情報処理学会論文誌　データベース　Vol.14　No.4　49–60 (Oct. 2021)

クエリと文書のフィールドを考慮した被引用統計データの検索

中野 優[1,a)]　加藤 誠[2,3]

受付日 2021年3月9日, 採録日 2021年7月1日

概要：本論文では文書中の数値の真偽を検証するために，数値が参照する統計データを自動的に検索する手法を提案する．我々はこの統計データ検索問題をアドホック検索問題と見なし，数値と数値を含む文書の組をクエリ，統計データを検索対象の文書ととらえて検索モデルを適用するというアプローチをとる．さらに，検索対象文書である統計データが列名などのフィールドを持つ点に加えて，クエリに含まれる文書もタイトルなどの付随する情報からフィールドを持つ点に着目し，クエリと検索対象統計データの双方のフィールドを考慮した検索手法である BM25FF を提案する．提案手法の有効性を検証するために，政府統計を引用する Wikipedia 記事を利用してデータセットを作成した．このデータセットを用いて，BM25 などのクエリや統計データのフィールドを用いないベースライン手法と提案手法の比較を行った．その結果，提案手法はベースライン手法と比較して最大 3.8 倍の性能を発揮することが判明し，クエリと文書の両方のフィールドを利用することが本検索タスクにおいては重要であることが示された．これに加えて，エンティティに関連するクエリのフィールドが統計データの検索に重要である可能性が示された．

キーワード：データ検索，統計データ引用，複数フィールド検索，表検索，メタデータ

Leveraging Query and Document Fields for Cited Dataset Retrieval

Yu Nakano[1,a)]　Makoto P. Kato[2,3]

Received: March 9, 2021, Accepted: July 1, 2021

Abstract: To verify the numerical values in the text, we propose a method for automatically retrieving the cited statistical dataset. We formulate this dataset retrieval problem as an ad hoc retrieval problem. In other words, our approach is to first consider pairs of a numerical value and a document containing the numeric value as the queries and the statistical dataset as the documents to be retrieved, and then apply the retrieval model. We propose BM25FF, a retrieval method that takes into account the fields of both the query and the statistical dataset to be retrieved, based on the fact that the statistical dataset to be retrieved has fields such as column names, and the document included in the query also has fields such as titles. To evaluate the effectiveness of the proposed method, we created a dataset using Wikipedia articles that cite government statistical dataset. Using this dataset, we compare the proposed method with the baseline methods that do not use query or statistical dataset fields, such as BM25. Our evaluation shows that the proposed method performs up to 3.8 times better than the baseline method, indicating that the use of both query and document fields is important in this retrieval task. In addition, the query fields related to the entities were found to be important for the retrieval of statistical dataset.

Keywords: dataset retrieval, dataset citation, multi-field retrieval, table retrieval, metadata

1. はじめに

文章を書く際において，数値的な情報の根拠として統計データを引用することは多い．たとえば，犯罪件数の傾向に関する文章を書く際には「警視庁が発表した犯罪統計に

[1] 筑波大学大学院人間総合科学学術院
Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[3] JST さきがけ
JST PRESTO, Kawaguchi, Saitama 332–0012, Japan

[a)] s2030514@s.tsukuba.ac.jp

よると，2017 年の刑法犯の認知件数は 91 万件でした」のような形で統計データが引用されると考えられる．

このように，文章中において統計データから数値を引用することを，本論文では**統計データ引用**と呼ぶ．Redi らの研究 [18] において公開されているデータ[*1]によると，英語版 Wikipedia における引用のうち，統計やデータなどからの引用はデータセット全体の約 6.7%を占めており，統計データ引用は数多く存在すると考えられる．また，統計データ引用は論文においても年々増加傾向にあり [24]，議論の土台ともなるため，正しく引用されることが重要であると考えられる．

しかしながら，統計データは正しく引用されない場合が存在する．正しく引用されない場合としては次の 2 つの場合が考えられる．1 つ目はどの統計データを引用したのかが明示的に示されない場合である．たとえば「総務省の統計によると」など，出典が示されずデータの提供者のみが明記されることもあると考えられる．

このように統計データ引用は，形式がほぼ固定されている学術論文における引用とは異なり，引用された統計データ（**被引用統計データ**）が曖昧な形でしか記述されず，文章を読んだ人が根拠となる統計データを見つけられない可能性がある．そのため，既存の学術論文における引用の特定技術 [3], [16], [23] では被引用統計データを特定することは難しいと考えられる．

統計データが正しく引用されない 2 つ目の場合は，統計データから誤った数値が引用される場合である．我々の調査によると，Wikipedia において統計データが引用されている場合に，実際の統計データ内の数値と Wikipedia において記述されている数値が異なっている例が複数存在することが判明している．また，誤った数値を引用することは，フェイクニュースにもつながる恐れがあると考えられる．たとえば，2016 年のイギリスの EU 離脱に関して，EU 離脱派は「離脱すれば毎週 3.5 億ポンドを国内の医療制度に使える」という宣伝を行い，これは国民投票に影響を与えた．しかしながら，後の分析で実際に使える額はその半分かそれ以下であることが報告され，フェイクニュースであったことが判明した[*2]．このようなフェイクニュースの脅威を回避するためには，誤った数値が引用されている場合においても，被引用統計データが特定できることが重要となると考えられる．

本論文では，統計データから引用されている数値に誤りがないかの確認を容易にすることを目的として，文書中の数値から被引用統計データを自動的に特定する問題に取り組む．この問題は，数値がどの統計データを引用しているかを特定する問題と，数値が統計データ内のどの箇所を引用しているかを特定する問題の 2 つに分割することが可能であるが，本論文ではまず前者の問題に取り組むこととする．本論文ではこの前者の問題を**統計データ検索問題**と呼ぶこととし，アドホック検索の問題として定式化を行う．つまり，検証対象の数値と数値を含む文書の組をクエリと見なし，被引用統計データを検索対象文書と見なして，検索モデルを適用するというアプローチをとる．後者の問題は既存の表データに対する質問応答の研究 [12] を応用することが考えられるため，本論文では扱わずに詳細を 6 章にて議論する．

さらに本論文では，クエリに含まれる文書と検索対象文書の統計データの双方のフィールドを考慮した検索モデルを提案する．本論文が取り組む統計データ検索問題においては，検索対象文書である統計データがメタデータや表などの複数のフィールドを持つ．情報検索においては，このように検索対象文書が複数のフィールドを持つ場合の検索はこれまで研究されており，BM25 を拡張した BM25F などの検索モデルが研究されてきた [6], [14], [19], [27], [28]．

一方で，本論文の統計データ検索問題においては，クエリに含まれる文書もタイトルなどの付随する情報から複数のフィールドを持つという特徴がある．そこで本研究では，検索対象文書の統計データのフィールドだけではなくクエリのフィールドまで考慮した検索モデルとして，BM25F を拡張した **BM25FF** を提案する．これにより，クエリと検索対象文書の統計データをより詳細にマッチさせることが可能になると考えられる．

本論文では提案手法を評価するために，データセットの構築と実験を行った．データセットの構築においては，クエリとして Wikipedia の記事を用い，被引用統計データとして政府統計データのポータルサイトである e-Stat [*3]の統計データを用いた．作成手順としては，まず Wikipedia の記事から e-Stat の統計データへのリンクを抽出し，そこから Wikipedia 記事内の数値がリンク先の統計データ内の数値を実際に引用しているかを人手でアノテーションした．さらに，構築したデータセットに対してベースライン手法と提案手法を比較した．

