

論文データベースのキーワード解析・可視化による論文執筆・投稿支援システムの提案と評価

佐野 祐作[†] Le Hieu Hanh[†] 横田 治夫[†]

[†] 東京工業大学 情報理工学院 〒152-8552 東京都目黒区大岡山 2-12-1

E-mail: †{daikichi,hanhlh}@de.cs.titech.ac.jp, ††yokota@cs.titech.ac.jp

あらまし 現在、非常に多くの学術誌や学術会議が存在し研究者は関連する出版物を追跡するのが困難である。そこで、検索可能な論文データベースが開発されたが、そのような情報に基づいて研究トピックの傾向の全体像を把握することは依然として困難である。これらのデータベースはいずれも、ユーザーが学会を時系列でまたはインタラクティブに比較することができない。したがって、研究者は論文の提出に適した会議を決定するためには多数の論文を読む必要がある。さらに、会議のトピックの最新の傾向を把握したり、よく使用されるキーワードを特定したりすることは困難である。そこで、本研究では KTrends と呼ばれるインタラクティブな視覚化システムを開発した。KTrends が提供する視覚化ツールにより学会間の比較が可能になり、研究者が論文を書いたり提出するのに貢献する。KTrends は一部会議からの情報を自動的に更新する。インタラクティブシステムに適した低コストのトピック検出手法を提案し、従来手法と比べた利点を実験を通じて示す。

キーワード 論文データベース、テキストマイニング、キーワード解析

1 はじめに

学術雑誌や国際会議に毎年発表される研究論文の数は非常に多く存在する。さらに、出版物の数は頻繁に変化し研究の傾向も大きく変化する。したがって、研究者が特定の研究分野に焦点を合わせていても、出版物の傾向を把握することは非常に困難である。

Google Scholar [3] や DBLP Computer Science Bibliography [2] など、様々な学術出版物に掲載されている研究論文に関する書誌情報を収集するデータベースが多く公開されている。

これらの論文データベースは、研究者がタイトルまたは著者名に特定のキーワードを含む研究論文を単に見つけたい場合に役立つ。しかし、この情報だけではトピックやトレンド傾向を把握するには情報量不足であった。特定の研究分野でこのような傾向を調査するには莫大な論文を読む必要がある。

一部の論文データベースは、参照する出版された論文の分類されたタイプに関する統計情報を提供しているが、キーワードの使用傾向や学会間比較のための情報は提供していない。研究者が様々な学会でのキーワード出現の傾向に関する情報や最新の注目すべき情報を入手できれば、論文の投稿先に適切な学会を探すことや論文執筆の負担を軽減することができる。方法の1つとして、学会のトピックキーワードを抽出することで、最新の注目すべき情報として提示することができる。

しかし、従来のトピックキーワード抽出手法は、対話的なユーザー入力がある場合、リアルタイム計算が必要になるものが多い。さらに学会の傾向やトピックを把握する情報としては抽出精度が不十分であった。

そこで、本研究はトピックキーワードの抽出のために

「Particularity-Score」を提案し、精度の向上を実験によって検証する。提案手法が従来手法に比べた利点は2つある。まず1つ目は、従来手法では難しかった出現頻度が低いが抽出率の高い単語を抽出し、対話的トピック解析が可能であること。2つ目は、対話的なシステムにおいてリアルタイムの大きな計算が発生しないためレスポンスが速いことである。

また、この Particularity-Score を導入し上記の論文執筆・投稿支援の情報不足のために KTrends と呼ばれる Web ベースのインタラクティブな視覚化システムを開発する。KTrends を使用すると、ユーザーはさまざまな視覚化ツールを使用して、学術会議を時系列で対話的に比較できる。トピックキーワード抽出の評価に加えて、このシステム全体の有用性を調べるためにユーザー調査を実施し評価を行う。

2 関連研究

2.1 背景知識

本研究では、既存手法の LDA や TF・IDF との比較を行う。そのため背景知識としてこれらの既存手法について説明する。

2.1.1 TF・IDF

TF-IDF は文書群において重要なキーワードを評価するための指標であり、それぞれ TF (Term Frequency) と IDF (Inverse Document Frequency) と呼ばれる指標である。

単語出現頻度 (TF) はある文書 d における単語 w の出現頻度を文書 d における全単語の出現頻度の総和で除算したものであり、単語 w が文書 d において多く使われるほど高くなる指標になっている。逆文書出現頻度 (IDF) は全体文書数を単語 w が出現する文書数で除算したものの対数をとったものであり多くの文書で出現する単語であるほど低くなる指標である。

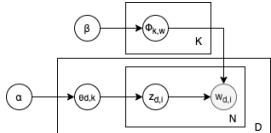


図 1 LDA のグラフィカルモデル

2.1.2 LDA (Latent Dirichlet Allocation)

LDA [4] は 1 つの文書が複数のトピックからなることを仮定したトピックモデルであり、文書群とトピック数 K を与えてキーワードに潜在的なトピックを決定する。文書 d でトピック k が出現する確率を $\theta_{d,k}$ としディリクレ分布に従うと仮定する。トピック k における単語 w の出現確率を $\Phi_{k,w}$ として $\theta_{d,k}$ と同様にディリクレ分布に従うと仮定する。文書 d の i 番目に出現する単語を $w_{d,i}$ として $w_{d,i}$ に対応する潜在トピックを $z_{d,i}$ とする。ここで α をトピック分布 $\theta_{d,k}$ のディリクレ分布に従うハイパラメータ、 β を単語分布 $\Phi_{k,w}$ のディリクレ分布に従うハイパラメータ、単語数を N 、文書数を D として以下のように LDA の生成過程をグラフィカルモデルで示す。

