデータを利用しており，ユーザの負担を抑えつつ柔軟なクエリ推薦を実現する．実験では，これらの操作パターンに基づくクエリ推薦において検索結果の再現率を向上させる語を提示できることを示した．

1 はじめに

学術論文検索において，適切な検索クエリの構築は，必要な情報を効率的に取得するために不可欠である．特に，検索クエリの考案は，論文の内容や研究テーマに基づく精密な検索を可能にする重要な作業である．しかし，最初に作成したクエリで関連論文を十分に収集できることは稀であり，ユーザは一部の検索結果を検証し，新たなクエリを作成または修正するという作業を繰り返し行う．すなわち，検索結果の品質向上には，クエリを修正・洗練する支援が求められる．従来のクエリ推薦システムでは，クエリ候補を提示する機能はあるものの，その推薦理由を十分に説明できないという課題がある．この欠点は，ユーザが推薦されたクエリを採用するか否かの判断を難しくし，最適な検索クエリの構築を妨げる可能性がある．

本研究では，学術論文検索を対象に，大規模言語モデルを用いたブーリアン型検索クエリの考案支援システムを提案する．本システムでは，ユーザが考案した初期クエリに基づき，検索精度を向上させるための候補クエリを提示するとともに，それぞれの推薦理由を説明文として付加する．また，ブーリアン型検索クエリの修正に必要となる 4 つの操作パターン「AND 結合」「OR 結合」「クエリ置換」「クエリ削除」をシステムの機能として取り入れることで，ユーザが直感的にクエリを修正できる環境を提供する．この機能により，ユーザは推薦理由を理解しながら，最適な検索クエリを選択・構築できるようになることが期待される．なお，クエリ推薦と理由文の生成においては，ユーザによる，情報要求を満たしているかどうかの判定を行った少量の論文のアノテーションデータを活用する仕組みを採用している．そのため，ユーザに多少の負担を求めるものの，アノテーションデータの内容を学習して，柔軟かつ説得力のあるクエリ推薦および理由文を生成できると考えられる．

本論文の構成は以下のとおりである．2 章で関連研究を述べ，3 章で提案システムについて述べる．4 章で提案手法の有効性を示すための実験について述べ，5 章で本論文をまとめる．

2 関連研究

学術情報検索では，検索結果に対して高い再現率を達成する場合，情報要求に関連する様々な検索クエリを考案することが求められる[1][2]．しかし，情報要求を端的に表し，かつ関連論文を網羅的に収集できるようなクエリの考案には，多くの時間や労力を要する．このような検索クエリの考案に対して，これまでに，ユーザの情報要求と関連するクエリを自動的に生成するシステムが開発されている[2][3]．しかし，推薦されたクエリがなぜ適切であるのかを説明する機能が不足している場合，ユーザにとって，それらのクエリを採用するか否かの判断が難しくなるという課題がある．また，推薦されたクエリに対して AND で結合した方が良いのか，既に考案されたクエリ内の語と置き換えた方がよいのかといった

具体的な操作が困難な場面が発生したとき，それが検索結果の品質向上を阻む要因となる可能性がある．特に，異分野や学際的な研究領域では，専門外の知識に基づいてクエリを評価する必要があるため，推薦されたクエリの意図を明確に提示することが重要といえる．

近年は，大規模言語モデルを用いた検索およびクエリの推薦・考案が活発に行われている．Wang ら[4]および Burau ら[5]は，ChatGPT を用いたシステマティックレビューのためのブーリアン型クエリの生成を行った．また，Wang ら[6]は，大規模言語モデルを用いて，ブーリアン型クエリを自然言語のクエリに変換し，ランク付けを改善する手法を提案した．Jagerman ら[7]は，さまざまなプロンプト（zero-shot, few-shot and Chain-of-Thought）を使用してクエリ拡張を行い，既存の Pseudo-Relevance Feedback 手法と比較を行った．本研究では，クエリ生成だけでなく，「AND 結合」「OR 結合」「クエリ置換」「クエリ削除」といった具体的なクエリ操作をユーザに支援することを目的としている点で，上記の研究とは，大規模言語モデルを用いる目的が異なる．

