DEIM2021 F25-4

誤引用検証ための被引用統計データの検索

中野　優†　加藤　誠††

† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s2030514@s.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　本論文では数値を含む文書の真偽を検証するために，文章が参照する統計データを検索する手法を提案する．特に，検索対象である統計データが構造化されている点に加えて，クエリとなる文章もタイトルなどの付随する情報から構造化されている点に着目し，クエリ文章と検索対象の統計データの双方の構造を考慮した検索手法を提案する．提案手法を検証するために，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．実験ではこのデータセットを用いて，ベースライン手法と提案手法を比較検討し，提案手法が最も高い性能を示すことがわかった．

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，メタデータ

1 はじめに

文章を執筆する際において，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．例えば，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 e-Stat[1] の統計データが引用されると考えられる．Redi らの研究 [15] において公開されているデータ [2]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% を占めている．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**とよぶ．統計データ引用は論文においても行われており [20]，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．例えば「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 [19] 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．例えば，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．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，文章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，文章がどの統計データを引用しているかを特定する問題と，文章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．特に本論文では前者の問題を検索の問題として定式化を行い，この問題を**統計データ検索問題**と呼ぶ．つまり，統計データから数値を引用している文章をクエリとみなし，被引用統計データを検索対象のアイテムとみなして，統計データを検索する問題に取り組む．

さらに本論文では，クエリとなる数値情報を含む文章と検索対象の統計データの双方の構造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，検索対象である統計データがメタデータなどの複数のフィールドを持つことに加えて，クエリとなる数値情報を含む文章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．情報検索においては，検索対象のアイテム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 [5, 12, 16, 23, 24]．そこで本研究では，統計データの構造だけではなくクエリの構造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，クエリとなる文章と検索対象の統計デー

1：https://www.e-stat.go.jp/
2：https://figshare.com/articles/Citation_Reason_Dataset/7756226

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，Wikipedia を統計データを引用する文章として用い，e-Stat の政府統計データを被引用統計データとして用いた．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．その結果，提案手法はクエリ文章や統計データ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．

本論文の貢献は以下のとおりである．

（1） 本論文では，文章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に対して，クエリとなる文章と検索対象の統計データの双方の構造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，クエリ・統計データの両方を用いることで，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わかった．また，どのクエリ・統計データのフィールドが提案手法において重要かについて検証した．

本論文の構成は次の通りである．第 2 節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．第 3 節では統計データ検索の問題設定と提案手法であるクエリ文章と検索対象の統計データの双方の構造を考慮した検索モデルについて説明する．第 4 節では Wikipedia と e-Stat を用いた統計データ検索のためのデータセット構築方法について説明し，第 5 節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する．第 6 節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 [19] や，被引用データセットを特定する研究 [2] 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．

また，文書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も行われている．推薦の対象となるアイテムとしては，学術論文を推薦する研究 [10] やニュースを推薦する研究 [7, 14] 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，データセット中の表を検索するタスクである [4, 18, 23]．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による semantic matching によって表を検索する手法を提案している [23]．また，Chen らは BERT による表検索を行っており，BERT の入力長制限を回避するために表から検索に有効な情報を抽出する手法を提案している [4]．本論文はクエリが（短い）キーワードではなく（長い）文章であるという点と，既存研究の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 [3] 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章である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，検索対象のアイテム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 [16, 22]，エンティティ検索 [24]，表検索 [4, 18, 23]，XML 検索 [12]，商品検索 [5] などが存在する．複数フィールドを持つアイテムに対する検索においては，フィールドごとにテキストの長さが異なったり，フィールドごと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，検索対象のアイテム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 [16, 23]．近年は

ニューラルネットワークを用いた検索モデルが研究されており，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 [1, 22]．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．例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 [6] 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 [17] 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 統計データ検索

本節ではまず 3.1 節で本論文が扱う問題設定について説明し，3.2 節で基本的な検索モデルについて説明した後，3.3 節で提案手法である検索モデルについて説明する．

3.1 問題設定

統計データ検索問題の定式化を説明する．本問題では，与えられたクエリ q と検索対象アイテムの集合 D に対して，本検索タスクは D に含まれるアイテムをランク付けする問題であり，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．