実験の結果，提案手法はクエリと統計データのうちの片方のフィールドのみを考慮する手法やいずれのフィールドも考慮しない手法と比較して，約 2.2〜3.8 倍の性能を発揮することが判明した．加えてフィールドごとの分析を行った結果，エンティティに関連するクエリのフィールドが統計データの検索に重要である可能性が示された．

本論文の貢献は以下のとおりである．

(1) 文書中の数値から被引用統計データを特定する問題である統計データ検索問題を提案し，この問題をアドホック検索問題として定式化を行った．さらに統計

[*1] https://figshare.com/articles/Citation_Reason_Dataset/7756226（2021/05/19 閲覧）
[*2] https://ifs.org.uk/publications/10354（2021/05/19 閲覧）
[*3] https://www.e-stat.go.jp/（2021/05/19 閲覧）

データ検索問題に対して，クエリと検索対象文書の統計データの双方のフィールドを考慮した検索モデルを提案した．

(2) 統計データ検索問題に対する新たなベンチマーク用データセットを構築した．また，構築したデータセットについて，どのような文書においてどのような統計データが引用されているかについて分析を行った．

(3) 構築したデータセットを用いた実験により，提案手法の検証を行った．その結果，ベースライン手法と比較して提案手法は約 2.2〜3.8 倍の性能を発揮することが判明した．また，フィールドごとの分析により，エンティティに関連するクエリのフィールドが統計データの検索に重要である可能性が示された．

本論文の構成は次のとおりである．2 章では関連研究として引用を特定・推薦する研究，表や統計データを対象とした検索の研究，複数フィールドを持つアイテムに対する検索の研究について説明する．3 章では統計データ検索問題において Wikipedia と e-Stat を用いてデータセットを構築する方法と，構築したデータセットを分析した結果について説明する．4 章では統計データ検索問題の問題設定と，提案手法であるクエリと文書の双方のフィールドを考慮した検索モデル BM25FF について説明する．5 章では構築したデータセットを用いた評価実験を通して提案手法の有用性を確認し，6 章では本論文の結論とともに今後の課題について説明する．

2. 関連研究

本節では統計データ検索に関連する研究として，2.1 節で類似する問題設定である引用の特定と推薦に関する研究を紹介した後，2.2 節と 2.3 節で関連する検索技術としてそれぞれ表検索に関する研究と複数フィールド検索に関する研究を紹介する．

2.1 引用の特定・推薦

文書から引用されたアイテム（被引用アイテム）を特定する研究としては，学術論文における被引用アイテムの特定の研究があげられる．特定する被引用アイテムの種類としては，被引用論文を特定する研究 [16], [23] や，被引用データセットを特定する研究 [3] が存在する．論文における引用においては，引用された文献が必ず明示され，かつ形式がほぼ固定である．一方で，本論文の統計データの引用においては「総務省の統計によると」など，被引用統計データが曖昧な形でしか提示されない場合もあるという点において本論文とは異なる．

本論文のように，明示的に文献が示されず周辺の情報などから引用を特定するという問題設定は，引用を推薦する研究と類似している．引用を推薦する研究とは，文脈となる文書が与えられた際に，その文脈の根拠を補強するために引用すべきアイテムを推薦する研究である．このような研究としては，学術論文を推薦する研究 [11] やニュースを推薦する研究 [8], [17] が行われている．既存の被引用アイテムの推薦を行う研究では，推薦の対象となるアイテムがテキストである場合が多いことに対して，本論文は統計データを対象とする点においてこれらの研究とは異なる．

2.2 表検索・統計データ検索

統計データに類似するアイテムとして，表を対象とした検索の研究を説明する．アドホック表検索タスクはキーワードをクエリとして，クエリに適合する Web 上の表を検索するタスクである [1], [5], [21], [22], [27]．Zhang と Balog はアドホック表検索タスクのデータセットである WikiTables データセットを提案し，さらにランキング学習を用いて表を検索する手法や単語埋め込みとエンティティ埋め込みを用いた semantic matching によって表を検索する手法を提案している [27]．アドホック表検索タスクにおいては Zhang と Balog が提案した WikiTables データセットを対象に研究が行われており，パッセージ検索と多様体学習を用いる手法 [21]，行列分解を用いる手法 [1]，マルチモーダル深層学習を用いる手法 [22]，BERT を用いる手法 [5] など様々な手法がこれまでに研究されてきた．本論文はクエリが（短い）キーワードではなく（長い）文書を含むという点と，3.4.1 項で後述するとおり，既存の表検索データセットに含まれる表は文字列が多い一方で本論文が扱う統計データは数値が多いという点において，これらの研究とは異なる．

また，キーワードから統計データを対象として検索を行う研究も存在する．Chen ら [4] は，統計データから抽出した列名を検索に用いることで，統計データの検索性能を向上させる手法を提案した．さらに，data.gov の統計データを対象として 6 つの検索タスクを設定して統計データ検索用のデータセットを構築し，ベースライン手法と比較して提案手法が良い検索結果を提示できることを示した．表検索の場合と同様に，本論文はクエリがキーワードではなく文書を含むという点において上記の研究とは異なる．

2.3 複数フィールド検索

情報検索において，検索対象文書が複数のフィールドを持つことは多い．そのようなアイテムに対する検索の研究としては，Web 検索 [19], [26]，エンティティ検索 [28]，表検索 [5], [21], [27]，XML 検索 [14]，商品検索 [6] などが存在する．複数フィールドを持つ文書に対する検索においては，フィールドごとに出現しやすい単語が異なるなど，検索アルゴリズムにおいて用いられる情報の傾向が異なる．そのためこれまでの研究においては，検索対象文書が持つフィールドごとにクエリとのスコアを計算した後に，スコアを統合するという検索モデルが研究されてきた [19], [27].

近年では深層学習を用いた複数フィールド検索モデルも研究されており，フィールドごとのスコアの計算からスコアを統合する部分まで end-to-end で行うモデルが提案されている [2], [26]．本論文は検索対象である統計データがメタデータや表など複数のフィールドを持つという点においてこれらの研究と類似する一方で，クエリもフィールドを持つという点においてこれらの研究とは異なる．

クエリがフィールドを持つ場合の検索の研究としては，上記のような複数フィールドを持つアイテムに対して，アイテム自体をクエリとして検索を行う研究が存在する．たとえばキーワードクエリに加えて例となるエンティティの集合をクエリとして与え，類似するエンティティを検索するタスクである類似エンティティ検索 [7] や，表をクエリとして与えて関連する表を検索するタスクである類似表検索 [20] などが研究されている．上記であげた研究はクエリのフィールドと文書のフィールドが同じであることが想定されている．そのため，クエリが持つ各フィールドについて，対応する検索対象アイテムのフィールドとマッチングさせれば良い．つまり，クエリと文書のフィールドに対して，どのフィールドどうしをマッチングさせるべきかが明確である．一方で，本論文はクエリが持つフィールドと検索対象のアイテムが持つフィールドが異なる．そのため，クエリと文書のどのフィールドどうしをマッチングさせるべきかが不明であるという点において，上記の研究とは異なる．

3. データセット

本章ではまず統計データ検索のベンチマーク用データセットを構築する方法について説明し，その後構築したデータセットについて分析を行う．

3.1 使用する統計データ

本論文では利用する統計データとして，e-Stat の統計データを利用する．ここでは NTCIR-15 Data Search Task [13] の Japanese Subtask のために提供された e-Stat のデータセットを利用する．この統計データの概要を**表 1** に示す．このデータセットには 1,338,402 個の統計データのファイルが含まれており，各統計データのファイルには政府統計名（例：国勢調査）やデータ提供者（例：総務省）などのメタデータが付随している．本論文では，Excel 形式と CSV 形式の統計データのみを使用し，PDF 形式の統計データは扱わないこととした．

表 1　統計データセットの形式ごとの統計情報
Table 1　Statistics of dataset collections.

