様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １９ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)（一般）
研究期間：2015～2017
課題番号：１５Ｈ０２７８９

研究課題名（和文）実験情報の抽出・可視化・推薦のための電子図書館システムの研究

研究課題名（英文）A Study on Digital Library System for Experimental Information Extraction, Visualization and Recommendation

研究代表者

高須　淳宏（Takasu, Atsuhiro）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２１６６４８

交付決定額（研究期間全体）：（直接経費）　　12,200,000円

研究成果の概要（和文）：研究者は、研究計画の立案、関連研究調査、論文執筆など、研究の各段階で学術情報の調査分析が必要になる。電子図書館は、学術論文本文へのアクセスを容易にした点で大きな役割を果たしてきた。学術情報はこれまでキーワードを用いた全文検索を用いて提供されることが多かった。本研究では、特に学術論文に含まれる実験情報に焦点を宛て、その情報を学術論文の全文や引用文献から情報を抽出するための系列分析モデルを提案した。また、学術情報を効果的に提示するための情報推薦法を考案した。

研究成果の学術的意義や社会的意義
学術論文に含まれる実験情報は、研究成果を定量的に提示するとともに、科学的発見の根拠を示す情報として重要である。本研究の実験情報の抽出・提示技術は、研究者の研究動向の調査および分析の効率化に大きく寄与することが期待される。また、本研究の成果である利用者のニーズにあわせて情報を能動的に提供する情報推薦法は、学術情報のみならず、多種多様な情報推薦に利用可能な汎用性を備えた技術となっている。

研究成果の概要（英文）：Researchers need to survey research trend in the related research fields in various tasks, such as research planning, research trend analysis, and writing papers. Digital libraries have been playing an important role in providing research papers fulltext. Fulltext search is a main technology for retrieving research papers. This study focuses on experiment information included in papers and developed sequence analysis models for extracting experiment information. We also developed a recommender system for actively providing scholarly information.

研究分野：情報工学

キーワード：電子図書館　情報抽出　情報推薦

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

研究者は、研究計画の立案、関連研究調査、論文執筆など、研究の各段階で学術情報の調査分析が必要になる。電子図書館は、学術論文本文へのアクセスを容易にした点で大きな役割を果たしてきた。また、電子図書館に集積された大量の学術情報は、研究動向調査や研究者のアクティビティの分析など学術政策立案のための情報源としても貴重なものとなっている。たとえば、大規模かつ多様な情報のなかから、研究者に有効な情報を提供する技術として、情報推薦システムが注目を集めている。電子図書館分野の主要国際会議で毎年関連研究が採択されており、また、情報推薦システムの主要会議である。しかし、多くの学術情報推薦システムの研究は、論文や研究者といった比較的扱いやすい情報の推薦にとどまっている。

学術研究において実験は重要な役割を担っている。純粋に理論的な研究を除けば、研究によって得られた知見の正しさを示すため、また、技術的な研究では提案技術の優位性を示すために実験が行われる。研究者は、実験計画の立案から結果の分析、比較評価に至るプロセスにおいて、様々な情報を収集する必要がある。たとえば情報処理の分野では、研究者は、「どのような評価指標を用いるのが適切か？」、「ベンチマークデータにはどのようなものがあるか？」、「比較すべき先行研究や手法にどのようなものがあるか？」、「現在、どの程度の性能が達成されているのか？」といったことを知った上で実験を進めていく。研究者はこれらの実験情報を常にアップデートしておく必要があるが、巨大化・多様化する学術情報に対しては電子図書館の支援が必要となる。そのため、技術の優位性を示すためには、多くの論文を収集し、論文内に記載されている実験結果を抜き出し、比較分析する必要がある。現在の電子図書館にはこのような実験の各フェーズで研究者を支援する機能が欠けている。

２．研究の目的

本研究は、日々増大する学術情報を研究者が効果的かつ効率的に活用するための電子図書館システムの基盤技術を研究開発することを目的としている。電子図書館は研究者が学術論文を入手するのを容易にしてきたが、多くのシステムは学術論文や研究者情報へのアクセス支援に機能が限定されている。本研究では、これまで電子図書館ではあまり扱われてこなかった実験情報に着目し、(1)学術論文から実験に関する情報を高い精度で抽出する技術を研究開発し、(2)結果を提示するための可視化・推薦法を構築することを目的としている。