3 システム

3.1 概要

本システムの全体図を図 1 に示す．ユーザはまず，システムに対してブーリアン型の検索クエリおよび 1 文程度の情報要求を入力する．その後，システムは，入力された検索クエリを完全に満たす論文集合をデータベースから検索する．そして，ユーザに対し，それぞれの論文が情報要求を満たすかどうかの判定を要求する．その後，システムは，ユーザによる判定結果を用いて，大規模言語モデル(LLM)を通じた 4 種類のクエリ推薦を行う．

3.2 クエリ推薦

検索クエリの推薦において，本研究では，「AND 結合させた方が良い候補語（AND 結合）」「OR 結合させた方が良い候補語（OR 結合）」「検索クエリ内の語に代わる候補語（クエリ置換）」「検索クエリ内から除去した方が良い語（クエリ削除）」の 4 パターンを扱う．システムでは，それぞれのパターンに応じたプロンプトを作成し，「AND 結合」「OR 結合」「クエリ置換」「クエリ削除」における候補語とその理由文を生成する．以下では，各パターンにおける共通のプロンプト部分を示す．

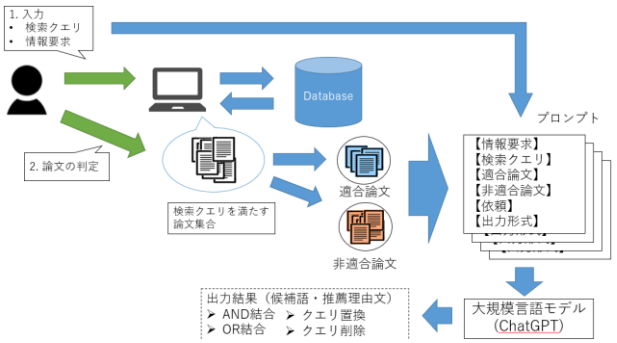

図 1: クエリ推薦までの全体図

・情報要求：図 1 において，ユーザが入力した情報要求を扱う．

・検索クエリ：図 1 において，ユーザが入力した検索クエリを扱う．

・適合論文：図 1 において，ユーザが適合と判定した論文集合を扱う．

・非適合論文：図 1 において，ユーザが非適合と判定した論文集合を扱う．

・出力形式：出力形式を指定する．具体的には，(1) tsv 形式の表として出力，(2)表以外の出力の必要はない，(3) 候補語は単数形，(4) 候補語は名詞，動詞，形容詞を指定，(5) 出力イメージに必ず従うこと．

そして，各パターンにおける特有のプロンプト部分を以下に示す．

[AND 結合]

・依頼：【情報要求】を満たす論文をより多く収集するために，【検索クエリ】で AND 結合すべき候補語と，その理由を，【適合論文】と【非適合論文】から推測し出力する．【検索クエリ】で既に使用されている語は出力しない．

[OR 結合]

・依頼：【情報要求】を満たす論文をより多く収集するために，【検索クエリ】における AND で結ばれた括弧ごとに，さらに OR で結ぶべき候補語とその理由を【適合論文】と【非適合論文】から推測し出力する．【検索クエリ】で既に使用されている語は出力しない．

[クエリ置換]

・依頼：【情報要求】を満たす論文をより多く収集するため，【検索クエリ】の中に他の単語に置き換えるべき候補語がある場合は，その理由とともに【適合論文】と【非適合論文】から推測し出力する．【検索クエリ】で既に使用されている単語は出力しない．

[クエリ削除]
・依頼：【情報要求】を満たす論文をより多く収集するために，【検索クエリ】の中に削除すべき単語がある場合は、その理由とともに【適合論文】と【非適合論文】から推測し出力する．