形式	拡張子	個数	割合
Excel	.xlsx, .xls	721,230	53.9%
CSV	.csv	568,042	42.4%
PDF	.pdf	49,124	3.7%
その他	.xlsm	6	< 0.1%
計		1,338,402	-
Excel+CSV		1,289,272	96.3%

3.2 数値情報を含む文書からのデータセット構築

数値情報を含む文書から統計データ検索のデータセットを構築した方法について説明する．

まず，数値情報を含む文書からその文書がどの統計データを引用しているかを自動的に特定した．本論文では数値情報を含む文書として Wikipedia を使用することとし，2020 年 6 月 1 日時点の Wikipedia 日本語版のダンプデータを取得した．このデータに含まれる Wikipedia 記事の各セクションから e-Stat の統計データへのリンクを抽出し，さらに特定の 1 つの統計データへのリンクのみを抽出した．これにより，記事のセクションの単位で被引用統計データの候補となる集合を特定することができた．

次に，記事の各セクションに対して，セクション中の各数値がどの統計データを引用しているかについて，候補となる統計データの中から人手で特定した．本データセットにおける「数値が統計データから引用されている」の定義としては，対象となる数値が統計データ中の 1 つのセルを引用する場合のみ，その数値が統計データを引用していると見なすこととした．この定義にしたがってアノテーションを行った結果，109 個の Wikipedia 記事と 40 個の統計データから計 443 個の (数値, 統計データ) の組を得ることができた．この組において数値と統計データは，それぞれアドホック検索におけるクエリと適合文書を表している．つまり，クエリは文章中の検証対象の数値であり，付随する情報としてページタイトルや数値周辺のテキスト（コンテキスト）などを持つ．さらに各クエリは，ただ 1 つの適合文書である統計データを持つ．

3.3 クエリ・文書のフィールド設定

1 章で述べたとおり，本論文のアドホック検索の問題設定においては，検索対象文書である統計データだけではなく，クエリに含まれる文書もフィールドを持つ．本節ではクエリと文書（統計データ）が持つフィールドについてそれぞれ説明する．

3.3.1 クエリのフィールド設定

本論文で用いるクエリのフィールドを**図 1** に示す（図の例は宇部市の Wikipedia 記事[*4]より）．前節でも述べたとおり，本論文におけるクエリは統計データを引用する検証対象の数値と，その数値を含む文書の組である．このうち，数値を含む文書（ここでは Wikipedia の記事）が持つページタイトルやカテゴリなどを考慮することにより，クエリがフィールドを持つと見なせる．各フィールドの詳細

[*4] https://ja.wikipedia.org/?curid=18837（2020/09/30 閲覧）

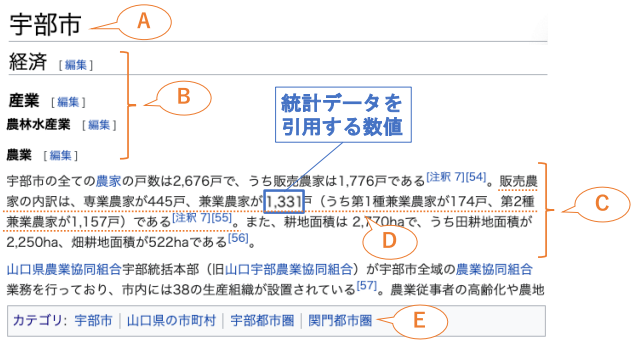


図 1　クエリのフィールド
Fig. 1　Fields of queries.

(a) メタデータ　　(b) 表

図 2　文書（統計データ）のフィールド
Fig. 2　Fields of documents (statistical dataset). The left and right figures represent metadata and a table, respectively.

は以下のとおりである．

A. ページタイトル　クエリに含まれる数値（以下，本節内においては単に数値と呼ぶ）を含む Wikipedia 記事のページのタイトルである．

B. セクションタイトル　数値を含むセクションのタイトルと，その祖先となるすべてのセクションのタイトルからなる．

C. パラグラフ　数値を含むパラグラフである．本論文では，改行が 2 連続する部分をパラグラフの分かれ目と見なした．ただし，パラグラフが数値の前後 200 文字を超える場合は，そこで打ち切った．

D. コンテキスト　数値の周辺のテキストである．本論文では，数値の前後 50 文字のテキストをコンテキストと設定した．

E. カテゴリ　数値を含むページのカテゴリからなる．

3.3.2　文書（統計データ）のフィールド設定

本論文で用いる文書（統計データ）のフィールドを**図 2** に示す（図の例は農林業センサス[*5]より）．統計データはメタデータと表を持ち，それぞれがフィールドを持つ．各フィールドの詳細は以下のとおりである．

a. タイトル　メタデータとして付与されたタイトルである．

b. 説明　メタデータとして付与された説明である．

c.（その他の）メタデータ　メタデータとして付与された情報のうち，タイトルと説明以外のメタデータからなる．

d. 列ヘッダかつ行ヘッダ　表のうち，列ヘッダ（列名）の行と，行ヘッダ（行名）の列の両方に該当するセルの文字列からなる．本実験では，表の先頭の 2 割の行のうち，数値のセルが 1 割未満の行を列ヘッダの行と見なす（行ヘッダも同様）．

e. 列ヘッダ，f. 行ヘッダ　それぞれ表の列ヘッダ（列名），行ヘッダ（行名）である．ただし，列ヘッダかつ行ヘッダの部分を除く．

g. データ　表のうち，列ヘッダかつ行ヘッダ，列ヘッダ，行ヘッダに該当しないセルの文字列からなる．

3.4　データセット分析

本節では 3.1 節で説明した統計データセットに関する分析と，3.2 節で構築した統計データ検索のデータセットに関する分析を行う．

3.4.1　統計データセットの分析

本項では，2.2 節であげた表検索の関連研究との違いを明確ことを目的として，3.1 節で説明した e-Stat の統計データのデータセットと，表検索に関する既存のデータセットである WikiTables データセット [27][*6] を比較する．WikiTables データセットは Wikipedia に含まれる表データを収集したデータセットである．また，このデータセットは 2.2 節で説明したアドホック表検索タスクにおいて検索対象となっている Web 上の表についてのデータセットの 1 つである．これら 2 つのデータセットに対して，列数，行数，数値のセルの割合のそれぞれについて，平均値と中央値を計算した結果を**表 2** に示す[*7,*8]．表 2 から分かることとしては，統計データは Web 上の表データと比較して，列数や行数が多く，かつ文字列のセルより数値のセルが多いということがいえる．

3.4.2　統計データ検索のデータセット分析

本項では，3.2 節で構築した統計データ検索用データセットを分析した結果について述べる．データセット中の Wikipedia 記事 109 件について，記事に付与されているカテゴリの出現数上位 5 件を**表 3** に示す．また，データセット中 e-Stat の統計データ 40 件について，統計データに付与されている政府統計名の出現数上位 5 件を**表 4** に示す．表 3 の第 1 位から第 4 位までのカテゴリを見ると，統計