３．研究の方法

(1)学術論文からの情報抽出

学術論文から情報抽出を行うためには、論文テキスト分析モデルの構築が必要になる。モデルの学習にはある程度の規模のラベル付き訓練データを用意する必要がある。そのため、訓練データの作成コストを抑えつつ、高抽出精度を実現する抽出モデルを学習する技術が鍵となる。本研究では、以前から取り組んでいる学術論文から著者名やタイトルといった書誌要素を抽出する研究を出発点として、抽出モデルを効果的に学習する方法の研究を行う。

(2)実験情報の抽出

実験情報は図や表の形式で論文中にまとめられていることが多い。そこで、本研究では、まず論文中の図表の抽出を行う。ここでは、上記の情報抽出技術を展開することを試みる。さらに表に記載されている実験情報を理解するための表の parsing 技術の研究を行う。

(3)情報提示

実験結果を記載した表の理解の結果を用いて、利用者にわかりやすく提示するための表・グラフ変換法の研究を行う。また、利用者の特性および対象情報の特性に基づいた情報推薦法の研究を行う。情報推薦において、利用者の特性を得るためには多くのデータが必要になる。本研究では、利用可能な補助情報も併せて用いることで効果的に特性を獲得する手法の構築を試みる。

４．研究成果

(1)学術論文からの情報抽出

自然言語処理などの様々な分野で利用されている識別モデルの一つである Conditional Random Field (CRF)を利用して、論文中の参考文献文字列から書誌情報を自動で抽出する手法の研究を行ってきた。この研究では、参考文献書誌情報抽出精度を評価し、その抽出誤りの詳細な分析を示したが、高い抽出精度を得るには、参考文献文字列の書式が異なる学術雑誌ごとに、少なくとも数百件の参考文献文字列を学習データとして用意する必要があり、その生成コストは無視できない。そこで本研究では、能動サンプリングと擬似学習データ、他雑誌のデータにおいて学習した書誌情報抽出器の推定結果を利用して、少量学習データでの書誌情報抽出精度の向上を図る。

本研究では、少量学習データでなるべく高い精度を得るため、能動サンプリングを行う。さらに、擬似学習データを利用したり、他雑誌で学習した CRF の書誌情報抽出器の推定ラベルを

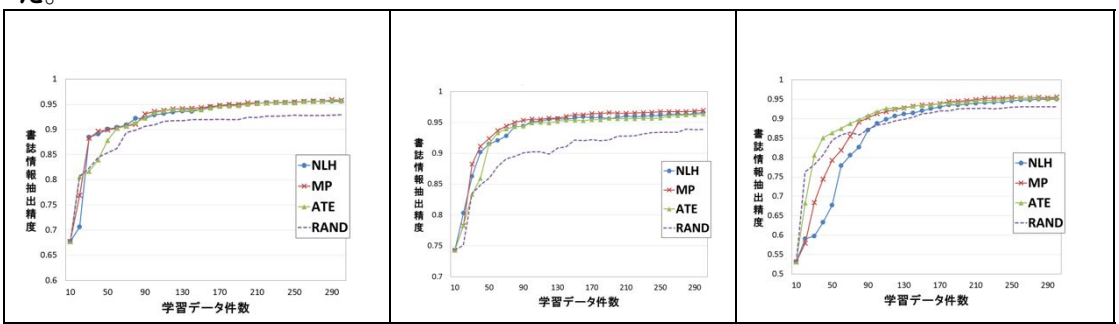
素性に加えたりすることで、更なる精度向上を図る。能動サンプリングは、ある時点の学習モデルで書誌情報抽出が困難な参考文献文字列を、優先的に選択して次回の学習データとし、逐次学習モデルを更新する。そのため、書誌情報抽出の困難さを表す尺度として、以下の三つの確信度を定義する。

一つ目の確信度（NLH）は CRF の出力する条件付き確率に基づいて定める。CRF は入力系列に対する条件付き確率が最大になるような出力ラベル系列を導出する。よって、入力系列を条件とする出力ラベル列の値が小さければ、その参考文献文字列に対するラベル付与は困難であるとみなすことができるので、この条件付き確率を確信度として利用する。ただし、この条件付き確率は入力系列の長さの影響を受けるため、入力系列の長さで正規化する。

