LDAを用いた図書館情報学の研究トピックの変化：2期間の雑誌論文の全文を対象に

宮田洋輔 帝京大学 m@miyay.org

山本通正 慶應義塾大学大学院

楊芳 慶應義塾大学大学院

岩瀬梓 慶應義塾大学大学院

石田栄美 九州大学

倉田敬子 慶應義塾大学

上田修一 元慶應義塾大学

はじめに

学問分野が扱うトピックや主題の動向は，内容分析[1)]や計量書誌学[2)]の手法によって分析されることが多かった。内容分析の場合は，比較的小規模になることが多く，分析の枠組みは，前もって作られたものを用いるか，調査者が質的に作り上げることになる。一方，計量書誌学的手法の場合は，分析対象から直接的に枠組みが構築されるが，著者や引用文献など間接的な情報の関係性から導かれることになる。

この課題に対して，最近ではLDA（潜在的ディリクレ配分法）に代表されるトピックモデリングのアプローチを用いることが増えてきている。トピックモデリングの手法を用いることによって，テキストデータなど大量のデータから，データの特徴に基づいて自動的に扱われていているトピックを導き出すことができる。また，LDAでは1つの文書が複数のトピックを持つことも可能であり，従来的なクラスタリングの手法に比べて柔軟さを持っている。

BleiとLaffertyは，*Science*誌に掲載された論文のデータを用いて，LDAによって，科学研究の動向の変化を分析した[3)]。その後，計算機科学，統計学など特定の分野に対しても同様のアプローチによる分析が発表されている。

図書館情報学でもいくつかの先行研究がある。Sugimotoらは，北米の図書館情報学分野の学位論文に対してLDAを用いた分析を行っている[4)]。Yanは，*Journal Citation Report*（*JCR*）の図書館情報学カテゴリの雑誌から得たタイトルを用いてLDAで研究の動向を測定した[5)]。Figuerolaらは，LISAから抽出した論文の抄録をLDAで分析している[6)]。

図書館情報学分野の先行研究は，いずれもメタデータレベルの分析である。学術雑誌の電子化が進む近年では，全文のテキストが入手できる雑誌も増えている。また，先行研究での分析対象にばらつきがある。*JCR*では*MIS Quarterly*のような経営情報システムの雑誌が上位に収録されていたり，*LISA*では*Library Journal*のような査読誌ではない雑誌に掲載された研究論文以外のデータが含まれ，図書館情報学研究のトピックを必ずしも反映しているとはいえない。

図書館情報学分野を代表する雑誌に掲載された原著論文を対象とすることによって，図書館情報学の第一線で扱われるトピックを見ることができる[7)]。本研究では，2期間における雑誌論文の全文テキストを取得し，LDAを用いて，図書館情報学の第一線で扱われるトピックの変化を分析する。

方法

図書館情報学分野での学術雑誌に関する調査などを参考に，図書館情報学の代表的な雑誌である*Journal of the Association for Information Science and Technology*（*JASIST*），*Information Processing & Management*（*IPM*），*Journal of Documentation*（*JDOC*），*Library Quarterly*（*LQ*），*Library & Information Science*

Research（*LISR*）の 5 誌を選んだ。2000 年から 2002 年と 2015 年から 2017 年の間に，この 5 誌に掲載された原著論文の全文テキストを収集した。2000 年から 2002 年は，ウェブと検索エンジンが普及し情報環境が大きく変化しはじめた時期であり，2015 年から 2017 年は最新の 3 年間として，今回の分析対象とした。2 期間での論文数を表 1 に示した。

表 1 期間ごとの論文数

雑誌	2000-2002	2015-2017
JASIST	280	567
IPM	108	183
JDOC	89	190
LISR	48	108
LQ	36	39
合計	561	1,087

各雑誌の電子ジャーナルサイトの HTML ファイルから全文テキストを得た。全文テキストに対して機能語と数字を含む語を除去し，語幹処理を行った。また，9 割より多くの文献に出現する語と 9 回未満の低出現語を除去した。

期間ごとに Python の gensim[8)]を用いて，LDA を実行した。反復数は 500 として，それ以外は gensim の標準パラメータを用いた。トピック数は，先行研究や用いたデータの規模に基づいてそれぞれ 30 トピックずつとした。

結果

LDA で得られた 2 期間のトピックを分析した。トピックを理解しやすくするためにラベルを付与した。トピックのラベルは，各トピックの頻出語とトピックに割り当てられる確率が高い論文のメタデータと本文に基づいて検討し，著者全員で合議して決定した。

2000 年から 2002 年のトピック 17 を例にラベル付与の過程を示す。このトピックの中で頻繁に用いられる語の上位 5 語とその出現確率を表 2 に示した。語幹処理を施す前には単数形，複数形などの形が考えられるが，その一例を語幹処理前の例の列に示した。ここから，学生または利用者がデータベースや IR システムを探索することとの関連が推察される。