*5　https://www.e-stat.go.jp/stat-search/files?stat_infid=000001186106（2020/09/30 閲覧）

*6　https://github.com/iai-group/www2018-table（2021/05/19 閲覧）

*7　数値のセルの割合について説明する．まず表の全セルの個数を表全体から空のセルを取り除いたときのセルの個数と定義し，数値のセルを整数もしくは実数の値のみが入っているセル，それ以外のセルを文字列のセルと定義する．このとき数値のセルの割合を，(数値のセルの個数)/(全セルの個数) で定義する．

*8　統計データセットについては，Excel 形式と CSV 形式のデータセットから合計して 1 割になるように統計データをサンプリングし，サンプリングしたデータセットに対して計算した．

表 2　e-Stat（統計データ）と WikiTables [27]（既存の表のデータセット）の比較．Ave. と Med. はそれぞれ平均値と中央値を表す

Table 2　Comparison of e-Stat (statistical dataset) and WikiTables (existing Webtable dataset). Ave. and Med. denote the mean and median, respectively.

データセット	列数		行数		数値セルの割合	
	Ave.	Med.	Ave.	Med.	Ave.	Med.
e-Stat	53.4	28	1,403.7	129	76.1%	81.9%
WikiTables [27]	5.0	4	12.1	6	26.0%	16.7%

表 3　データセット中の Wikipedia 記事のカテゴリの上位 5 件

Table 3　Top 5 categories of Wikipedia articles in the dataset.

カテゴリ名	個数
日本の町・字のスタブ項目	64
鹿児島県関連のスタブ項目	54
日置市の大字	36
指宿市の町・字	17
ISBN マジックリンクを使用しているページ	11

表 4　データセット中の e-Stat の統計データの政府統計名一覧

Table 4　List of government statistics names for e-Stat statistical dataset in the dataset.

政府統計名	個数
国勢調査	10
漁業センサス	4
経済センサス - 基礎調査，福祉行政報告例，作物統計調査，農林業センサス，矯正統計調査，食料需給表，特産果樹生産動態等調査，人口動態調査，被保護者調査	2
社会・人口統計体系，在留外国人統計（旧登録外国人統計），特定作物統計調査，労使関係総合調査（労働組合基礎調査），登記統計，全国消費実態調査，木材統計調査，就業構造基本調査	1

データが引用されている Wikipedia 記事としては市町村に関する記事が多いことが分かる．さらに，これらの記事について実際に統計データが引用されている箇所を確認すると，国勢調査から人口などを引用している場合が多いことが判明した．この傾向は表 4 の最上位に国勢調査が位置しているという事実とも一致する．

また，データセット中に含まれる Wikipedia 記事が最後に更新された年と e-Stat の統計データが公開された年の頻度を**表 5** に示す．Wikipedia の記事に関しては最終更新年が比較的最近であることに対して，e-Stat の統計データの公開年は 2015 年以前のものも多く存在している．このことから，Wikipedia の記事の内容は頻繁に更新されるものの，統計データに関する記述については古い統計を引用したままであり，あまり頻繁には更新されないと推測される．公開年度の古い被引用統計データに対しては，より新しい統計データを推薦することが適切であるが，これに関しては政府統計名などのメタデータから容易に公開年度の新しい統計データを特定することが可能であると考えられる．

表 5　データセット中に含まれる Wikipedia 記事が最後に更新された年と e-Stat の統計データが公開された年の頻度

Table 5　Frequency of the last updated year of the Wikipedia articles (left) and the published year of the e-Stat statistical dataset (right), in out dataset.

(a) Wikipedia の最終更新年

最終更新年	個数
2017 以前	24
2017	4
2018	11
2019	30
2020	40

(b) 統計データの公開年

公開年	個数
2015 以前	17
2015	5
2016	5
2017	7
2018	6

4. 統計データ検索

本章ではまず 4.1 節で本論文が扱う問題設定について説明し，4.2 節で基礎となる検索モデルについて説明した後，4.3 節で提案手法である検索モデルについて説明する．

4.1 問題設定

統計データ検索問題のアドホック検索問題としての定式化を説明し，必要となる記号を導入する．本論文で使用する記号を**表 6** に示す．

入力として与えられたクエリ q と検索対象文書の集合 D に対して，本検索タスクは D に含まれる文書をランク付けする問題であり，出力としてランキング $(d_1, d_2, \ldots, d_k)$ を返す．ただし，各 i について $d_i \in D$ である．

本問題におけるクエリ $q \in Q$ は検証対象の数値と数値を含む文書の組である．ただし，Q は任意のクエリの集合である．クエリ q に含まれる文書はタイトルやパラグラフなど複数のフィールドを持ち，このフィールドの集合を F_Q と表す．本論文における F_Q は 3.3.1 項で説明したタイトルやパラグラフなどの 5 つのフィールドである $(|F_Q| = 5)$．

本論文における検索対象文書 $d \in D$ は統計データである．統計データはメタデータや表など複数のフィールドを持ち，このフィールドの集合を F_D と表す．本論文における F_D は 3.3.2 項で説明したタイトルや列名などの 7 つのフィールドである $(|F_D| = 7)$．

本論文ではこのランキング問題を，クエリ $q \in Q$ と検索対象文書の統計データ $d \in D$ を入力として，スコアを出力

表 6　本論文で用いる記号
Table 6　Notations in this paper.

記号	定義
Q	任意のクエリの集合
D	検索対象文書の集合
$q \in Q$	クエリ
$d \in D$	検索対象文書
F_Q	クエリのフィールドの集合
F_D	検索対象文書のフィールドの集合
$f \in F_Q$	クエリのフィールド
$f' \in F_D$	検索対象文書のフィールド
$s(q,d)$	クエリ q に対する文書 d のスコア
$w \in q$	クエリ q に含まれる単語
$\mathrm{tf}(w,d)$	検索対象文書 d における単語 w の出現頻度
$\mathrm{idf}(w)$	検索対象文書集合 D における単語 w の逆文書頻度
$\mathrm{dl}(d)$	検索対象文書 d の長さ（単語数）
$\mathrm{ql}(q)$	クエリ q の長さ（単語数）
avgdl	全文書の平均文書長（平均単語数）
k_1, b	BM25 のパラメータ
α	クエリフィールドの重み
β	文書フィールドの重み

するスコア関数 $s : Q \times D \to \mathbb{R}$ を設計する問題と見なす．つまり，スコア関数 s の出力するスコアの降順に結果を並べることで最終的なランキングを得る．

4.2　基礎となる検索モデル

本節では次節で説明する提案手法の基礎となる検索モデルについて説明する．

4.2.1　単一フィールド化モデル

まず最も基本的な検索モデルとして単一フィールド化モデルを説明する．このモデルは，クエリのフィールドや，検索対象文書の統計データのフィールドを無視してそれぞれ 1 つのテキストとして表現することにより，通常のアドホック文書検索と同様に検索を行うモデルである．たとえば，検索対象となる統計データはメタデータや表などの構造を持っているが，単一フィールド化モデルではその統計データに含まれるテキストのみを抽出し，1 つの文書とする．これにより通常の文書検索のスコア関数が適用可能となる．本論文では単一フィールド化モデルのスコア関数として BM25 を用いることとする．BM25 のスコア関数 s は以下で表される．

$$s(q,d) = \sum_{w \in q} \mathrm{idf}(w) \frac{\mathrm{tf}(w,d)}{k_1 \left((1-b) + b \frac{\mathrm{dl}(d)}{\mathrm{avgdl}} \right) + \mathrm{tf}(w,d)} \tag{1}$$

