COTER: Conditional Optimal Transport meets Table Retrieval

Anonymous Author(s)

ABSTRACT

Ad hoc table retrieval refers to the task of performing semantic matching between given queries and candidate tables. In recent years, the approach to addressing this retrieval task has undergone significant shifts, transitioning from utilizing hand-crafted features to leveraging the power of Pre-trained Language Models (PLMs). However, key challenges arise when candidate tables contain shared items, and/or queries may refer to only a subset of table items rather than the entire one. Existing models often struggle to distinguish the most informative items and fail to accurately identify the relevant items required to match with the query.

To bridge this gap, we propose **C**onditional **O**ptimal **T**ransport based table retriev**ER** (COTER). The proposed algorithm is characterized by simplifying candidate tables, where the semantic meaning of one or several words (from the original table) is enabled to be effectively "transported" to individual words (from the simplified table), under the prior condition of the query. COTER achieves two essential goals simultaneously: minimizing the semantic loss during the table simplification and ensuring that retained items from simplified tables effectively match the given query. Importantly, the theoretical foundation of COTER empowers it to adapt dynamically to different queries and enhances the overall performance of the table retrieval. Experiments on two popular Web-table retrieval benchmarks show that COTER can effectively identify informative table items without sacrificing retrieval accuracy. This leads to the new state-of-the-art with substantial gains of up to 0.48 absolute Mean Average Precision (MAP) points, compared to the previously reported best result.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

Conditional Optimal Transport, Table Retrieval, Semantic Matching, Table representation, Web search

ACM Reference Format:
Anonymous Author(s). 2024. COTER: Conditional Optimal Transport meets Table Retrieval. In *WSDM '24: The 17th ACM International Conference on Web Search And Data Mining, March 4 to March 8, 2024, Mérida, Mexico.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

WSDM '24, March 4 to March 8, 2024, Mérida, Mexico

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Web tables are immensely valuable tools for effectively interpreting, manipulating, and managing data. Composed of multiple *cells* organized into *rows* and *columns*, tables offer a versatile format that can accommodate a wide range of *items*, including numbers, characters, and strings. There are billions of tables available across the Web (with various markup formats like HTML, Wikitext, Markdown, *etc*.), providing a wealth of information resource to Web queries in search engines [13, 35].

Due to the popularity of tabular data, the necessity of automating table retrieval in response to a user query becomes essential for numerous downstream tasks, including table fact verification [29], table question answering [3], and table-to-text generation [20], among others. Extensive and diverse research efforts then have been dedicated to assessing the relevance between a given user query and candidate tables, underscoring the significance of table retrieval.

Early research treats tables as documents and applies document retrieval techniques [24, 26, 33], and employs hand-crafted features (such as term frequencies and number of rows/columns), making the process inefficient and time-consuming. Recently, the adoption of Pre-trained Language Models (PLMs, like BERT [7]) has achieved notable advancements in table retrieval [4, 25, 28], since they can automatically extract features with the help of multi-head attention and thus provide better context/layout representation.

Search Query: countries capital

Country or territory	Languages (official in bold)	Capital
Canada	Languages of Canada English	Ottawa
Mexico	Languages of Mexico Spanish	Mexico City
United States	Languages of the United States English	Washington, D.C.

Status	Country	Legal since
Marriage performed nation wide (1 country)	Same-sex marriage in Canada Canada	2005
Limited recognition (2 countries)	Same-sex marriage in Mexico Mexico	2000
Limited recognition (2 countries)	Same-sex unions in the United States United States	2013

Figure 1: Shared items between tables can mislead the retrieval model. In this example, both tables (from WikiTables [33]) include country names (highlighted in green), but only the top table accurately represents respective capitals (highlighted in blue) in response to the query. In contrast, the bottom table discusses marriage laws (highlighted in yellow) while also mentioning country names.

Although neural models have dramatically advanced the table retrieval task, there are three major drawbacks of prior approaches. Firstly, the use of tables with multiple rows and columns often leads to lengthy inputs to retrieval models, which significantly affects their efficiency and scalability, not to mention the computational overhead. Secondly, not all table items are equally relevant to a given query, and essential semantic information can be overlooked. Thirdly, the existence of shared items across different tables introduces an additional layer of complexity, as overlapping contents could potentially mislead the retrieval models. An illustration is shown in Fig. 1 featuring two tables, both mentioning country names. However, it is important to note that only the top table provides accurate information regarding the query on "countries capitals", while the bottom one primarily pertains to marriage laws.