二つ目の確信度（MP）は、参考文献文字列中の各トークンに付与されたラベルの周辺確率そのものを利用する。入力系列に対して、参考文献文字列中の i 番目の文字に対するラベル確率分布の最大値を P_i としたとき、参考文献文字列中の各トークンに対するこのラベル付与の確信度の中で最小のものを、その参考文献文字列の書誌情報抽出の確信度とする。

三つ目の確信度（ATE）は全ラベル候補の周辺確率のエントロピーに基づいて定める。この確信度では、各トークンに付与された周辺確率が最大のラベルだけでなく、その他のラベルも考慮できる。エントロピーの値が大きいほど、より多くの書誌要素ラベルに確率が分散しているため、ラベル付与が困難であると判断する。

能動サンプリングの有無による抽出精度の比較のため、ラベルを付与する参考文献文字列を無作為に選出した場合を RAND と記し、ベースラインとする。以下の図の縦軸は書誌情報抽出精度、横軸は学習データ件数、凡例はサンプリングに使用した確信度の種類、またはその方法を表す。図 1 に示すように、確信度の種類によって差はあるものの、概ね RAND より少ない学習データ件数で高い抽出精度が得られた。特に、図 1(b)の IEICE-E では、その効果が顕著だった。また、確信度の種類で比較すると、IEICE-J と IEICE-E では MP、IPSJ では ATE が効果的であった。



(a)IEICE-J　(b)IEICE-E　(c)IPSJ

図 1 能動サンプリングにおける学習データサイズと抽出精度

(2)実験情報の抽出

論文中の表の構造を解析し、それをグラフへと自動変換する手法を開発した。表の罫線の引き方やセルの構成など表の構造は著者によって異なる。特に、罫線は表の構造解析にあたって非常に重要な情報となるが、罫線を引くのか、それとも罫線でなくセルの配置によって行、列を表現するのかは論文の著者に任されている。そのため、我々は罫線を用いずセルの配置と隣接するトークンの重心などから表の構造を解析、グラフ化する方法を提案した。

提案手法は、表中に陽に使用されている罫線に加えて、陽には現れないがセルの境界を**補助罫線**と呼び、表中のオブジェクト間のスペースやまわりのオブジェクトとの位置揃えの情報を用いて補助罫線を推定する。次に、罫線と補助罫線に基づいて表の構造を解析する手法を提案する。まず論文 PDF ファイルを変換して得られる XML ファイル中のトークンをマージしていくことでセルを生成する。セパレータがトークン間にあるかどうかを基準にマージするため、エッジを利用することで補助罫線を推定する。セルの生成後、そのセルが何行、何列目であるかの範囲を決定する。最後に、セルの過分割やその範囲の誤りが存在するため、ルールを用いてセルを結合、拡張し、表の構造を決定する。図２は提案手法によって得られた表構造の例をしめしている。複数のセルにまたがる複合的なセルにも対応できることが示されている。

	Method Used								Problem¥$9
	(CFA_1)	(CFA_2)	MTFA	MINRES	WLS	PA	PC	MLE	(R/p)
	0	0	0	699.5	700.1	700	639.9	699.6	3/200
\	0.6889	0.6889	-	0.2898	0.2898	0.2898	0.2956	0.2898	3/200

図 2 表構造の抽出の例

(3)情報推薦のためのユーザモデリング

利用者が必要とする情報を的確に提供するためのユーザモデリングと情報推薦法の研究を行なった。ユーザモデリングでは、通常、ユーザからシステムに陽に与えられるフィードバックを用いるが、大規模なフィードバックを獲得することは難しい。そこで、ここでは、ユーザのアクセスログのように大規模に収集することが可能なデータも併せて利用する方法を考案した。