ただし，w はクエリ q に含まれる単語，$\mathrm{tf}(w,d)$ は文書 d における単語 w の出現頻度，$\mathrm{idf}(w)$ は単語 w の逆文書頻度を表す．また，k_1, b はパラメータであり，avgdl は全文書の平均文書長（平均単語数），$\mathrm{dl}(d)$ は文書 d の文書長（単語数）を表す．

4.2.2　複数フィールド化モデル

次に複数フィールド化モデルについて説明する．2.3 節で説明したとおり，情報検索において検索対象文書が複数のフィールドを持つことは多い．また，2.2 節で説明したアドホック表検索においても，表をタイトルや列名など複数フィールドを持つ文書と見なして検索することが一般的である [5], [21], [27]．

複数フィールドを持つ文書の検索においては，文書フィールドごとに長さや出現しやすい単語などが異なるなど，検索アルゴリズムにおいて用いられる情報の傾向が異なるため，文書フィールドごとにスコアを計算し，それを統合するというアプローチが用いられることが多い．たとえば BM25F [19] は以下のスコア関数 s によりスコア付けを行う．

$$s(q,d) = \sum_{w \in q} \mathrm{idf}(w) \frac{\widetilde{\mathrm{tf}}(w,d)}{k_1 \left((1-b) + b \frac{\mathrm{dl}(d)}{\mathrm{avgdl}} \right) + \widetilde{\mathrm{tf}}(w,d)} \tag{2}$$

$$\widetilde{\mathrm{tf}}(w,d) = \sum_{f' \in F_D} \beta_{f'} \cdot \mathrm{tf}(w, d_{f'}) \tag{3}$$

ただし，$\mathrm{tf}(w, d_{f'})$ は文書 d の文書フィールド f' における単語 w の出現頻度を表し，$\beta_{f'}$ は文書フィールド f' ごとのパラメータである．BM25 の式 (1) と BM25F の式 (2) は単語頻度の計算方法のみが異なっている．BM25F の単語頻度は式 (3) で計算され，この式は文書フィールドごとの単語頻度の重み付き和を計算している．よって，BM25F は式 (3) を擬似的な単語頻度と見なして BM25 のスコアを計算する手法であり，式 (3) の $\beta_{f'}$ でどの文書フィールド f' を重要視するかが設定可能となっているといえる．

4.3　BM25FF

本節では新たな検索モデルである **BM25FF** について説明する．以下では，まず BM25FF のスコア関数を説明したのち，BM25FF が持つパラメータのチューニング方法について説明する．

4.3.1　スコア関数

提案手法の BM25FF のスコア関数について説明する．既存のフィールドを考慮した検索モデルは，検索対象文書のフィールドのみを考慮したモデルであった．しかしながら本論文の問題設定においては，クエリに含まれる文書も複数のフィールドを持ちうる．そこで本論文では，文書フィールドのみを考慮する BM25F をクエリのフィールドも考慮できるように拡張した手法である **BM25FF** を提案する．

BM25FF のように，クエリと文書の両方がフィールドを持つ問題設定においては，クエリのフィールドと文書のフィールドの交互作用が重要である．そこで，BM25F の

重み付き単語頻度を表す式 (3) における，文書フィールドの重みのパラメータ β を，クエリフィールドごとに設定する．具体的には BM25FF のスコア関数 s を以下の式で定義する．

$$s(q,d) = \sum_{f \in F_Q} \alpha_f \cdot \frac{s_f(q_f, d)}{\mathrm{ql}(q_f)} \tag{4}$$

$$s_f(q_f,d) = \sum_{w \in q_f} \mathrm{idf}(w) \frac{\widetilde{\mathrm{tf}}(w,d)}{k_{1f}\left((1-b_f)+b_f\frac{\mathrm{dl}(d)}{\mathrm{avgdl}}\right)+\widetilde{\mathrm{tf}}(w,d)} \tag{5}$$

$$\widetilde{\mathrm{tf}}(w,d) = \sum_{f' \in F_D} \beta_{f,f'} \cdot \mathrm{tf}(w, d_{f'}) \tag{6}$$

ただし，q_f はクエリ中のフィールド $f \in F_Q$ に含まれる単語の集合であり，α_f はクエリのフィールド f に関するパラメータである．また，$\mathrm{ql}(q)$ はクエリ q のクエリ長（単語数）を表す．BM25FF のスコア関数である式 (4) はクエリフィールドごとに BM25F のスコアを計算し，その重み付け和をとる式となっている．また，式 (5) と式 (6) は BM25F においてはそれぞれ式 (2) と式 (3) に対応している．

4.3.2　パラメータチューニング

提案手法のパラメータチューニングについて説明する．提案手法は式 (4) における α_f（計 $|F_Q|$ 個），式 (5) における k_{1f}, b_f（計 $2|F_Q|$ 個），式 (6) における $\beta_{f,f'}$（計 $|F_Q||F_D|$ 個）のように，多数のパラメータを持つ．これらは合計すると $|F_Q|(|F_D|+3)$ となり，グリッドサーチのようなナイーブなチューニング方法では時間がかかりすぎると考えられる．そこで本研究では提案手法のパラメータを最適化するために，座標上昇法（Coordinate Ascent，CA）[15] を用いる．CA は最適化手法の 1 つであり，情報検索においては線形モデルを用いたランキング学習において，パラメータを最適化するために用いられる手法である．一方で BM25 や BM25F のような非線形なモデルにおいても CA によるパラメータのチューニングは良い性能を示すことが経験的に知られており [10]，BM25FF のパラメータチューニングにおいても用いることとした．

5.　実験

本章では前章で構築したデータセットに対して，提案手法の性能をベースライン手法との比較により評価する．まずは検証すべき研究課題を列挙した後，比較手法と実験設定について述べ，最後に実験結果について述べる．

5.1　研究課題

RQ1　クエリのフィールドと統計データのフィールドの両方を用いることで統計データ検索の性能は改善するか？

RQ2　クエリと統計データのどのフィールドが統計データ検索の性能に影響を与えるか？

表 7　本実験で比較する手法．チューニング方法の GS はグリッドサーチ，CA は座標上昇法を表す

Table 7　The methods compared in this experiment. The tuning method GS and CA means grid search and coordinate ascent, respectively.