In this paper, we concentrate on how to choose significant table items provided a given user query with satisfactory retrieval performance. We first introduce the concept of conditional optimal transport (COT), and further propose a novel retrieval algorithm, termed **C**onditional **O**ptimal **T**ransport based table retriev**ER** (COTER). The proposed algorithm considers significant table items as a subset of the original table data and formulates the item selection as a conditional optimal transport model. The process of selecting items is achieved by minimizing the overall (semantic) transportation cost, subject to the specified condition or, more precisely, the query in our context. This ensures that the condensed subset conveys nearly identical meaning or, in simpler terms, offers similar "semantic coverage".

The cooperation with conditional OT-based selection yields several benefits. Firstly, the proposed algorithm efficiently identifies a subset of items, rather than processing the entire table, making it a computationally affordable option. Secondly, selected items adapt dynamically to the given query, ensuring that they precisely align with the query context. Lastly, chosen items, which substantially contribute to subsequent retrieval tasks, are easily interpretable, thereby enhancing the overall usability and interpretability of the retrieval model. To the best of our knowledge, this is the first study explicitly investigating conditional optimal transport theory in table retrieval tasks. The main contributions of our proposed work are summarized as follows[1]:

- The proposed method introduces the concept of conditional optimal transport (COT), extending the traditional optimal transport theory to accommodate a transport plan dependent on a conditional variable;
- The proposed method utilizes COT to identify informative table items by minimizing semantic loss during the table simplification and dynamically adapting the retained items to different queries;
- Empirically, our proposed method exhibits significant improvements, surpassing state-of-the-arts on two Web-table retrieval benchmarks. Specifically, we achieve a substantial gain of +0.48 absolute Mean Average Precision (MAP) points, solidifying the effectiveness and superiority of our approach.

2 RELATED WORK

2.1 Table Retrieval

This task aims to retrieve semantically relevant tables from a table corpus based on a given user query. Zhang and Balog [33] present systematically the task of ad hoc table retrieval from an information retrieval perspective and propose the **STR** method as the retrieval approach. STR employs hand-crafted terms (either words or entities from the query/table) and produces their semantic representation to estimate the query-table relevance. Extending the STR approach, **TRUKTQ** [34] introduces three semantic representations from discrete sparse and continuous dense spaces, along with four similarity measures. This allows for the generation of matching scores between the query and candidate tables by exploring all possible combinations of these semantic representations and similarity measures.

Yet, the semantic vectors derived from STR/TRUKTQ rely on hand-crafted features and lack contextual and task-specific information. To address this limitation, **Table2Vec** is proposed in [32], which introduces four embedding variants that leverage various table elements such as captions and headings. Similarly, **MCON** [26] proposes encompass table metadata in order to learn their embeddings. Bagheri *et al*. [2] suggest representing tabular data using low-dimensional latent factor matrices, and Trabelsi *et al*. propose **DSRMM** [23] via utilizing a combination of convolutional neural networks (CNNs), kernel pooling and Term-Gating network to identify query-table pairs. Additionally, **MTR** [19] utilizes recurrent neural networks (RNNs) and CNNs to acquire representations of queries and tables. **RGTR** [24] incorporates external knowledge from WordNet and pre-trained Glove embeddings, and further employs multi-relational graph convolutional networks to extract embeddings. By incorporating these fine-grained contextual-aware embeddings, the effectiveness of table retrieval is significantly improved.

In recent years, significant advancements have been made in the field of table retrieval by leveraging Pre-trained Language Models (PLMs), such as BERT [7], which have demonstrated promising results. One approach, proposed by Chen *et al*. [4], is **TSDM**, which utilizes BERT to encode flattened tables and combines resultant features with conventional embeddings from STR. A few studies explore the pre-training of BERT-like models specifically designed to handle table structures. Examples include Tapas [11], Tapex [16], and TaBERT [30].