LDA による文書の生成過程は以下のようになっており単語のトピック分類がされる。ディリクレ分布を $\text{Dir}(\alpha)$ 、多項分布を $\text{Multinomial}(\theta_d)$ とした。

(1) トピックの確率分布を選択

$$\theta_d \sim \text{Dir}(\alpha)$$

(2) N 個の各単語に対して

(a) 潜在トピック $z_{d,i}$ を θ_d に基づいて選択

$$z_{d,i} \sim \text{Multinomial}(\theta_d)$$

(b) 潜在トピック $z_{d,i}$ に対する単語の確率分布に従って単語を選択

$$w_{d,i} \sim \text{Multinomial}(\theta_d)$$

2.2 関連研究

関連研究として, He ら [5] は、潜在的ディリクレ割り当て (LDA) モデル [4] を採用して、科学文献についてトピックを検出するためのフレームワークを提案した。伊藤ら [6] は、非負行列因子分解に基づいてトピックを検出するための手法を提案した。これらの 2 つのアプローチは効果的だが、特定のキーワードの傾向をインタラクティブに示すことができる視覚化インターフェースを提供しない。さらに、従来の方法はインタラクティブ機能の側面に必要なリアルタイム計算により、応答時間が長くなる可能性がある。

2.2.1 関連研究の課題

Google Scholar など [2, 3] が研究論文を収集するデータベースが多く公開されている。これら論文データベースと従来の関連研究のアプローチには、簡単に対処できない 3 つの問題がある。まず第一に、論文投稿を考えるユーザーは、提出に適した会議を探すときに多数の論文を読む必要があること。第二に、ユーザーはトピックの最近の傾向を把握するために多大な労力が必要であること。最後に、ユーザーは論文執筆の際にさまざまな会議で使用される一般的なキーワードや伝わりやすい表現を把握するのが難しいこと。

2.2.2 本研究のアプローチ

提案システムである KTrends は、低コストで高速なトピック分析方法を介して、システムのデータを up-to-date の状態に保ち、インタラクティブにすることでこれらの問題に対処する。

KTrends は「Conference Selection」、「Topic Discovery」、「Word Selection」の 3 つのユースケースで視覚化ツールを提供する。

「Conference Selection」の視覚化ツールは、論文を提出する研究者が提出する適切な会議を特定するのに役立つ。「Topic Discovery」の視覚化ツールを使用すると、ユーザーは会議に適したトピックキーワードを識別し、会議の性質を理解できる。「Word Selection」の視覚化ツールは、研究者が論文執筆時に使用する適切なキーワードを選択するのに役立つ。

KTrends は、これら 3 つのユースケースの視覚化ツールを提供する。KTrends とその使用法の概要は公開されている 5 分のビデオで示されている¹。

3 提案手法

本研究では論文データベースから書籍情報を取得し、トピックなどを可視化する手法を提案する。まず、トピックの可視化や解析に用いる手法を説明し、その後に実際のデータを利用した提案システムでの実現例とユースケースを説明する。

3.1 解析手法

システムに利用する TF/DN、関連語抽出手法 CoScore、トピックキーワード抽出手法 Particularity-Score についてまず説明する。TF/DN は UseCase1,2,3 に渡って利用されるトピックの強さを示す指標、CoScore は UseCase1 で用いられる関連語抽出手法、Particularity-Score は UseCase2 で用いるトピックキーワードの抽出のために利用される。

3.1.1 TF/DN

単に論文の数を示す Document Number (DN) と、トピックとしての「強さ」を示す Term Frequency (TF) という指標を算出する。 w が特定の単語で、 v が特定の会議として、

$$DN(v, i, w) = \begin{aligned} &\text{Number of papers} \\ &\text{for which } w \text{ appears in } v \text{ in year } i \end{aligned} \quad (1)$$

$$TF(v, w) = \frac{\text{Frequency of appearance of } w \text{ in } v}{\text{Frequency of appearance of all words in } v} \quad (2)$$

と示される。

3.1.2 CoScore

CoScore [7] は、関連するフレーズを抽出するために使用される。

$$CoScore = f(w_1, w_2) \cdot \log \frac{\frac{f(w_1, w_2)}{N}}{\frac{f(w_1)}{N} \cdot \frac{f(w_2)}{N}} \quad (3)$$

ここで、 $f(w_1, w_2)$ は、単語 w_1 と w_2 が同時に出現する回数を表している。 $f(w)$ は、単語 w が出現する回数を示し、 N は、すべての単語が出現する合計回数を示している。

1 : <http://yokota-www.cs.titech.ac.jp/ktrends#overview>

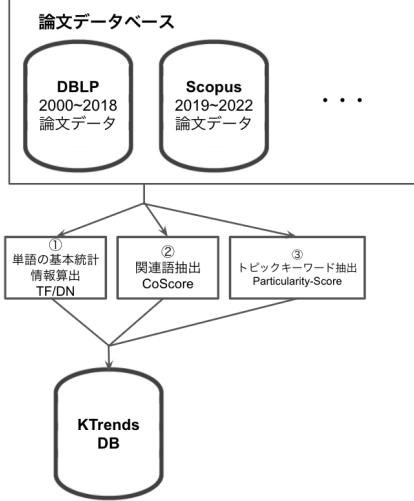


図 2 論文分析基盤

3.1.3 Particularity-Score

キーワードをシステムに入力することにより、共起する論文に限定してトピックを分析することが可能なトピックキーワード抽出提案手法として「Particularity-Score」を以下のように定義する。

$$\text{Particularity-Score}(v, i.w) = \frac{DN(v, i.w)}{\sum_{v' \in \mathbb{V}} DN(v', i.w)}, \quad (4)$$