表 2 トピック 17 の頻出語と出現確率

頻出語	出現確率	語幹処理前の例
student	0.020	student
search	0.018	search
devic	0.012	device
user	0.011	user
databa	0.011	database

さらにこのトピックに割り当てられる確率が 0.5 以上の論文のタイトル，抄録，本文を見た。割り当てれる確率が 0.7 以上の論文を表 3 に示した。利用者と情報検索システムの相互作用を Kintsh の理論を使ってモデル化する，拡張した情報検索システムの実際の学習場面で

表 3 トピック 17 に割り当てられる確率の高い論文

確率	著者名	論文名	掲載誌
.9997	Cole, C; Mandelblatt, B	Using Kintsch's discourse comprehension theory to model the user's coding of an informative message from an enabling information retrieval system	*JASIST*
.9997	Cole, C	Intelligent information retrieval: Part IV. Testing the timing of two information retrieval devices in a naturalistic setting	*IPM*
.9997	Cole, C; Cantero, P; Ungar, A	The development of a diagnostic-prescriptive tool for undergraduates seeking information for a social science/humanities assignment. III. Enabling devices	*IPM*
.9996	Cole, C	Interaction with an enabling information retrieval system: Modeling the user's decoding and encoding operations	*JASIST*
.8619	Hood, WW; Wilson, CS	The scatter of documents over databases in different subject domains: How many databases are needed?	*JASIST*
.7221	Nicholson, S	Raising reliability of Web search tool research through replication and chaos theory	*JASIST*

の効果の実証, 学部生の情報探索に活用できる診断ツールの開発などが上位に挙がった。ここから，合議の上，学生を中心とする検索システムの探索行動をモデル化しているトピックと考え，ラベルを「学生の検索システム探索行動モデル化」とした。

さらに，トピック間の関係を分析し，トピックをまとめるカテゴリについて検討した。pyLDAvis[9]を用いて 2 次元プロットを作成した。トピック間の距離の近さに基づいて，それぞれのトピックをまとめるカテゴリを付与した。2 次元プロット図とカテゴリ，トピックのラベルを図 1 に示した。図中の番号は図下部のトピックラベルの一覧と一致している。また円の大きさは, 全体に対するトピックの比率に比例している。

2 期間でのカテゴリを比較すると，以下のような変化を挙げることができる。

- 「図書館」カテゴリのトピックが減少し，図書館サービスに関する研究から, 図書館の社会的役割や歴史に関する研究が増加した。
- 「情報検索」ではアルゴリズムに関するトピックが減少し, 利用者志向からの分析が増加した。これにより，「情報探索行動」のカテゴリに含まれるトピックが増え，「利用者と情報探索」となった。
- 「情報探索行動」でも情報探索行動のモデル化から，特定の状況・コンテキストでの情報探索への変化が見られた。
- 「学術コミュニケーション」には 2 期間とも同程度のトピックが含まれていたが, 計量書誌学的法則に関する研究が減少し, 研究評価に関する研究が増加した。また 2000 年から 2002 年にはビブリオメトリクスのトピックが存在していたが，2015 年から 20017 年には見られなくなった。
- 「情報検索」カテゴリに近いトピックとして「ツイート分析」のように SNS から得たデータを分析したトピックが出現した。

引用文献

1） Tuomaala, O., et al. Evolution of Library and Information Science, 1965–2005. *JASIST*. 2014, vol. 65, no. 7, p. 1446-1462.

2） Börner, K., Chen, C. & Boyack, K. W., Visualizing knowledge domains. *ARIST*. 2003, vol. 37, p. 179-255.

3） Blei, D. M., & Lafferty, J. D.. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*. 2006, p. 113-120.

4） Sugimito CR, et al. The shifting sands of disciplinary development. *JASIST*. 2011, vol. 62, no. 1, p. 185-204.