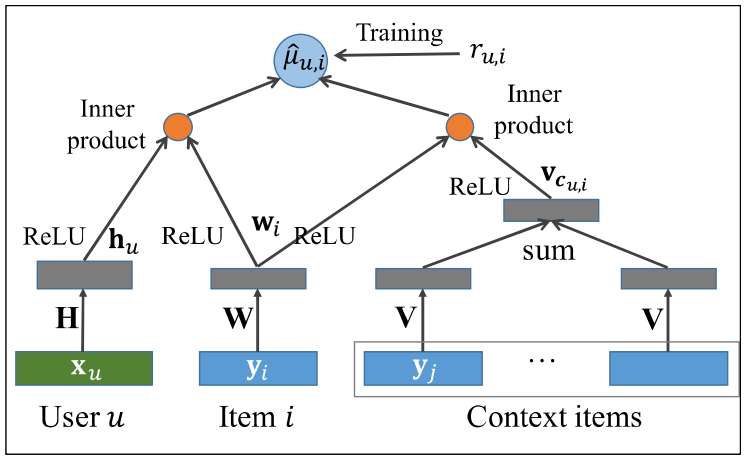


図 3　情報推薦のためのモデル

図３は本研究で提案したモデルを示している。左側は推薦システムで用いられるユーザからの陽なフィードバックに基づいてユーザモデルとオブジェクトモデルを獲得するためのencoder を表している。一方、右側はアクセスログなどの情報をコンテキストと考え、それらのコンテクストの埋め込み表現を統合するモデルとなっている。以下の表は、３つの評価用データセットを用いて提案手法を評価した結果をしめしている。表に示されるように提案手法は近年提案された３つの方法と比較して優れた推薦性能を達成した。

Methods	ML-10m		OnlineRetail		TasteProfile	
	Re@20	nDCG@20	Re@20	nDCG@20	Re@20	nDCG@20
SLIM	0.1342	0.1289	0.2085	0.1015	0.1513	0.1422
BPR	0.1314	0.1253	0.2137	0.0943	0.1598	0.1398
NeuCF	0.1388	0.1337	0.2199	0.0911	0.1609	0.1471
NPE (our)	**0.1497**	**0.1449**	**0.2296**	**0.1742**	**0.1788**	**0.1594**

表１　提案手法の推薦性能

５．主な発表論文等

〔雑誌論文〕（計 20 件）

1. Hung Nghiep Tran, Atsuhiro Takasu: Analyzing Knowledge Graph Embedding Methods from a Multi-Embedding Interaction Perspective, International Workshop on Data Science for Industry 4.0 in conjunction with EDBT 2019, p. 8, 2019.
2. Thai Binh Nguyen, Atsuhiro Takasu: Learning Representations from Product Titles for Modeling Shopping Transactions, AAAI-19 Workshop on Recommender Systems and Natural Language Processing, p.8, 2019.
3. Ryoya Yamada, Manabu Ohta, Atsuhiro Takasu: An Automatic Graph Generation Method for Scholarly Papers Based on Table Structure Analysis, International Conference on Management of Digital EcoSystems, pp. 132 - 140, 2018, 10.1145/3281375.3281389.
4. Thai Binh Nguyen, Atsuhiro Takasu: NPE: Neural Personalized Embedding for Collaborative Filtering, International Joint Conference on Artificial Intelligence, pp. 1583 - 1589, 2018, DOI: 10.24963/ijcai.2018/219.
5. Thai Binh Nguyen, Atsuhiro Takasu: A Probabilistic Model for the Cold-Start Problem in Rating Prediction Using Click Data, International Conference on Neural Information Processing, Vol. LNCS 10638, 2017, DOI: 10.1007/978-3-319-70139-4_20.
6. Thai Binh Nguyen, Atsuhiro Takasu: A Hierarchical Bayesian Factorization Model for Implicit and Explicit Feedback Data, 13th International Conference on Advanced Data Mining and Applications, pp. 104 - 118, 2017, DOI: 10.1007/978-3-319-69179-4_8.
7. MD Mostafizur Rahman, Atsuhiro Takasu: Entity Oriented Action Recommendations for Actionable Knowledge Graph Generation, EEE/WIC/ACM International Conference on Web Intelligence, pp. 686 - 693, 2017, DOI: 10.1145/3106426.3106546.
8. ThaiBinh Nguyen, Kenro Aihara, Atsuhiro Takasu: Collaborative Item Embedding Model for Implicit Feedback Data, International Conference on Web Engineering, Vol. LNCS 10360, pp. 336 - 348, 2017, DOI: 10.1007/978-3-319-60131-1_19.
9. Baptiste de La Robertie、 L. Ermakova, Yoann Pitarch, Atsuhiro Takasu, Olivier Teste: A Unified Approach for Learning Expertise and Authority in Digital Libraries, Springer Lecture Notes in Computer Science, pp. 354 - 368, 2017, DOI: 10.1007/978-3-319-55699-4_22.