\mathbb{V} は論文データセット全体の集合を示している。

Particularity-Score は、出現頻度は低いが抽出ニーズが高い単語を抽出できるため、従来の方法では困難であったインタラクティブなトピック分析が可能である。さらに、Particularity-Score は対話型システムにおいて高速なレスポンスを得ることができる。

3.2 提案システムでの実現例

提案手法は様々な論文データベースや学会のデータに対応することができるが、本研究の中で開発した KTrends では、AMiner [1] が提供する DBLP に 2000~2018 年の間に投稿された論文に関するデータを取得し、Scopus から 2019 年以降のデータを取得する。このデータセットは DBLP に蓄積された論文データの情報がデータマイニング目的のために公開がされている。また、Scopus の提供する API を使用して、SIGMOD、VLDB、ICDE から最新のデータを自動的に取得する。論文のデータは、タイトルとアブストラクトを利用した。

図 2 は、DBLP と Scopus から取得したデータを実際に 3 つのユースケースで実際に利用できる解析し、web アプリのサーバーで利用される KTrends データベースに格納するまでの流れと概要を提示する。

図 3 は、KTrends の UI である。それらは次のように説明される。

- (1) ユースケースに対応する視覚化ツールの選択
- (2) 選択した視覚化ツールを操作できる領域
- (3) 会議名やジャーナル名の入力フィールド



図 3 ユーザーインターフェース概要

また、実際にサーバーで行われるユーザーがキーワードを入力をしてから返答が返ってくるまでのリアルタイム計算手順を図 4 に示す。

4 ユースケース

ユーザーに利用用途を明瞭に伝え、可視化ツールを整理するために提案システム KTrends は 3 つのユースケースを想定する。

4.1 Use Case 1: Conference Selection

図 5 は、各視覚化ツールの UI を示している。

「Conference Selection」の視覚化ツールは、論文を提出する研究者が提出する適切な会議を特定するのに役立つ。例えば “database” に関わる論文を投稿したいユーザーに、システムがどここの学会で “database” に関するトピックが大きく盛り上がりをみせているのか明らかにすることによって、ユーザーは投稿先学会の判断材料を得ることができる。

4.1.1 Focus on one conference

図 5 (A) に示すように、ユーザーがシステムにキーワードを入力した後、選択した会議のキーワード頻度を関連するフレーズの頻度とともに視覚化される。研究者はその会議で提出が検討されている論文のテーマの人気の可能性を把握することができる。緑の線は、ターゲットキーワードの利用頻度の遷移を示している。黄色の線は、ターゲットキーワードの関連フレーズの利用頻度の遷移を示している。

ここで CoScore [7](式 (1)) は、関連するフレーズを抽出するために使用される。ユーザーは、図 5 (A) に示す黄色のボックスから平均化する関連フレーズを選択できる。また、1 つのグラフで複数のキーワードを視覚化して比較することもできる。

ユーザーは、ターゲットキーワードの時系列比較のために 2 つの指標が利用可能である。単に論文の数を示す DN(式 (2)) と、トピックとしての「強さ」を示す TF(式 (3)) である。