本論文におけるクエリ $q \in Q$ は数値情報を含む文章であり，統計データを引用する数値 q_{num} を含む．ただし，Q は任意のクエリ（ここでは特定すべき数値情報を含む文章）の集合である．クエリ q は文脈として数値 q_{num} が記述されている文書 q_{con} を持つ．さらに，q_{con} 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．

本論文における検索対象のアイテム $d \in D$ 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．

本論文ではこのランキング問題を，クエリとなる文章 $q \in Q$ と検索対象の統計データ $d \in D$ を入力として，スコアを出力するスコア関数 $s : Q \times D \to \mathbb{R}$ を設計する問題とみなす．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．

3.2 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．

a） 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について説明する．このモデルは，クエリとなる数値情報を含む文章の構造や，検索対象の統計データの構造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，通常の文書検索と同様に検索を行うモデルである．例えば，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 のスコア関数は以下で表される．

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

ただし，w はクエリ q に含まれる単語，$\mathrm{idf}(w)$ 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），$\mathrm{dl} : D \to \mathbb{N}$ は文章 d を引数として d の文書長（単語数）を返す関数を表す．

b） 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 節で説明した通り，情報検索において検索対象のアイテムが複数のフィールドを持つことは多い．また，統計データに類似するアイテムである表の検索においても，表をタイトルや列名など複数フィールドを持つアイテムとみなして検索することが一般的である [4, 18, 23]．

複数フィールドを持つアイテムの検索においては，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，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．例えば BM25F [16] はフィールドごとの単語頻度の重み付け和を BM25 の単語頻度の代わりに用いることで，フィールドごとの特徴を考慮した検索を行う．具体的には，以下のスコア関数 s によりフィールドごとの特徴を考慮したスコア付けを行う．

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

$$\widetilde{\mathrm{tf}}(w) = \sum_{f \in F_D} \beta_f \cdot \mathrm{tf}_f(w) \quad (3)$$

ただし，tf_f は統計データのフィールド f における単語頻度を表し，β_f はフィールドごとのパラメータである．

3.3 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．

3.3.1 スコア関数

既存のフィールドを考慮した検索モデルは，検索対象のアイテム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，クエリとなる数値情報を含む文章も複数のフィールドを持ちうる．そこで本論文では，BM25F をクエリのフィールドも考慮できるように拡張した **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．

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

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

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

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，$\mathrm{ql}: Q \to \mathbb{N}$ はクエリ q を引数にとり q のクエリ長を返す関数である．式 5 と式 6 は BM25F においてはそれぞれ式 2 と式 3 に対応しており，式 4 はクエリのフィールドごとに 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．

3.3.2 パラメータチューニング

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）[13] 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 [9]，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．

4 データセット

本節では統計データ検索のベンチマーク用データセットの構築方法について説明する．

4.1 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 [11] 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

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%

表 2: e-Stat（統計データ）と WikiTables [23]（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．

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 [23]	5.0	4	12.1	6	26.0%	16.7%

4.2 数値情報を含む文書からのデータセット構築

数値情報を含む文章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，数値情報を含む文章からその文章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どの統計データを引用しているかについて特定することができた．

次に，統計データを引用するセクション中の数値が実際に統計データから引用された数値かどうかを人手で特定した．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，対象の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することとした．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．

4.3 データセット分析

本節では統計データセットに関する分析と，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．

a） 統計データセットの分析

まず，4.1 節で説明した e-Stat の統計データのデータセット

表 3: データセット中の Wikipedia 記事のカテゴリの上位 5 件

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

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

と，既存の表に関するデータセットである WikiTables データセット [23] 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り，2.2 節で説明したアドホック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．ただし数値のセルの割合とは，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したときの，(数値のセルの個数)/(全セルの個数) のことである．また，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たデータセットに対する結果である．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言える．

b） 統計データ検索のデータセット分析

次に構築した統計データ検索用データセットの分析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．

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

(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

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．そのため，統計データの数値に関して誤りが含まれていたとしても修正が行われる可能性は低いと考えられる．よって，数値が引用する統計データを自動的に特定することで，数値に誤りがあるかどうかの判定が容易になり，統計データの引用に関する誤りの修正も容易になると考えられる．

5 実　　験

本節では前節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，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？