10. Phannakan Tengkiattrakul, Saranya Maneeroj, Atsuhiro Takasu: Applying Ant-colony Concepts to Trust-based Recommender Systems, International Conference on Information Integration and Web- based Applications & Services, pp. 34 – 41, 2016, DOI: 10.1145/3011141.3011161.
11. Daiki Matsuoka, Manabu Ohta, Atsuhiro Takasu, Jun Adachi: Examination of Effective Features for CRF-Based Bibliography Extraction from Reference String, International Conference on Digital Information Management, pp. 259 – 264, 2016.
12. Junki Tanijiri, Manabu Ohta, Atsuhiro Takasu, Jun Adachi: Important Word Organization for Support of Browsing Scholarly Papers Using Author Keywords, ACM Symposium on Document Engineering, pp. 135 – 138, 2016, DOI: 10.1145/2960811.2967163.
13. Phanucheep Chotnithi, Atsuhiro Takasu, Frequent Multi-Byte Character Substring Extraction Using a Succinct Data Structure, ACM Symposium on Document Engineering, pp. 103 – 106, 2016. DOI: 10.1145/2960811.2967161.
14. Tomonari Masada, Atsuhiro Takasu: A Simple Stochastic Gradient Variational Bayes for the Correlated Topic Model, Springer Lecture Notes in Computer Science, Vol. 9932, pp. 424 – 428, 2016, DOI: 10.1007/978-3-319-45817-5_39.
15. Tomonari Masada, Atsuhiro Takasu: A Simple Stochastic Gradient Variational Bayes for the Latent Dirichlet Allocation, Springer Lecture Notes in Computer Science, Vol. 9789, pp. 232 – 245, 2016, DOI: 10.1007/978-3-319-42089-9_17.
16. 川上 尚慶、 太田 学、 高須 淳宏、 安達 淳: 少量学習データによる参考文献書誌情報抽出精度の向上, 情報処理学会論文誌データベース, Vol. 8, pp. 18 – 29, 2015.
17. Atsuhiro Takasu, Manabu Ohta: Utilization of Multiple Sequence Analyzers for Bibliographic Information Extraction, Lecture Notes in Computer Science, Vol. 9443, pp. 222 – 236, 2015, DOI: 10.1007/978-3-319-25530-9_15.
18. Tomonari Masada, Atsuhiro Takasu: Heuristic Pretraining for Topic Models、 Proc. of IEA/AIE 2015, Springer Lecture Notes in Computer Science、 Vol. 9101, pp. 123 – 134, 2015, DOI: 10.1007/978-3-319-19066-2_34.
19. Padipat Sitkrongwong, Saranya Maneeroj, Pannawit Samatthiyadikun, Atsuhiro Takasu: Bayesian Probabilistic Model for Context-Aware Recommendations, International Conference on Information Integration and Web- based Applications & Services, pp.166 – 175, 2015, DOI: 10.1145/2837185-2837223.
20. Tomoya Mori, Atsuhiro Takasu, Jesper Jansson, Jaewook Hwang, Takeyuki Tamura, Tatsuya Akutsu: Similar Subtree Search Using Extended Tree Inclusion, IEEE Transactions on Knowledge and Data Engineering, Vol. 27, pp. 3360 – 3373, 2015, DOI: 10.1109/TKDE.2015.2457922.

〔学会発表〕(計 5 件)
1. 田邊俊介、 太田 学、 高須 淳宏、 安達 淳：doc2vec による学術論文の被引用箇所推定の一手法、10th Forum on Data Engineering and Information Management、2018.
2. 山田凌也、太田 学、高須 淳宏: 学術論文の表の解析によるグラフの自動生成の一手法、10th Forum on Data Engineering and Information Management、 2018.
3. Atsuhiro Takasu: Similar Subtree Search Using Extended Tree Inclusion, IEEE International Conference on Data Engineering, 2016.
4. 高須 淳宏：書誌情報抽出および統合のためのテキストマイニング、人工知能学会合同研究会(招待講演)、 2015.
5. 松岡大樹、 太田 学、 高須 淳宏、 安達 淳： CRF による参考文献書誌情報抽出のための有効な素性の検討と拡充、 第 162 回データベースシステム研究会、 2015.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
特になし

６.研究組織

(1)研究分担者
研究分担者氏名：正田　備也
ローマ字氏名：Tomonari Masada
所属研究機関名：長崎大学
部局名：工学研究科
職名：准教授
研究者番号（8 桁）：60413928

(2)研究協力者
研究協力者氏名：太田　学
ローマ字氏名：Manabu Ohta

研究協力者氏名：Saranya Maneeroj
ローマ字氏名：Saranya Maneeroj