4.1.2 Focus on three conferences

図 5 (B) に示すように、ユーザーは選択した複数の会議の使用状況の変化を 1 つのグラフで視覚化できる。図の左下には、

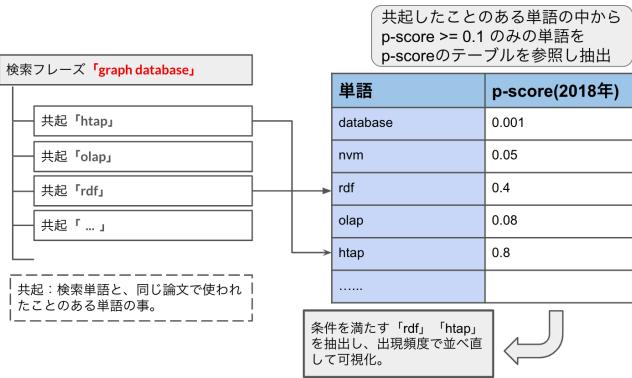


図 4 Particularity-Score の解析についてサーバーで行われるリアルタイム計算手順

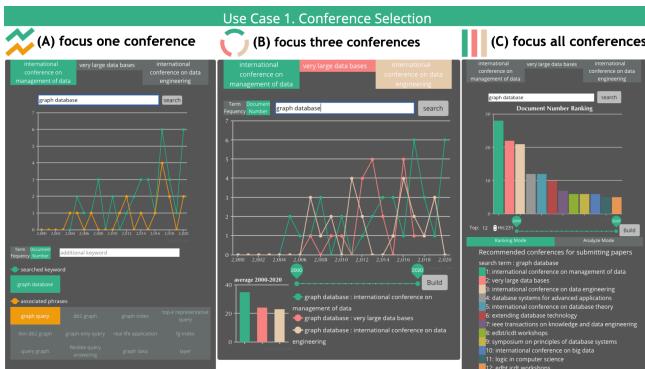


図 5 Use Case 1: 「Conference Selection」可視化ツール

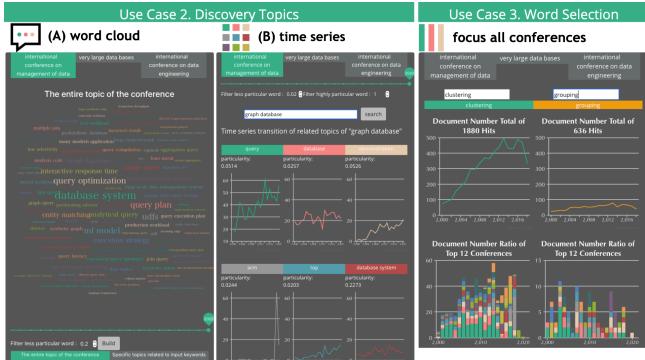


図 6 Use Cases 2 and 3: 「Topic Discovery」「Word Selection」可視化ツール

指定された期間中の各会議の平均出現頻度が棒グラフで視覚化されている。

図 5 (C) に示すように、システムはユーザーが完全な会議名を入力しなくても、ユーザーが入力したキーワードから送信するための適切な会議を自動的に推奨する。推奨プロセスでは、DN の上位ランキングに対応する学術会議が一覧表示される。

4.2 Use Case 2: Topic Discovery

「Topic Discovery」の視覚化ツールを使用すると、ユーザーは指定した会議で盛り上がっているトピックキーワードを発見し、その会議の一般的な性質や傾向を理解することができる。

例えば “database” に関する論文を投稿したいユーザーが投

表 1 実行環境

CPU	Intel(R) Xeon(R) CPU E5-2695 v2
Memory	264 GB
OS	Ubuntu 18.04.5
Python.ver	Python 3.6.9
DataBase	psql (PostgreSQL) 10.19

稿を考えている学会に対して、その学会のリストやキーワードをシステムに入力することによって感心のある分野のトピックキーワードを発見できる。

図 6 は、3 つの視覚化ツールの UI を示している。

4.2.1 Word cloud

図 6 (A) に示すように、KTrends は、会議のトピックキーワードを抽出し、それらをワードクラウドとして視覚化する。特定の単語のフィルタリングを強化することにより、システムは特徴的な単語のみに焦点を当てることができます。つまり、ユーザーは一般的で無関係な単語を除外できる。

また、これらのキーワードをシステムに入力することにより、共起する論文に限定してトピックを分析することも可能である。このフィルタリングの指標は、Particularity-Score (式 (4)) を利用する。

4.2.2 Time series

図 6 (B) に示すように、KTrends はワードクラウドの視覚化を使用して理解するのが難しい時系列比較のための視覚化ツールも提供する。この視覚化ツールを使用すると、ユーザーは会議のキーワードの注目すべき遷移を発見することができる。

4.3 Use Case 3: Word Selection

図 6 のユースケース 3 に示されているように、KTrends は研究者が論文を作成するときに適切なキーワードを選択するのに役立つ。研究者は論文執筆で言葉選びに迷っている際に、指定したキーワードがどれほど広く頻繁に使用されているかを比較することによって、理解しやすい適切な言葉や表現を選択することができる。例えば、論文執筆をしているユーザーが “clustering” と “grouping” という似た意味の言葉を利用するか迷っていた場合、これらのキーワードを入力することによって利用頻度や広い学会で使われているかを調べて判断のための情報を得ることができる。

5 評価実験

提案手法 Particularity-Score の効果と、システムに導入した際にレスポンスおよびサーバーへの負荷について、従来手法と比較評価を行う。また、それぞれのユースケースとそれらで提供している UI の有用性をユーザー調査によって評価を行う。実験環境を表 1 に表す。

なお、ユーザー調査に関しては、東京工業大学の人を対象とする研究倫理審査委員会の承認を得ている。

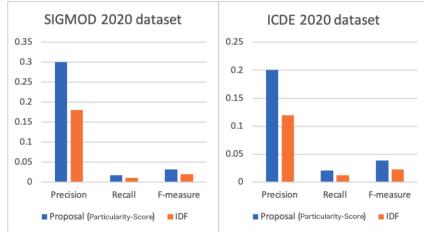


図 7 「Particularity-Score」と「IDF」の精度評価

5.1 トピックキーワード抽出精度評価

5.1.1 実験目的

提案手法 Particularity-Score のトピックキーワード抽出手法として精度の点からの有用性を調査するために、従来手法の IDF との比較実験を行い評価する。

5.1.2 実験対象のデータ

Scopus に掲載されている 2 つの国際会議 (SIGMOD2020 と ICDE2020) の実際のデータセットからトピックキーワードを抽出した。これら 2 つの会議の論文の著者によって指定されたキーワードを正解のデータセットとした。正解データセットは、SIGMOD 2020 (ICDE 2020) の 307 (241) の論文から抽出された 887 (469) のキーワードで構成されている。

5.1.3 実験内容

本研究の提案の適合率と再現率の結果は、標準の逆文書頻度 (IDF) ベースの計算を使用した単純なアプローチと比較される。提案する Particularity-Score では、抽出されたトピックキーワードのセットとして閾値未満のキーワードを抽出し検討した。IDF では、スコアが閾値よりも大きいキーワードのみを抽出し検討した。それぞれの方法について、計算されたスコアのからランキング付けされた 50 のトピックキーワードを定義した。

この実験では、Particularity-Score の閾値は各データセットに対してそれぞれ 0.045 および 0.055, IDF 閾値はそれぞれ 0.025 と 0.035 に設定されている。