Additionally, another line of research also takes the table layout/structure into account, and formulates tables as hypergraphs by defining different types of nodes and edges. **MGNEST** [5] models the tables' relation using one or more graphs and applies the point-wise mutual information to estimate the semantic correlation between queries and tables. **GTR** [28] constructs one tabular graph with cell, row and column nodes covering different granularities. The Graph Transformer model is then utilized to calculate the latent representation for both table item and layout structures.

A more recent advancement is **StruBERT** [25], which introduces the horizontal self-attention (via extending the vertical self-attention from TaBERT) to integrate the tabular structure and textual item. **QMTR** [14] investigates three qualitative metrics, *i.e.*, coherence, interpretability and exactness. By interpolating these metrics, QMTR offers a systematic approach to address the table retrieval task, and shows promise in improving the retrieval performance.

Nevertheless, previous studies have treated all table items as equally important, resulting in an inability to distinguish the informative items. Shared items among various tables further exacerbates the complexity, potentially confusing the retrieval models. In contrast, our proposed method aims to preserve the utmost informative items by considering the provided query as prior conditions. This deliberate approach significantly enriches the efficacy of the retrieval models.

[1] The source code will be made available upon acceptance.

2.2 Optimal Transport (OT)

The OT theory has recently garnered significant attention and has been applied to several tasks, such as cross-lingual information retrieval [12], text matching [15], text style transfer [18], and document ranking [31], among others. Formally, OT addresses a transportation problem that transports goods from a collection of suppliers $\mathbf{U} = \{u_i | i = 1, ..., |\mathbf{U}|\}$ to a collection of customers $\mathbf{V} = \{v_j | j = 1, ..., |\mathbf{V}|\}$, where u_i and v_j indicate the supply quantity of the i-th supplier and the order quantity of the j-th customer, respectively. Additionally, let $p_{ij}(\geq 0)$ and $c_{ij}(\geq 0)$ be the quantity transported from the i-th supplier to the j-th customer and its cost, an optimal transport plan $\mathbf{P}^* = \{p^*_{i,j}\} \in \mathbb{R}^{|\mathbf{U}|\times|\mathbf{V}|}$ in pursuit of minimizing the transportation cost can be obtained by solving the following optimization problem:

$$\mathbf{P}^* = \operatorname*{argmin}_{\mathbf{P}} \sum_i^{|\mathbf{U}|} \sum_j^{|\mathbf{V}|} p_{ij} c_{ij}, \quad s.t. \sum_{j=1}^{|\mathbf{V}|} p_{ij} = u_i, \quad \sum_{i=1}^{|\mathbf{U}|} p_{ij} = v_j, \quad (1)$$

where the constraints indicate the quantity requirements for both suppliers and customers.

By contrast, the proposed method is different from existing OT approaches in the sense that the transportation plan dynamically adapts based on a prior condition, for which we introduce the `conditional optimal transport`. In this context, the transport quantity p_{ij} is no longer self deterministic but becomes dependent on a conditional variable q (*i.e.*, the query in the context of table retrieval). Consequently, the total optimal transport plan $\mathbf{P}^*$ also changes accordingly as q varies.

3 PROPOSED METHOD

To begin with, we first describe the baseline model. Then we elaborate the proposed COTER with the conditional optimal transport, to simplify the input tables while minimizing the semantic loss and ensuring strong alignment with the given query. The COTER is integrated with mainstream encoders (such as BERT) into a unified framework and fully end-to-end trainable.

3.1 Preliminary

Let $C = \{T_1, \cdots, T_k, \cdots, T_N\}$ be the table corpus with N tables, where T_k represents the k-th table. Furthermore, T_k is linearized by concatenating all its context fields (items), such as table caption, page title, headers, and cell values. That is, $T_k = \{t_1, \ldots, t_i, \ldots, t_{|T_k|}\}$ and t_i represents the i-th tokenized item (although one item may contain more than one token, we treat them as a cohesive unit).

Then, given a tokenized user query q, the ad hoc table retrieval task is to assess T_k independently ($\forall T_k \in C$) and estimate the relevance, say $\mathtt{score}(q, T_k)$. Specifically, the q is concatenated with T_k to form the model input, *i.e.*,

$$\texttt{[CLS]}\, t_1, \ldots, t_i, \ldots, t_{|T_k|}\, \texttt{[SEP]}\, q_1 \ldots q_{|q|}\, \texttt{[SEP]}.$$

We further utilize the embedding of `[CLS]` from the last layer as the representation for the query-table input. Then the relevance $\mathtt{score}(q, T_k)$ is calculated by stacking an Multilayer Perceptron (MLP) scorer with the `[CLS]` representation.