本研究では，大規模言語モデルとして，ChatGPT-4o を用いた．また，候補語がデータベースに出現していない場合，その語はユーザに出力しない処理を行っている．

4 実験

本実験では，推薦された候補語を用いることで，検索性能がどの程度改善されるのかに着目した．

4.1 データセット

実験には, NTCIR-1, -2 で提供されている情報検索用テストコレクション[8][9]を用いた．これらのコレクションには, 264,513 件の英語論文と図 2 の上部のような検索課題が 132 件含まれている．図 2 において，<TOPIC>は課題番号，<TITLE>は課題のタイトル，<DESCRIPTION>は情報要求を 1 文で説明したもの，<NARRATIVE>は情報要求の詳細な説明，<CONCEPT>は情報要求に関連するキーワードを表している．各検索課題には，約 1,000 から 5,000 件の学術論文に対して，その課題に対する「高適合」「適合」「部分適合」「不適合」のラベルが付与されている．なお，ラベル付けされた論文集合は，プーリングと呼ばれる手法を通じて，関連論文候補として収集されたものである[10][11].

本研究では，各検索課題において，「高適合」「適合」「部分適合」「不適合」が付与されている論文を，4 節で述べている検索対象集合として扱った．そして，「高適合」「適合」「部分適合」が付与された論文を関連論文，「不適合」のラベルが付与された論文を非関連論文とした．これは，情報要求と関連する可能性がある論文を網羅的に収集することが，研究の新規性を確認することを目的とした学術論文検索では求められていることに基づいている．また，ユーザがチェックできる論文の数の限界を 1,000 件と考え，さらに，ユーザが研究の新規性を確認する場合に，関連論文の数が極端に少ないというケースは稀であると考え，ラベルが付与されている論文が 1,000 件未満または関連論文を 10 件未満である検索課題を除いた．そして，一人の被験者が課題

<TOPIC q=0061>
<TITLE>階層関係の自動抽出</TITLE>
<DESCRIPTION>言語資源から語の階層関係を自動抽出する手法について論じられている文献。</DESCRIPTION>
<NARRATIVE>コーパス，辞書等の言語資源から，語の上位下位の階層関係に関する知識を獲得し，語の階層構造を構築する方法について知りたい。 (略) </NARRATIVE>
<CONCEPT>自然言語処理，コーパス，テキスト，辞書，言語資源，知識獲得，属種関係，全体部分関係，階層関係，階層構造，上位下位関係，シソーラス, (略) </CONCEPT>

検索クエリ: (hierarchical relationship OR hierarchical structure) AND (extract OR extraction OR acquire OR acquisition)

図 2　検索課題および被験者が考案した検索クエリの一例

内容を読み，図 2 下部のような検索クエリを考案した後，図 1 におけるユーザによる論文の判定に対する負担を考慮し，検索クエリを完全に満たす論文の数が 30 件以下であった 20 件の検索課題を用いた．また，TreeTagger[12]を用いて 264,513 件の英語抄録内の各単語を原形に戻し，品詞の解析および小文字への変形を行った．そして，品詞が名詞，動詞，形容詞，数詞以外の単語，ストップワードリストに含まれる単語[*2], 1 件の抄録にしか出現しない単語を除去した．その後，同一の抄録を持つ論文集合および 2 単語以下の抄録を除外し, 最終的に 259,546 件の英語論文抄録を用いた．なお，図 1 における，ユーザが入力として与える情報要求には，<DESCRIPTION>で記述されている文章を用いた．

4.2 評価尺度，比較手法

論文検索では，ユーザがチェックできる数の検索結果に含まれる関連論文の数が重要になる. そこで, 100, 200, 300, 400, 500 件の論文を検索結果として獲得したときの再現率で評価した．なお，*n* 位タイの論文が *k* 個ある場合，*n* 位から *n+k-1* 位の論文はこれらの *k* 個の論文をランダムに出力するものとした. そして，出力結果に含まれる関連論文の数に対する期待値を用いて再現率の計算を行った.