5.1.4 実験結果

図 7 は、ランキング付けされた 50 個のキーワードを使用した提案手法 (Particularity-Score) と IDF の適合率、再現率、および F-measure の結果を示している。この図は、提案手法がより多くの適切なトピックキーワードを抽出していることが分かる。提案手法の適合率は IDF よりも優れており、両方のデータセットで 1.67 倍の精度の結果を達成できている。

5.2 トピックキーワード抽出レスポンス評価

5.2.1 実験目的

提案手法 Particularity-Score のトピックキーワード抽出手法として、実際にシステムに採用した際のレスポンスの点での有用性を調査するために、従来手法の LDA との比較実験を行い評価する。

5.2.2 実験内容

SIGMOD, VLDB, ICDE の 2018 年に投稿された論文に利用されたキーワードをシステムに 5 回の試行回数で入力し、提案手法 Particularity-Score と従来手法 LDA でトピックキーワード

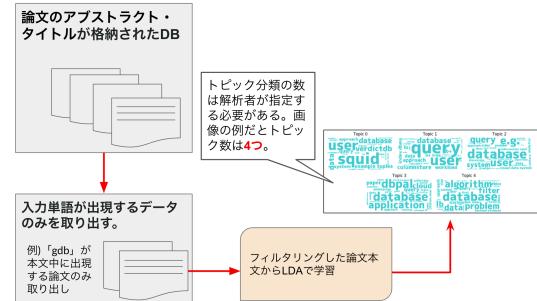


図 8 LDA の解析についてサーバーで行われるリアルタイム計算手順

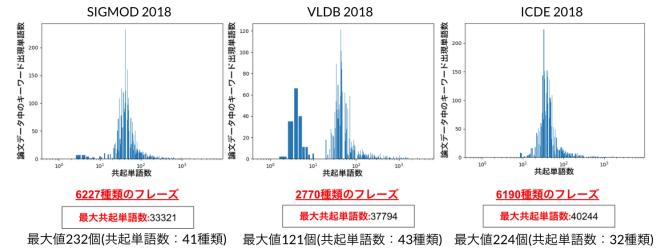


図 9 共起単語数分布

抽出を行なった際のレスポンス速度を比較した。Particularity-Score は同じ論文で共起を起こしたことのある単語数でレスポンスが変化するため、共起数に区分を分けて比較をした。

測定する実際にユーザーが入力をしてから返答が返ってくるまでのリアルタイム計算手順を、従来の手法の Particularity-Score を図 4、従来の手法の LDA を図 8 に示す。

5.2.3 実験対象のデータ

SIGMOD, VLDB, ICDE の 2018 年に投稿された論文のタイトルとアブストラクトを利用する。キーワードと共にしたことのある共起単語数の分布は図 9 に示されている。

5.2.4 実験結果

提案手法 (Particularity-Score) と LDA のレスポンス結果を図 10 に示している。横軸はエリとして入力したキーワードの共起単語数で縦軸は応答速度を示し、エラーバーは 5 回の試行回数のうち最大値と最小値を示している。

SIMOD2018 の 20,000 以上共起数がある 2 個のキーワードを除いて、提案手法がレスポンスが優れていることが分かる。また、Particularity-Score は共起単語数の増加に合わせて単調に応答速度が増加する傾向にあることが分かる。また、LDA はキーワードの共起単語数がレスポンスや CPU 使用率にそこまで影響を出さないことが分かる。

また、各データセットについて共起単語数の最大値をとるキーワードをシステムに入力し性能を調べた。

まず、SIGMOD の共起単語数で最も種類が多い、41 キーワードの共起単語を持つ 121 種類の中の “high ingest rate” というキーワードを 5 回入力し性能を調べた。レスポンスは平均 0.232(sec)、CPU 使用率は平均 0.310(%)、メモリ使用量は平均 83.300(MByte) であった。次に、VLDB の共起単語数で最も種類が多い、43 キーワードの共起単語を持つ 232 種類の中の “lim-

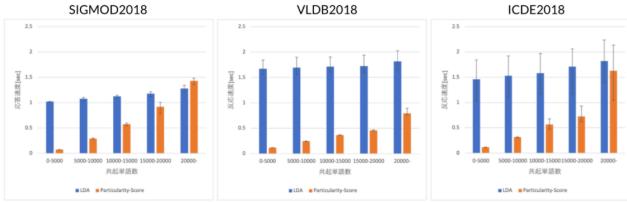


図 10 トピックキーワード抽出レスポンス比較

iting”というキーワードを5回入力し性能を調べた。レスポンスは平均0.233(sec), CPU使用率は平均0.233(%), メモリ使用量は平均78.473(MByte)であった。最後に,ICDEの共起単語数で最も種類が多い,32キーワードの共起単語を持つ224種類の中の“eccentricity computation algorithm”というキーワードを5回入力し性能を調べた。レスポンス0.0304(sec), CPU使用率は平均0.231(%), メモリ使用量は平均78.634(MByte)であった。

これらの結果からほとんどのクエリにおいて,0.25s以下でレスポンスを返すことができ,CPUは1%以下, メモリは80MB以下で解析できることが分かる。

表 2 共起単語数分布

	0-5,000	5,000-10,000	10,000-15,000	15,000-20,000	20,000-
SIGMOD	5785	29	11	5	2
VLDB	2,721	30	10	7	2
ICDE	5,797	26	12	5	2

5.3 トピックキーワード抽出 CPU・メモリー使用率比較

5.3.1 実験目的

提案手法Particularity-Scoreのトピックキーワード抽出手法として,実際にシステムに採用しwebアプリとして運用した際のサーバーへの負担の点での有用性を調査するために,従来手法のLDAとの比較実験を行い評価する。