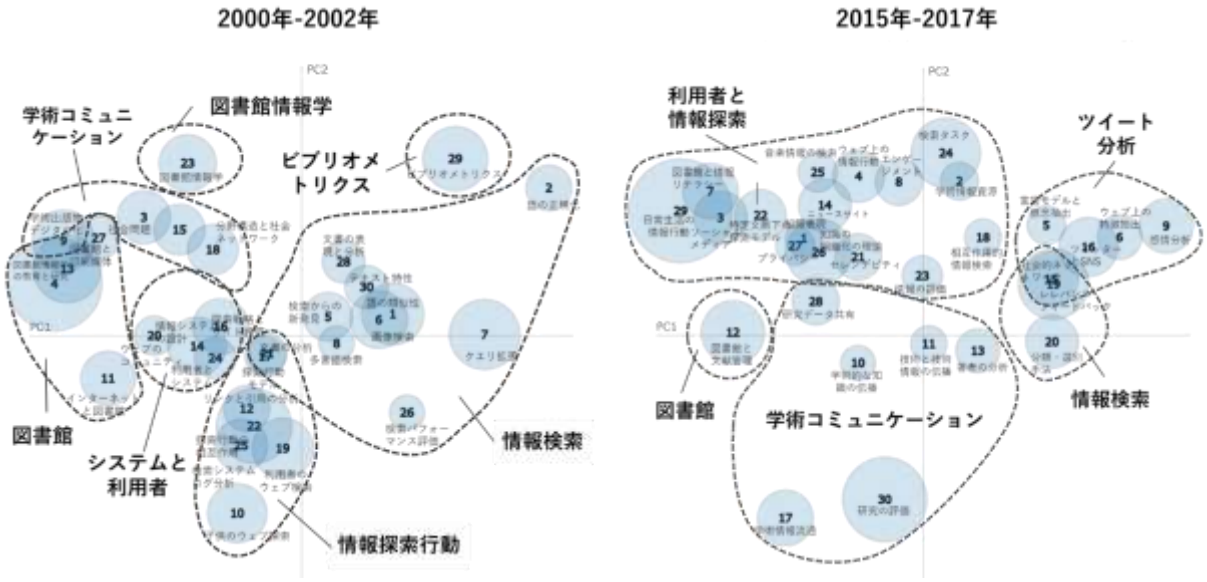
5） Yan, E. Research dynamics, impact, and dissemination: A topic‐level analysis. *JASIST*. 2015, vol. 66, no. 11, p. 2357-2372.

6） Figuerola, C. G., Marco, F. J. G., & Pinto, M.. Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*. 2017, vol. 112, no. 3, p, 1507-1535.

7） Kurata, K., et al. Analyzing Library and Information Science Full-Text Articles using a Topic Modeling Approach. *Proceedings of 2018 Annual Meeting of ASIS&T*. 2018, To Appear.

8） Řehůřek, R. genism. https://radimrehurek.com/gensim/, (accessed 2018-09-25)

9） bmabey. pyLDAvis. https://github.com/bmabey/pyLDAvis, (accessed 2018-09-25)



- **システムと利用者:** 14 インターフェースを考慮した情報システムの設計，16 ウェブ探索戦略とウェブサイトの評価，20 ウェブ上のコミュニティ，24 利用者と情報システム間の分析，29 ビブリオメトリクスとテキストへの統計的アプローチ
- **ビブリオメトリクス:** 29 ビブリオメトリクスとテキストへの統計的アプローチ
- **学術コミュニケーション:** 3 知識の伝達における社会問題，9 学術出版物のデジタル化における経済的側面，15 学術コミュニケーションへのインターネットの影響，18 分野構造と社会ネットワーク
- **情報検索:** 1 文献検索における語の類似性，2 レンマ化とステミング，5 情報検索から新たな発見を得る技術，6 画像検索，7 クエリ拡張とデータベース圧縮，8 多言語環境における情報検索，21 文書の分析，26 検索のパフォーマンス評価，28 情報検索のための文書表現と分析，30 モデル化を用いたテキストの特性
- **情報探索行動:** 10 児童，大学生のウェブ情報探索行動，12 ウェブのリンク分析と引用分析，17 学生の情報探索行動のモデル化，19 利用者ベースのウェブ検索，22 情報探索行動における相互作用，25 検索システムのログ分析
- **図書館:** 4 公共図書館の意義と役割，11 インターネットによる図書館サービスの利用，13 図書館情報学の教育と研究，27 図書館と印刷媒体
- **図書館情報学:** 23 図書館情報学における認識論
- **ツイート分析:** 5 言語モデルと概念抽出，6 ウェブ上の特徴抽出，9 多言語を含むテキストの感情分析，15 社会的ネットワークの応用，16 ツイッターと SNS
- **学術コミュニケーション:** 10 学術的な知識の伝播，11 技術と技術情報の伝播，13 著者の分析，17 学術情報流通，28 研究データ共有と知識共有，30 研究の評価
- **情報検索:** 19 レレバンスフィードバック，20 適切な分類・選別のためのアルゴリズム
- **図書館:** 12 図書館と文献管理
- **利用者と情報探索:** 1 メディアから得た知識表現，2 学術情報資源を使った分析，3 デジタルテキストとソーシャルメディアの情動，4 ウェブ上のコミュニティにおける情報行動，7 図書館と情報リテラシー，8 デジタル環境でのユーザーのエンゲージメント，14 ニュースサイトへのエンゲージメント，18 インタラクティブな情報検索，21 研究環境における情報探索：特にセレンディピティ，22 特定文脈における情報の探索と獲得のモデル化，23 情報の評価，24 検索タスク，25 音楽情報の検索，26 情報知識の組織化に対する新しい理論的アプローチ，27 ウェブコミュニケーションにおける相互作用：特にプライバシー，29 健康情報を中心とした日常生活の情報ニーズと情報行動

図 1　LDA 結果の 2 次元プロットとカテゴリ，トピックの一覧