比較として，ユーザが初期にシステムに入力した検索クエリを用いて検索を行った場合をベースラインとした. そして, 3.2 節で述べたそれぞれのクエリ推薦において出力された候補語を用いた場合，置き換えた場合，削除した場合における，再現率の差を比較することで，図 1 で述べたシステムにより推薦されたクエリの性能を調べた．実験では，各検索課

[*2] Default English stopwords list (https://www.ranks.nl/stopwords)に基づいて構築した.

題において，再現率が向上または低下した候補語の割合と再現率の差の平均値を測定した．そして，41 件の検索課題に対する候補語の割合および再現率の差のマクロ平均を算出した．

4.3 検索方法，パラメータ設定

本研究では，情報検索の分野で広く研究されているクエリ尤度モデルに着目し，Zhai ら[13]の手法を適用した．このモデルは，ディリクレ平滑化によって文書に出現しない単語に確率値を割り当てるクエリ尤度モデルの一種である．以下では，このモデルを LM と名付ける．LM では，ディリクレ平滑化パラメータ μ と線形補間 λ の 2 つのパラメータを設定する必要がある．μ を 10，λ を 10 と設定した．

4.4 実験結果，考察

AND 結合，OR 結合，クエリ置換，クエリ削除の各パターンにおける実験結果を表 1 示す．表 1 における括弧内の数値は，再現率が向上または低下した候補語の数の割合を示しており，括弧の上の数値は，ベースラインとの再現率の差の平均値を示している．なお，AND 結合，OR 結合，クエリ置換，クエリ削除において，平均で 8.45 語，7.00 語，4.75 語，1.00 語が推薦された．また，AND 結合，OR 結合，クエリ置換においては，20 件の検索課題に対して候補語が出力された一方で，クエリ削除では 3 件の検索課題のみで候補語が出力された．表 1 では，候補語が出力されなかった検索課題では，0 語が候補語として出力されたとみなして平均値を算出した．

表 1 から，システムが出力した候補語に対して適切な選択を行うことで，検索結果として獲得する論文数の各条件において，AND 結合では平均で 7.02%，OR 結合では 4.24%，クエリ置換では 4.64%の向上が期待できることが確認された．一方で，不適当な選択を行った場合，AND 結合では平均で 5.50%，OR 結合では 8.58%，クエリ置換では 13.4%低下することも示された．なお，クエリ削除では，検索性能を改善する候補語が出力されておらず，推薦された語はいずれも再現率を大幅に低下させるものであった．今後は，プロンプトの改善を行うことで，検索性能を向上させる語がより多く出力されることを目指す．

また，AND 結合，OR 結合，クエリ置換において，同一の語が候補として出力されているケースが発生していた．これは，各パターンにおける候補語の推定を独立して行っているためである．今後は，どのパターンで推薦する方が有効かを判定するアプローチを導入し，ユーザに対し，より精巧なクエリ作成支援を行うためのシステムの改善を計画している．

表 1　LM 手法における検索結果として獲得する論文集合に対する再現率

		検索結果として獲得する論文数				
		100	200	300	400	500
Baseline		0.499	0.636	0.715	0.797	0.828
AND 結合	Improve	0.091 (0.327)	0.083 (0.288)	0.070 (0.249)	0.052 (0.180)	0.055 (0.161)
	Decrease	-0.059 (0.384)	-0.062 (0.392)	-0.054 (0.344)	-0.056 (0.392)	-0.044 (0.234)
OR 結合	Improve	0.048 (0.193)	0.054 (0.240)	0.050 (0.216)	0.026 (0.107)	0.034 (0.198)
	Decrease	-0.094 (0.333)	-0.091 (0.296)	-0.086 (0.281)	-0.094 (0.305)	-0.064 (0.197)
置換	Improve	0.055 (0.246)	0.050 (0.222)	0.056 (0.268)	0.038 (0.206)	0.033 (0.227)
	Decrease	-0.135 (0.423)	-0.131 (0.405)	-0.141 (0.356)	-0.126 (0.342)	-0.138 (0.266)
削除	Improve	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)
	Decrease	-0.066 (0.150)	-0.083 (0.150)	-0.088 (0.150)	-0.092 (0.150)	-0.083 (0.150)