5.2 フィールド設定

本実験のフィールド設定について説明する．

5.2.1 クエリ文章のフィールド設定

本実験で用いるクエリ文章のフィールドを図 1 に示す．本実験で用いるクエリ文章のフィールドの集合は $F_Q = \{$ ページタイトル, セクションタイトル, パラグラフ, コンテキスト, カテゴリ $\}$ であり，$|F_Q| = 5$ である．各フィールドの詳細は以下の通りである．

A. ページタイトル 統計データから引用された数値 q_{num} を含む Wikipedia のページのタイトルである．

B. セクションタイトル q_{num} を含むセクションのタイトルと，その祖先のセクションのタイトルからなる．

C. パラグラフ q_{num} を含むパラグラフである．本実験では，改行が 2 連続する部分をセクションの分かれ目とみなした．た

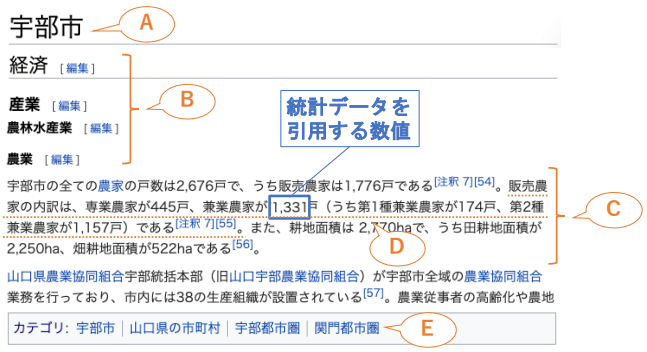

図 1: クエリである Wikipedia 文章のフィールド

(a) メタデータ　　(b) 表

図 2: 検索対象の統計データのフィールド

だし，パラグラフが q_{num} 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　q_{num} の周辺のテキストである．本実験では，q_{num} 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　q_{num} を含むページのカテゴリからなる．

5.2.2　統計データのフィールド設定

本実験で用いる統計データのフィールドを図 2 に示す．本実験で用いる統計データのフィールドの集合は $F_Q = \{$ タイトル, 説明, メタデータ, 列ヘッダかつ行ヘッダ, 列ヘッダ, 行ヘッダ, データ $\}$ であり，$|F_Q| = 7$ である．各フィールドの詳細は以下の通りである．

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　統計データの表のうち，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，統計データの先頭の 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みなす（行ヘッダも同様）．

e. 列ヘッダ, **f. 行ヘッダ**　統計データの表の列ヘッダ（列名）もしくは行ヘッダ（行名）である．ただし，行ヘッダかつ列ヘッダの部分を除く．

g. データ　統計データの表のうち，行ヘッダかつ列ヘッダ，行ヘッダ，列ヘッダ以外のセルの文字列からなる．

5.3　実験設定

本節では実験設定として，評価指標とベースライン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．

表 6: 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．

	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

a）評価指標

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 つしか存在しないため，既存のデータセット検索や表検索で用いられる 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．

b）比較手法

本実験で比較する手法を表 6 に示す．各手法の詳細は以下の通りである．

BM25　このベースラインはクエリ文章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 [21][3] を用いる．

BM25F [16]　このベースラインはクエリ文章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 3.3 節で説明した通り，クエリ文章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．

c）評価手順

本実験では上記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．

d）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それ以外の手法は，全コレクション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

3：http://anserini.io

表 7: 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）

	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	219/111/113	0.318	0.345	0.442
QF-BM25	0.080	243/ 65/135	0.237	0.266	0.637
BM25FF	**0.305**	311/ 41/ 91	**0.395**	**0.444**	**0.731**

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．パラメータの範囲としては，$k_1, k_{1f} \in [0.0, 2.0], b, b_f \in [0.0, 1.0], \alpha_f \in [0.0, 1.0], \beta_f, \beta_{f,f'} \in [0.0, \infty)$ 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．4 節で構築したデータセットは，クエリ（=Wikipedia）とただ 1 つの適合文書（=統計データ）のみを持つ．そこで，クエリごとにデフォルトパラメータの BM25 の top-100 とそのクエリの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-文書ペアを高々 101 個持つ．これらのクエリ-文書ペアに対してクエリごとに各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各検索モデルはパラメータを更新する．本実験では CA で最適化に使用する評価指標として MRR を用いることとした．

e） 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，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．提案手法を含むそれ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 [4]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，大半のクエリで適合文書であ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．

a） 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 7 に交差検証の結果を示す．提案手法である BM25FF は全ての評価指標において，フィールドを考慮しない BM25 やクエリ文章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 224.8%, 119.6% の性能の改善を得ることができている．この結果は，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改善することができることを示している．