	フィールド		パラメータ数	チューニング
	クエリ	統計データ		
BM25	-	-	2	GS
BM25F	-	✔	9	CA
QF-BM25	✔	-	15	CA
BM25FF	✔	✔	50	CA

5.2　比較手法

本実験で比較する手法を**表 7** に示す．各手法の詳細は以下のとおりである．

BM25　このベースライン手法はクエリのフィールドと統計データのフィールドをそれぞれ 1 つのフィールドに集約し，検索を行う．BM25 は k_1, b の 2 つのパラメータを持つ．BM25 の実装としては，Anserini [25] [*9] を用いる．

BM25F [19]　このベースライン手法はクエリのフィールドを考慮せずに，検索対象の統計データのフィールドを考慮して検索を行う．スコア関数は式 (2) で表され，$|F_D|+2=9$ 個のパラメータを持つ．

QF-BM25　このベースライン手法は，検索対象の統計データのフィールドを考慮せずに，クエリのフィールドのみを考慮する検索を行う．これは BM25FF において $|F_Q|=1$ のとき，つまり検索対象の統計データのフィールドが 1 つの場合として表現され，本実験ではこの手法を Query Fielded BM25（QF-BM25）と呼ぶ．QF-BM25 は $|F_Q|(2+1)=15$ 個のパラメータを持つ．

BM25FF（提案手法）　この手法は 4.3 節で説明したとおり，クエリのフィールドと統計データのフィールドの両方を考慮して検索を行う．BM25FF のスコア関数は式 (4) で表され，$|F_Q|(|F_D|+3)=50$ 個のパラメータを持つ．

5.3　実験設定

本節では実験設定として，評価指標と評価手順について説明したのち，パラメータチューニングと推論の方法について説明する．

5.3.1　評価指標・評価手順

評価指標としては，平均逆数順位（Mean Reciprocal Rank，MRR）と Hit@k（$k=10,20,100$）を用いることとする．本データセットは適合文書となる統計データが 1 つ

[*9] http://anserini.io（2021/05/19 閲覧）

表 8　実験結果．v.s. BM25 W/T/L は BM25 と比較して MRR が上回った/同じ/下回ったクエリの個数を表す．括弧内の数値は BM25 からの相対的な改善率を表す．記号 † は Tukey HSD 検定（$p < 0.01$）において，その手法以外のすべての手法と比較して統計的に有意な差が認められたことを表す

Table 8　Experimental results. v.s. BM25 W/T/L (Win/Tie/Loss) are the number of queries improved, unchanged, or hurt, compared to BM25 on MRR. Relative improvements over BM25 are shown in parentheses. The symbol † denotes that a Tukey HSD test ($p < 0.01$) shows the differences of the method and all other methods are statistically significant.

	MRR		v.s. BM25 W/T/L	Hit@10		Hit@20		Hit@100	
BM25	0.094		-/-/-	0.129		0.266		0.422	
BM25F	0.139	(+48.0%)	219/111/113	0.318	(+147.4%)	0.345	(+29.7%)	0.442	(+4.8%)
QF-BM25	0.080	(−14.8%)	243/ 65/135	0.237	(+84.0%)	0.266	(−0.1%)	0.637	(+50.9%)
BM25FF	**0.305**†	(+225.0%)	311/ 41/ 91	**0.395**†	(+206.9%)	**0.444**†	(+66.8%)	**0.731**†	(+73.3%)

しか存在しないため，既存のアドホック表検索やアドホックデータセット検索で用いられる NDCG や MAP などの評価指標ではなく，これらの指標を用いることとした．

また，本実験では前節で説明した比較手法に対して，5-fold 交差検証を用いて性能を検証する．つまり，データセットのクエリを 5 つの fold に分割して，そのうち 4 つの fold のクエリを用いて各検索手法のパラメータをチューニングし，1 つの fold のクエリを用いて評価を行う，という手順を 5 回繰り返す．

5.3.2　パラメータチューニング方法（訓練方法）

各検索手法の持つパラメータのチューニング方法について説明する．パラメータ数の少ない BM25 はコレクション全体に対するグリッドサーチ（GS）で最適化を行う．一方でパラメータ数の多い BM25 以外の手法は，統計データコレクション全体に対するグリッドサーチでは時間がかかりすぎるため，チューニング用データセットを作成し，これを用いて提案手法と同様に座標上昇法（CA）で最適化を行う．パラメータの範囲は，$k_1, k_{1f} \in [0.0, 2.0]$，$b, b_f \in [0.0, 1.0]$，$\alpha_f \in [0.0, 1.0]$，$\beta_f, \beta_{f,f'} \in [0.0, \infty)$ とした．

チューニング用データセットの作成手順について説明する．3 章で構築したデータセットにおいては，各クエリに対して適合する文書（統計データ）はただ 1 つである．そこで，クエリごとにデフォルトパラメータの BM25 の top-100 と適合文書をプーリングし，これをチューニング用のデータセットとする．つまり，チューニング用の各クエリは，適合度が付与された (クエリ, 文書) のペアをたかだか 101 個持つ[*10]．パラメータチューニングにおいては，このデータセットに対してクエリごとに検索モデルでリランキングを行い，評価指標を計算した結果をもとに検索モデルのパラメータを更新する．本実験ではチューニング時に CA の最適化に利用する評価指標として MRR を用いることとした．

5.3.3　推論方法（テスト方法）

テスト時の推論方法（検索方法）について説明する．まず BM25 についてはチューニング時と同様に，統計データコレクション全体に対してテスト用クエリで top-1000 を検索し，評価指標を算出する．これに対して，提案手法を含む BM25 以外の手法については，プーリングした一部の文書集合に対してリランキングを行い，評価指標を算出することとした．リランキング用の文書集合は，ページタイトル+セクションタイトル，パラグラフ，コンテスト，カテゴリの 4 つのクエリフィールド[*11]について，それぞれ Anserini のデフォルトパラメータの BM25（$k_1 = 0.9$, $b = 0.4$）で top-1000 を検索し，その結果をプーリングすることによって作成した．このプーリング結果の文書集合の再現率は 0.937 であり，大半のクエリにおいて適合文書となる統計データを含む．

5.4　結果

本節では 5.1 節で提示した研究課題に回答する．

5.4.1　RQ1. クエリのフィールドと統計データのフィールドの両方を用いることで統計データ検索の性能は改善するか？

表 8 に交差検証の結果を示す．提案手法である BM25FF はすべての評価指標において，フィールドを考慮しない BM25 やクエリと統計データの片方のフィールドのみを考慮するモデルである BM25F と QF-BM25 を上回っている．特に MRR に関しては，BM25 や BM25F と比較してそれぞれ 224.1%，119.6%の性能の改善を得ることができている．

さらに，データセットの 443 トピック（クエリ）におけるベースライン手法と提案手法の計 4 システムの評価値に

[*10] top-100 に正解データが含まれる場合はクエリ–文書ペアは 100 個となり，含まれない場合は 101 個となる．

[*11] セクションタイトルは存在しない場合があるため，ここではページタイトルと 1 つにまとめることとした．

表 9　Ablation Study の結果（「-カテゴリ」はカテゴリのフィールドを除外したうえでパラメータチューニングと交差検証を行った結果を表す）

Table 9　Results of Ablation Study (“-カテゴリ” indicates the results of parameter tuning and cross-validation after excluding the category field).