5.3.2 実験内容

共起単語数が表2の区分に属する5種類のクエリをシステムに入力し,応答を返すまでの性能を提案手法と従来手法LDAとの比較実験を行い評価する。

5.3.3 実験対象のデータ

SIGMOD, VLDB, ICDEの2018年に投稿された論文のタイトルとアブストラクトを利用する。

5.3.4 実験結果

提案手法(Particularity-Score)とLDAの性能を図11に示している。これら図によると共起単語数が少ない単語を除いては提案手法がサーバーへの負荷を抑えて解析をすることができていることが分かる。また,従来手法のLDAは共起単語数が増加すると大きくCPU使用率が増加する。

横軸はクエリとして入力したキーワードの共起単語数で,縦軸は応答速度,CPU使用率, メモリ使用率・量を各図で示し, エラーバーは5回の試行回数のうち最大値と最小値を示している。

これらの実験結果を見ると, 提案手法はおよそそのクエリで

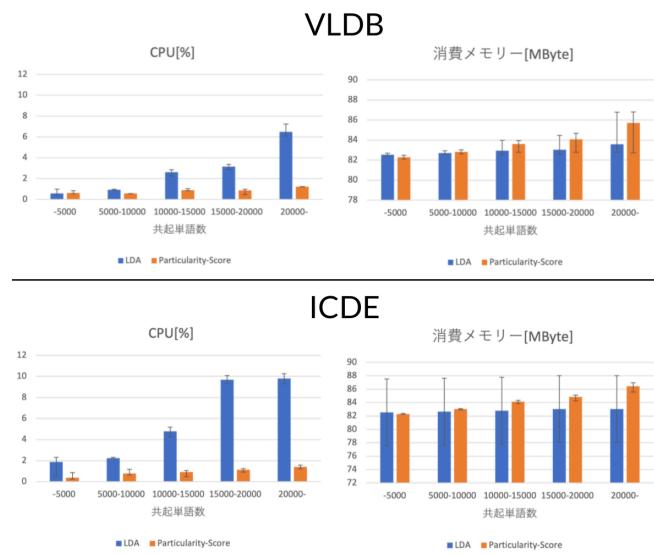
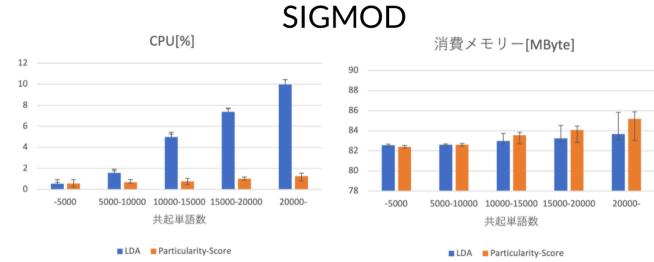


図 11 トピックキーワード抽出 CPU 使用率・メモリー使用量比較

CPU負荷がLDAより少ないことが分かる。その一方で、消費メモリに関しては、提案手法はややLDAよりもサーバーにかける負担が大きい。

5.4 事前学習性能比較

5.4.1 実験目的

提案手法Particularity-Scoreのトピックキーワード抽出手法として,実際にシステムに導入するまでの事前学習コストの点での有用性を調査するために,従来手法のLDAとの比較実験を行い評価する。

5.4.2 実験内容

システムで利用できる様になるための事前学習の性能を,論文数100,500,1000,2000のデータ量で提案手法と従来手法LDAで5回の試行回数で比較した。

5.4.3 実験対象のデータ

SIGMOD, VLDB, ICDEの2000~2018年の間に投稿された論文から無作為に選んだ論文の中のタイトルとアブストラクトを利用する。

5.4.4 実験結果

図12は提案手法とLDAの事前学習の性能を図12で示している。横軸は解析を行う論文数で, 縦軸は応答速度を示し, エラーバーは5回の試行のうち最大値と最小値を示している。

これらの実験結果から, 事前学習の点では提案手法は従来手法のLDAに比べコストがかかってしまうことが分かる。また, 解析する論文数が増加するにつれて大きく解析時間も増加する

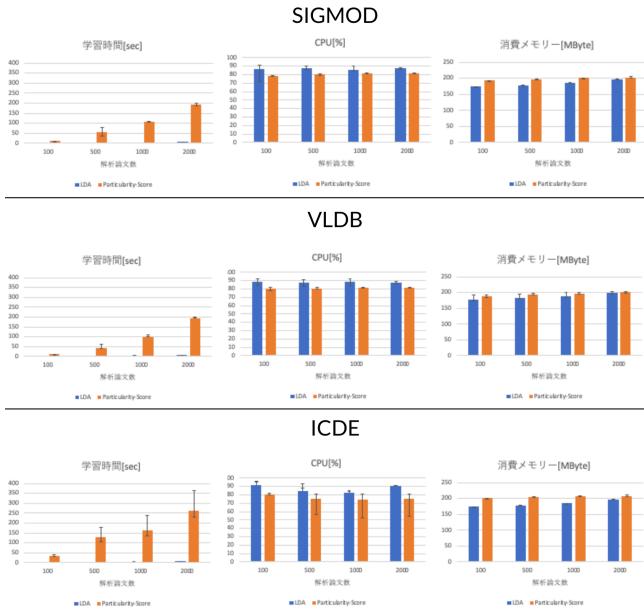


図 12 事前学習速度・CPU 使用率・メモリー使用量比較



図 13 トピックキーワード可視化 UI(A:従来手法 LDA B:提案手法 Particularity-Score)

ことが分かる。

5.5 ユーザー調査による評価

5.5.1 実験目的

提案システム UI のユーザビリティを調査するために、アンケートによるユーザー調査を行い評価する。

5.5.2 実験内容

評価者を用意し google form を利用して以下の設問に回答してもらう。

a) 評価 1. 3つの課題設定は適切か?