4：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 8: 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上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）

	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

b） 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 8 に BM25FF に対して，1 つのフィールドを除外した上でパラメータチューニングと交差検証を行うという形で Ablation Study を行った結果を示す．この表からは，そのフィールド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あまり重要でないフィールドであるという考察が可能である．

まずクエリ文章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上で重要な概念である．よって，データ検索においては，クエリ内の語を知識グラフやエンティティと関連付けた上で検索することで，よりよい検索結果になる可能性があると考えられる．一方でコンテキストを除外した場合はあまり性能が下がらず，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．これは，コンテキストやパラグラフはクエリ長が大きく，アドホック検索において Verbose Query [8] として知られている問題と関連していると推測される．

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．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，数値情報を含む文章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，クエリ文章と検索対象の統計データの両方の構造を考慮した 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，クエリ文章と統計データの両方の構造を用いた提案手法は，既存手法を上回る性能を見せることが判明した．

今後の課題としては，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．

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．

文　献

[1] Saeid Balaneshinkordan, Alexander Kotov, and Fedor Nikolaev. Attentive neural architecture for ad-hoc structured document retrieval. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1173–1182. ACM, 2018.

[2] Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. Identifying references to datasets in publications. In *Proceedings of the 2nd International Conference of Theory and Practice of Digital Libraries*, pages 150–161. Springer, 2012.

[3] Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D. Davison. Leveraging schema labels to enhance dataset search. In *Proceedings of the 42nd European Conference on IR Research, Part I*, pages 267–280. Springer, 2020.

[4] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D. Davison. Table search using a deep contextualized language model. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 589–598. ACM, 2020.

[5] Jason Ingyu Choi, Surya Kallumadi, Bhaskar Mitra, Eugene Agichtein, and Faizan Javed. Semantic product search for matching structured product catalogs in e-commerce. *arXiv*, abs/2008.08180, 2020.

[6] Gianluca Demartini, Tereza Iofciu, and Arjen P. de Vries. Overview of the INEX 2009 entity ranking track. In *Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 254–264. Springer, 2009.

[7] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. Finding news citations for wikipedia. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 337–346. ACM, 2016.

[8] Manish Gupta and Michael Bendersky. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval*, 9(3-4):91–208, 2015.

[9] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268. ACM, 2017.

[10] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 421–430. ACM, 2010.

[11] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. Overview of the NTCIR-15 data search task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.

[12] Jinyoung Kim, Xiaobing Xue, and W. Bruce Croft. A probabilistic retrieval model for semistructured data. In *Proceedings of the 31th European Conference on IR Research*, pages 228–239. Springer, 2009.

[13] Donald Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[14] Hao Peng, Jing Liu, and Chin-Yew Lin. News citation recommendation with implicit and explicit semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[15] Miriam Redi, Besnik Fetahu, Jonathan T. Morgan, and Dario Taraborelli. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *Proceedings of the 2019 World Wide Web Conference*, pages 1567–1578. ACM, 2019.

[16] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 42–49. ACM, 2004.

[17] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Y. Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 817–828. ACM, 2012.

[18] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Canim. Ad hoc table retrieval using intrinsic and extrinsic similarities. In *Proceedings of the Web Conference 2020*, pages 2479–2485. ACM / IW3C2, 2020.

[19] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Jöran Beel. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 99–108. ACM, 2018.

[20] An Yan and Nicholas M. Weber. Mining open government data used in scientific research. In *Proceedings of the 13th International Conference on Information*, pages 303–313. Springer, 2018.

[21] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using lucene. *ACM Journal of Data and Information Quality*, 10(4):16:1–16:20, 2018.

[22] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. Neural ranking models with multiple document fields. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 700–708. ACM, 2018.

[23] Shuo Zhang and Krisztian Balog. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web,*, pages 1553–1562. ACM, 2018.

[24] Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–262. ACM, 2015.