	MRR	diff
BM25FF	0.305	
-ページタイトル	0.263	−0.043
-セクションタイトル	0.279	−0.027
-パラグラフ	0.317	0.012
-コンテキスト	0.298	−0.007
-カテゴリ	0.223	−0.083
-タイトル	0.327	0.021
-説明	0.308	0.003
-メタデータ	0.281	−0.024
-列ヘッダかつ行ヘッダ	0.281	−0.024
-列ヘッダ	0.278	−0.027
-行ヘッダ	0.257	−0.048
-データ	0.158	−0.148

ついて，繰り返しのない二元配置分散分析を行った．その結果，評価指標 MRR，Hit@10，Hit@20，Hit@100 に対するシステムの F 統計量 $F(3, 1326)$ はそれぞれ 84.6，54.1，28.5，70.6 であり，すべての評価指標においてシステム効果が統計的に有意であった（$p < 0.01$）．さらに Tukey HSD 検定により，すべての評価指標において提案手法である BM25FF と比較した 3 つのベースライン手法との間で統計的に有意な差が認められた（$p < 0.01$）．

以上の結果から，統計データ検索においてはクエリと統計データの両方のフィールドを用いることで性能が改善することが示された．

5.4.2　RQ2. クエリと統計データのどのフィールドが統計データ検索の性能に影響を与えるか？

表 9 に BM25FF に対して，1 つのフィールドを除外したうえでパラメータチューニングと交差検証を行う形で Ablation Study を行った結果を示す．この表からは，あるフィールド f を除外した際にもとの BM25FF に対して大きく性能が下がっている場合は BM25FF によるデータ検索おいてフィールド f は重要なフィールドであり，性能があまり下がっていない・上がっている場合はフィールド f はあまり重要でないフィールドである可能性が示唆される．

5.4.2.1　クエリフィールド

まずクエリのフィールドに関しては，最も大きく性能が下がったのはカテゴリを除外した場合であり，その次がページタイトルを除外した場合であった．ページタイトルは DBpedia の知識グラフにおいては 1 つのエンティティとして表現され，カテゴリも知識グラフを整理するうえで重要な概念である．よってデータ検索においてはクエリ中のエンティティが重要であり，クエリ内の語を知識グラフやエンティティと関連付けたうえで検索することで，よりよい検索結果になる可能性があると考えられる．

表 10　クエリフィールドごとのクエリ長と性能の関係．相関係数はクエリごとにクエリ長と MRR についてピアソンの積率相関係数を求めた

Table 10　Relationship between query length and performance for each query field. Correlation coefficients are Pearson’s r between query length and MRR for each query.

	単語数の平均値	相関係数
ページタイトル	2.6	0.130
セクションタイトル	1.9	−0.188
パラグラフ	90.2	−0.438
コンテキスト	37.8	0.221
カテゴリ	12.9	0.196

一方でパラグラフを除外した場合に関してはむしろ性能が上がってしまっている．これに関して，クエリフィールドごとの平均クエリ長（平均単語数）と，MRR と平均クエリ長の相関係数を計算した結果を**表 10** に示す．この表から，パラグラフは他のクエリフィールドと比較してクエリ長が大きく，かつ相関係数も小さくなっている．アドホック検索においてはクエリ長が大きい場合に検索性能が低くなることが知られており，これは Verbose クエリ [9] として知られている問題である．本実験においてもパラグラフは Verbose クエリと見なせるため，クエリ中の重要な単語を選別するなど既存の Verbose クエリへの対処を適用することで性能を改善できる可能性があると考えられる．

5.4.2.2　統計データフィールド

また，統計データのフィールドに関しては，最も大きく性能が下がったのはデータのフィールドを除外した場合であった．データのフィールドに関して実際に統計データを確認してみたところ，いくつかのデータにおいて，本来は列ヘッダや行ヘッダのフィールドである認識されるべき部分がデータのフィールドであると判定されてしまっていた．3.3.2 項で述べたとおり，現在の列ヘッダや行ヘッダのフィールド判定には簡単なルールベースの手法を用いており，フィールドの誤判定が発生してしまっていると考えられる．その結果，今回の実験においては誤ってデータが過剰に重要と判定されてしまっていると考えられる．

さらにフィールドの誤判定に関する追加の検証として，ルールベースによる列ヘッダと行ヘッダのフィールド判定精度の検証を行った．検証方法は次のとおりである．まず一部の統計データについて著者の 1 人が列ヘッダと行ヘッダのアノテーションを行った．次に，列ヘッダと行ヘッダのそれぞれについて，正解となるアノテーション結果とルールベースで得られた結果の一致度を Cohen’s κ で計算した．

検証対象の統計データとしては，次の 2 種類の統計デー

表 11　ヘッダ判定に関するアノテーション結果とルールベースの結果の一致度（Cohen's κ）

Table 11　Agreement between annotation results and rule-based results for header recognition (Cohen's κ).

統計データ	列ヘッダ	行ヘッダ
被引用統計データ（40 個）	0.68	0.52
その他の統計データ（40 個）	0.61	0.59
計（80 個）	0.64	0.55

タ（計 80 個）を用いることとする．1 種類目は 3 章で作成したデータセットに含まれる被引用統計データ 40 個である．しかしながら，これらの統計データは表 4 に示したとおり，一部の政府統計名を持つ統計データに偏っている．そのため，ルールベースのより正確な精度を検証するためには，より多様な統計データを対象として検証を行う必要があると考えられる．そこで 2 種類目の統計データとして，前述の統計データとは別の統計データを 40 個選択し，検証に用いることとした．選択の方法としては，まず前述の統計データに含まれない政府統計名をランダムに 40 個選択し，選択した政府統計名からそれぞれ 1 つずつランダムに統計データを選択した．

列ヘッダおよび行ヘッダに対するアノテーションは次の手順で行った．まず，対象とした統計データのファイルを開き，表が含まれる範囲を特定した．その後，表の範囲のうち，列ヘッダとなる行と行ヘッダとなる列を特定した．ある行が列ヘッダに該当するかは，その行に含まれるセル値の大半が各列の値を表現する属性であるかを基準として判定した．ただし，都道府県 ID などの ID の行や，列番号の行は列ヘッダに該当しないこととした．行ヘッダについても同様の基準で判定を行った．

列ヘッダと行ヘッダのそれぞれについて，アノテーション結果とルールベースで得られた結果の一致度を，Cohen's κ を用いて計算した結果を**表 11** に示す．一致度は列ヘッダで 0.64，行ヘッダで 0.55 と高いとはいえない結果であり，ヘッダの認識結果には誤りが含まれていると考えられる．これは「データ」のフィールドが重要であるという結果の原因として，フィールド判定の誤りが原因である可能性を示唆していると考えられる．また，被引用統計データとその他の統計データとの間に著しい差は見られなかった．

一方でタイトルや説明を除外した場合は性能が上がってしまっている．これはタイトルや説明は，他の多くの統計データと同一の場合が多く，かつメタデータに同じ情報が含まれている場合が多いため，上位の絞り込みにあまり効果がなかった可能性が考えられる．

6. まとめ

本論文では，数値から被引用統計データを特定するタスクである統計データ特定問題を提案し，この問題をアドホック検索の問題として定式化を行った．次に，この問題に対する新たな検索モデルとして，クエリと検索対象文書の統計データの両方のフィールドを考慮した BM25FF を提案した．さらに，Wikipedia と e-Stat を用いて統計データ特定のための新たなデータセットを構築し，提案手法とベースライン手法の比較実験を行った．その結果，クエリと統計データの両方のフィールドを用いた提案手法は，ベースライン手法を上回る性能を発揮することが判明した．

今後の課題としては，統計データのフィールドの認識精度の向上させることや，統計データ中のどのセルを参照しているかを特定する問題であるセル特定の問題に取り組むことが考えられる．後者の統計データ中のどのセルを参照しているかを特定する問題（ここではセル特定問題と呼ぶ）に関しては，1 節で述べたとおり，表に対する質問応答が適用できる可能性がある．たとえば，TaPas [12] モデルは，入力として質問と表の組が与えられた際に，まず質問に対する答えとして表のどのセルが用いられるかの確率について，各セルごとに予測を行う．ここで質問と表の組を文書と統計データの組に置き換えると，セル特定問題にも TaPas モデルが適用でき，文書から各セルが参照されている確率が予測できると考えられる．この適用可能性の検証については，今後の課題とする．

謝辞　本研究は JSPS 科研費 18H03244，18H03243，および，JST さきがけ JPMJPR1853 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Bagheri, E. and Al-Obeidat, F.N.: A Latent Model for Ad Hoc Table Retrieval, *Proc. 42nd European Conference on IR Research*, pp.86–93, Springer (2020).