(1) KTrends が想定する 3 つのユースケースについて五段階評価で適切か点数を付けてもらう。

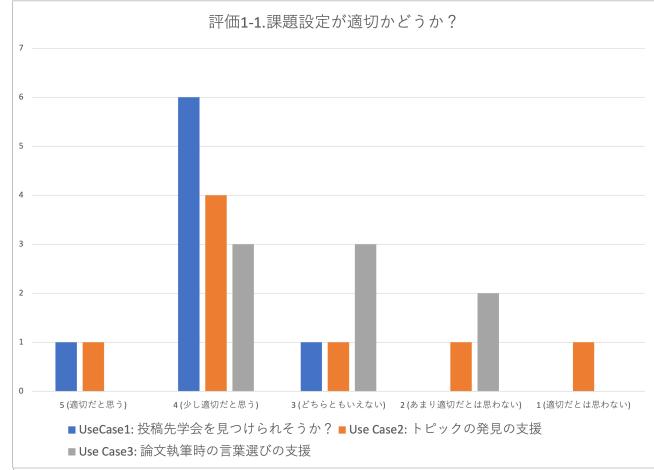
(2) 1 で回答した点数の理由

(3) 図13の A(従来手法 LDA) と B(提案手法 Particularity-Score) のどちらの可視化が役立つか。

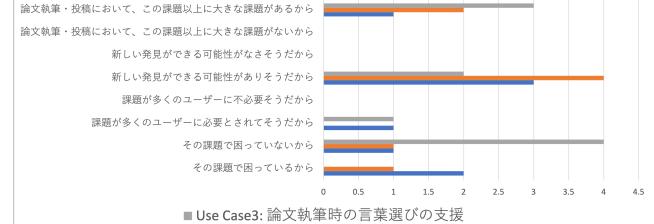
図 13 の A と B の可視化はどちらも SIGMOD2020 の「data base」に関するトピックを抽出している。

b) 評価 2. 従来システムとの比較

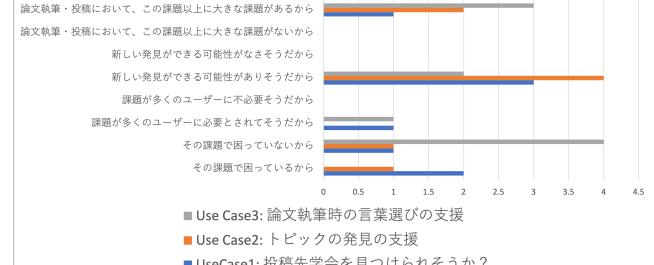
(1) KTrends が想定する 3 つのユースケースのそれぞれについて、論文執筆・投稿支援をする提案システム (KTrends) と、論文検索システムである従来システム (Google Scholar, DBLP, Scopus など) を比較して、3 つのユースケースで利用価値があ



評価1-1.課題設定が適切かどうか？



評価1-2 1-1の回答の理由



評価1-3.役に立つ可視化は？

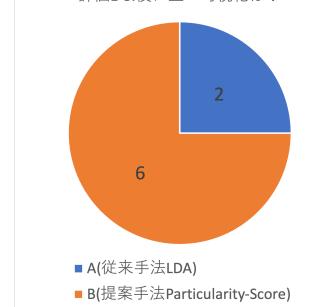


図 14 アンケート結果 (評価 1 点数)

るか五段階で回答してもらう。

(2) 1 で回答した点数の理由

5.5.3 実験対象のデータ

日本データベース学会が会員相互の情報交換を目的として運用しているメーリングリストである DBJapan において、2022 年 1 月 6 日～2022 年 1 月 18 日までの間で評価者を募り、4 名の協力者を得た。それに東京工業大学の学生・教員を加え、全体で 8 名から評価結果を得た。

5.5.4 実験結果 (点数)

図 14 に評価 1 の点数のアンケート結果と、図 15 に評価 2 の点数のアンケート結果を掲載する。

5.6 ユーザー調査全体の考察

十分な人数の評価者を用意することはできなかったが、トピックキーワード抽出について提案手法による可視化が 8 人の評価者のうち 6 人が実用的であるという回答を得ることができた。また、システム全体の有用性については、UseCase1 は肯定的な回答を得ることができたが、UseCase1,2 では評価が分散したこ