5 おわりに

本研究では，学術論文検索において，ユーザが考案した検索クエリを効果的に洗練できる支援を提供するためのシステムを提案した．本手法では，ブーリアン型の検索クエリの再考案を支援するために，大規模言語モデルを用いて，「AND 結合させた方が良い候補語」「OR 結合させた方が良い候補語」「検索クエリ内の語に代わる候補語」「検索クエリ内から除去した方が良い語」の 4 つのパターンに対応したクエリ推薦を，推薦理由とともにユーザに提示する．実験では，各パターンにおいて推薦されたクエリを用いた場合における検索性能の改善を検証し，再現率を向上させるような候補語が出力されていることを確認した．

今後は，本システムのさらなる検証を行う．具体的には，様々なユーザに本システムを利用してもらい，システムを通じてユーザが最終的に考案した検索クエリを用いたときにおける検索性能を検証する．このとき，システムがユーザに提示した推薦理由がどの程度の妥当性および信頼性があるのかを，5 段階評価などを活用したアンケートを通じて明らかにしていく．

謝辞

この研究は科研費 JP19K20629 の助成を受けたものである.

参考文献

1. Verberne, S., Sappelli, M. and Kraaij, W.: Query Term Suggestion in Academic Search, *Proc. ECIR*, pp.560–566 (2014).

2. Kim, Y., Seo, J. and Croft, W.B.: Automatic Boolean Query Suggestion for Professional Search, *Proc. SIGIR*, pp.825–834 (2011).

3. Scells, H., Zuccon, G., Koopman, B. and Clark, J.: Automatic Boolean Query Formulation for Systematic Review Literature Search, *Proc. WWW*, pp.1071–1081 (2020).

4. Wang, S., Scells, H., Koopman, B. and Zuccon, G.: Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? *Proc.* SIGIR, pp. 1426–1436 (2023).

5. Budau, L. and Ensan, F.: Benchmarking Fully Automated Scholarly Search for Biomedical Systematic Literature Reviews, IEEE Access, Vol. 12, pp. 83764–83773 (2024).

6. Wang, S., Scells, H., Koopman, B., Potthast, M. and Zuccon, G.: Generating Natural Language Queries for More Effective Systematic Review Screening Prioritisation, *Proc.* SIGIR-AP, pp.73–83 (2023).

7. Jagerman, R., Zhuang, H., Qin, Z., Wang, X. and Bendersky, M.: Query Expansion by Prompting Large Language Models, arXiv preprint arXiv:2305.03653 (2023).

8. Kando, N., Kuriyama, K., Nozue, T., et al.: The NTCIR Workshop: The First Evaluation Workshop on Japanese Text Retrieval and Cross-lingual Information Retrieval, *Proc. Information Retrieval with Asian Languages Workshop* (1999).

9. Kando, N.: Overview of the Second NTCIR Workshop, *Proc. NTCIR Workshop*, pp.35–43 (2001).

10. Kuriyama, K., Yoshioka, M., and Kando, N.: The Effect of Cross-Lingual Pooling on Evaluation. *Proc.* NTCIR-2, pp. 297–310 (2001).

11. Kuriyama, K., Kando, N., Nozue, T., and Eguchi, K.: Pooling for a large-scale test collection: an analysis of the search results from the first NTCIR workshop. Information Retrieval, Vol.5, No.1, pp. 41–59 (2002).

12. TreeTagger Homepage, available from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (accessed 2025-01-09)

13. Zhai, C. and Lafferty, J.: A study of Smoothing Methods for Language Models Applied to Information Retrieval, *ACM Transactions on Information Systems*, Vol.22, No.2, pp.179–214 (2004).