[2] Balaneshinkordan, S., Kotov, A. and Nikolaev, F.: Attentive Neural Architecture for Ad-hoc Structured Document Retrieval, *Proc. 27th ACM International Conference on Information and Knowledge Management*, pp.1173–1182, ACM (2018).

[3] Boland, K., Ritze, D., Eckert, K. and Mathiak, B.: Identifying References to Datasets in Publications, *Proc. 2nd International Conference of Theory and Practice of Digital Libraries*, pp.150–161, Springer (2012).

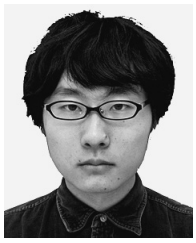
[4] Chen, Z., Jia, H., Heflin, J. and Davison, B.D.: Leveraging Schema Labels to Enhance Dataset Search, *Proc. 42nd European Conference on IR Research, Part I*, pp.267–280, Springer (2020).

[5] Chen, Z., Trabelsi, M., Heflin, J., Xu, Y. and Davison, B.D.: Table Search Using a Deep Contextualized Language Model, *Proc. 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.589–598, ACM (2020).

[6] Choi, J.I., Kallumadi, S., Mitra, B., Agichtein, E. and Javed, F.: Semantic Product Search for Matching Structured Product Catalogs in E-Commerce, arXiv, Vol.abs/2008.08180 (2020).

[7] Demartini, G., Iofciu, T. and de Vries, A.P.: Overview of the INEX 2009 Entity Ranking Track, *Focused Retrieval and Evaluation, 8th International Workshop of the Ini-*

tiative for the Evaluation of XML Retrieval, pp.254–264, Springer (2009).
[8] Fetahu, B., Markert, K., Nejdl, W. and Anand, A.: Finding News Citations for Wikipedia, *Proc. 25th ACM International Conference on Information and Knowledge Management*, pp.337–346, ACM (2016).
[9] Gupta, M. and Bendersky, M.: Information Retrieval with Verbose Queries, *Foundations and Trends in Information Retrieval*, Vol.9, No.3-4, pp.91–208 (2015).
[10] Hasibi, F., Nikolaev, F., Xiong, C., Balog, K., Bratsberg, S.E., Kotov, A. and Callan, J.: DBpedia-Entity v2: A Test Collection for Entity Search, *Proc. 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1265–1268, ACM (2017).
[11] He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, C.L.: Context-aware citation recommendation, *Proc. 19th International Conference on World Wide Web*, pp.421–430, ACM (2010).
[12] Herzig, J., Nowak, P.K., Müller, T., Piccinno, F. and Eisenschlos, J.M.: TaPas: Weakly Supervised Table Parsing via Pre-training, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.4320–4333, Association for Computational Linguistics (2020).
[13] Kato, M.P., Ohshima, H., Liu, Y.-H. and Chen, H.-L.: Overview of the NTCIR-15 Data Search Task, *Proc. 15th NTCIR Conference on Evaluation of Information Access Technologies*, pp.267–273, NII (2020).
[14] Kim, J., Xue, X. and Croft, W.B.: A Probabilistic Retrieval Model for Semistructured Data, *Proc. 31th European Conference on IR Research*, pp.228–239, Springer (2009).
[15] Metzler, D. and Croft, W.B.: Linear feature-based models for information retrieval, *Information Retrieval*, Vol.10, No.3, pp.257–274 (2007).
[16] Pasula, H., Marthi, B., Milch, B., Russell, S.J. and Shpitser, I.: Identity Uncertainty and Citation Matching, *Proc. 15th International Conference on Neural Information Processing Systems*, pp.1425–1432, MIT Press (2002).
[17] Peng, H., Liu, J. and Lin, C.: News Citation Recommendation with Implicit and Explicit Semantics, *Proc. 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, The Association for Computer Linguistics (2016).
[18] Redi, M., Fetahu, B., Morgan, J.T. and Taraborelli, D.: Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability, *Proc. 2019 World Wide Web Conference*, pp.1567–1578, ACM (2019).
[19] Robertson, S.E., Zaragoza, H. and Taylor, M.J.: Simple BM25 extension to multiple weighted fields, *Proc. 13th ACM International Conference on Information and Knowledge Management*, pp.42–49, ACM (2004).
[20] Sarma, A.D., Fang, L., Gupta, N., Halevy, A.Y., Lee, H., Wu, F., Xin, R. and Yu, C.: Finding related tables, *Proc. 2012 ACM SIGMOD International Conference on Management of Data*, pp.817–828, ACM (2012).
[21] Shraga, R., Roitman, H., Feigenblat, G. and Canim, M.: Ad Hoc Table Retrieval using Intrinsic and Extrinsic Similarities, *Proc. Web Conference 2020*, pp.2479–2485, ACM (2020).
[22] Shraga, R., Roitman, H., Feigenblat, G. and Canim, M.: Web Table Retrieval using Multimodal Deep Learning, *Proc. 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1399–1408, ACM (2020).
[23] Tkaczyk, D., Collins, A., Sheridan, P. and Beel, J.: Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers, *Proc. 18th ACM/IEEE on Joint Conference on Digital Libraries*, pp.99–108, ACM (2018).
[24] Yan, A. and Weber, N.M.: Mining Open Government Data Used in Scientific Research, *Proc. 13th International Conference on Information*, pp.303–313, Springer (2018).
[25] Yang, P., Fang, H. and Lin, J.: Anserini: Reproducible Ranking Baselines Using Lucene, *ACM Journal of Data and Information Quality*, Vol.10, No.4, pp.16:1–16:20 (2018).
[26] Zamani, H., Mitra, B., Song, X., Craswell, N. and Tiwary, S.: Neural Ranking Models with Multiple Document Fields, *Proc. 11th ACM International Conference on Web Search and Data Mining*, pp.700–708, ACM (2018).
[27] Zhang, S. and Balog, K.: Ad Hoc Table Retrieval using Semantic Similarity, *Proc. 2018 World Wide Web Conference on World Wide Web*, pp.1553–1562, ACM (2018).
[28] Zhiltsov, N., Kotov, A. and Nikolaev, F.: Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data, *Proc. 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.253–262, ACM (2015).



中野 優　（学生会員）

2016 年京都大学工学部情報学科卒業．2018 年同大学大学院修士課程修了．同年株式会社 NTT データ数理システムに入社．データ分析ソフトウェア開発に従事．2020 年筑波大学大学院に入学．現在，筑波大学大学院人間総合科学学術院人間総合科学研究群情報学学位プログラム博士後期課程在学中．日本データベース学会，ACM 各学生会員．ACM SIGIR 東京支部会員．

加藤 誠　（正会員）

2012 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．現在，筑波大学図書館情報メディア系准教授．情報検索の研究に従事．電子情報通信学会，日本データベース学会，ACM，ACM SIGIR 東京支部各会員．

（担当編集委員　欅 惇志）