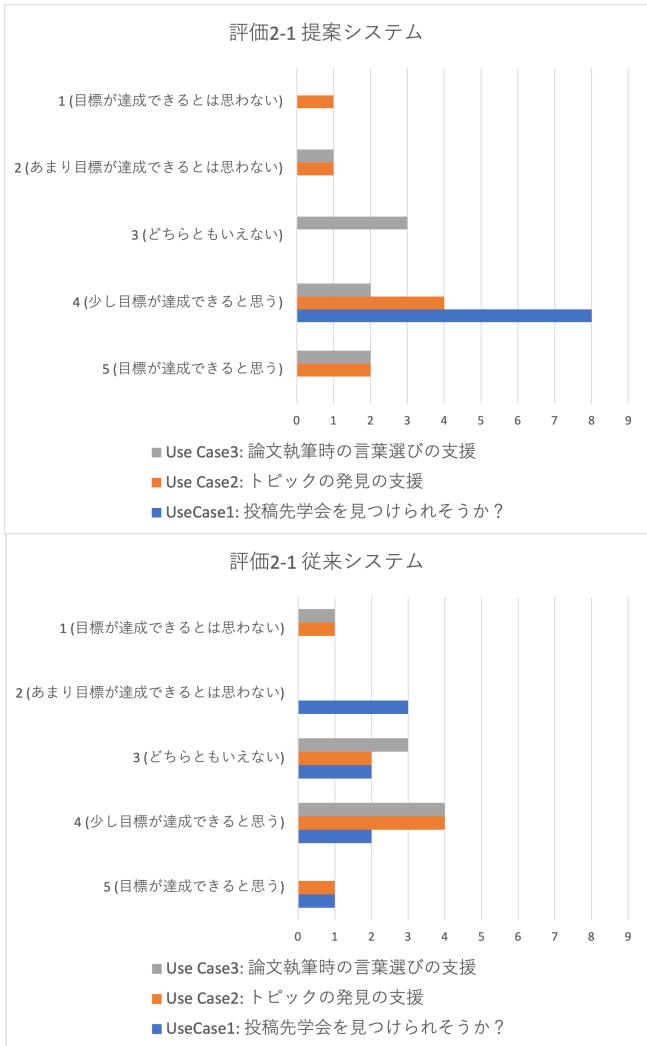


図 15 アンケート結果(評価 2 点数)

とが分かった。「新しい発見ができそうだから」という理由で肯定的な回答をした回答者が多いかった。

6 おわりに

6.1 まとめ

本研究では、学会やジャーナルのトレンドをインタラクティブに視覚化することを目的とした KTrends と呼ばれる新しいシステムを開発した。システムで提供されるインタラクティブな視覚化機能は、研究者が適切なキーワードに関する論文を書いたり、適切なジャーナルや年次会議に提出したりするのに役立つ。この論文では、KTrends の使用が研究者の活動をサポートするのに役立つ 3 つのユースケースを検討した。KTrends は 3 つの会議から最新の論文を自動的に取得し、ユーザーが最新の情報でトピックを分析できるようにする。インタラクティブなシステムで、迅速な応答と低い運用コストでトピックキーワードを提案手法「Particularity-Score」を用いて抽出する。

このような機能を提供する既存のシステムはなかった。従来の関連研究は、学会間でキーワードの傾向を比較するためのインタラクティブな視覚化ツールを提供せずに、トピックキー

ワードを検出することに焦点を当てていたものなどが存在していた。KTrends ではキーワードと調べたいターゲット学会やジャーナルを動的に調整して学会間の比較を行うことができる。KTrends は、Web アプリとして以下の URL にリリースされている²。

これらの提案手法とシステムの評価を行なった。「Particularity-Score」の精度のための評価では、論文上部のキーワードを正解データセットとした評価で IDF よりも高い精度を出すことができた。また、評価者が十分な人数ではないが提案手法と従来手法の可視化の比較では 8 人の評価者のうち 6 人が提案手法が実用的であるという回答を得ることができた。

提案手法はインタラクティブシステムにおいて、従来手法の LDA に比べてほとんどのクエリでレスポンスが速く、CPU 使用率の負担も小さいことが評価実験によって分かった。しかし、システムでリアルタイムの計算ができる様になるまでの、事前学習のコストは従来手法よりも大きくかかることが分かった。

システム全体の有用性についてのユーザー評価では、UseCase1 が課題設定で優位に評価された。

6.2 今後の課題

ユーザー評価の評価者が少なく、提案システムの有用性についての評価・比較が不十分であることが今後の課題として挙げられる。特に、Particularity-Score の精度評価のための正解データセットをユーザー調査によって収集し精度評価する予定だったが、正解データセットとして利用できるほど十分な量が収集できなかった。精度評価として論文上部のキーワードを正解データセットとして利用しているが、これは「トピックキーワード」と呼べるものを見たときに選んだものではないため信用性に欠ける。ユーザー評価の評価者を増やして正解データセットを収集できれば、提案手法の有用性をより正確に示すことができたと考えられる。

文 献

- [1] AMiner. <https://www.aminer.cn/citation>.
- [2] DBLP: Computer science bibliography. <https://dblp.uni-trier.de/>.
- [3] Google Scholar. <https://scholar.google.com/>.
- [4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [5] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009*, pp. 957–966, 2009.
- [6] Hiroyoshi Ito, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Detecting Topic Evolutions in Bibliographic Databases Exploiting Citations. In *Proceedings of the 26th International Conference on Information Modelling and Knowledge Bases (EJC 2016)*, pp. 465–480, 2016.
- [7] Yuichiro Machida, Daisuke Kawahara, Sadao Kurohashi, and Manabu Sasano. Design of Word Association Games Using Dialog Systems for Acquisition of Word Association Knowledge Latent Dirichlet Allocation. *IPSJ Journal*, Vol. 57, No. 3, pp. 1058–1068, 2016.

2 : <http://yokota-www.cs.titech.ac.jp/ktrends#overview>