3.2 COTER

We first present an abstract but more convenient form of Optimal Transport (OT), based on the following Wasserstein distance:

Definition 1 (Wasserstein distance [27]). *Let $(\mathcal{X}, d)$ be a Polish metric space where x and y are from, and $m \in [1, \infty]$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of the order m between μ and ν is defined by*

$$W_m(\mu, \nu) = \left(\inf_{p \in \Pi(\mu,\nu)} \int_{\mathcal{X}} \gamma(x, y)^m dp(x, y) \right)^{\frac{1}{m}}, \quad (2)$$
$$s.t. \int_{\mathcal{X}} p(x, y) dy = \mu, \int_{\mathcal{X}} p(x, y) dx = \nu.$$

where $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν.

The Wasserstein distance represents a metrized version of the OT cost, where the traditional distance function $\gamma(x, y)$ is replaced by a more general cost function $c(x, y) \geq 0$ (not necessarily satisfy all metric properties). Among the different choices for m, the most widely used is referred to as the Kantorovich-Rubinstein distance (KR distance), commonly denoted as $W_1(\mu, \nu)$ or simply $W(\mu, \nu)$ (with $m = 1$). Intuitively, the Wasserstein distance, or the total OT cost, measures the expected dissimilarity between two sets by accounting for all possible joint probabilities on the Cartesian product of these two sets while keeping the marginal probabilities fixed. Furthermore, recall the original OT formulation in Eq. (1). If taking probability measures and appropriately normalizing u_i and v_j to ensure they represent proper probabilities, it is straightforward to recognize Eq. (1) as a discrete version of the KR distance.

Given our primary objective is to identify the parsimonious semantic structure of a table based on the given query q, we aim to incorporate the query dependence by transforming it into a conditional probability measure, as depicted in Eq. (2). To maintain simplicity, we introduce a conditional KR distance, denoted as $W(\mu, \nu|q)$, which accounts for the given condition q:

$$W(\mu, \nu|q) = \inf_{p \in \Pi(\mu,\nu|q)} \int_{\mathcal{X}} \gamma(x, y) dp(x, y|q), \quad (3)$$
$$s.t. \int_{\mathcal{X}} p(x, y|q) dy = \mu|q, \int_{\mathcal{X}} p(x, y|q) dx = \nu|q,$$

where $\Pi(\mu, \nu|q)$ represents all couplings conditioned on q, with the marginal probabilities remaining consistent with those defined in Eq. (2). The distinction lies in the substitution of the condition probabilities, *e.g.*, replacing μ (ν) with $\mu|q$ ($\nu|q$). That is, $\Pi(\mu, \nu|q)$ comprises all possible ways of coupling the probability distributions μ and ν while taking into account q, for which we introduce the concept of *Conditional Optimal Transport* (COT). In this context, the condition q influences the coupling process, ensuring that the resulting distributions reflect the relevant semantic structures that align with the provided query. This conditional coupling allows us to tailor our analysis to focus specifically on a "slice" of the entire space, manifested by q, while preserving the overall structure defined by the marginal probabilities. Similarly, the transport quantity p_{ij} from Eq. (1) becomes a function of q, so does the total optimal transport plan $\mathbf{P}^*$. This implies that the optimal transport plan dynamically adapts and changes according to q, providing crucial context information that influences subsequent decisions.

It is also noteworthy that the conditional KR distance exhibits the connection, to some extent, with the original KR distance due to

Theorem 4.6 in [27]. That is, under the certain condition/constraint[2] specified in that theorem, the normalized conditional joint probability indeed represents the optimal transference. However, in this paper, we purposely refrain from imposing such restrictions to Eq. (3). This choice grants us a greater flexibility in considering different optimal transport plans under a given condition, and allows us to explore and leverage the context information from q.

The above establishes the theoretical foundation for the proposed **C**ondi-tional **O**ptimal **T**ransport based table retriev**ER** (COTER). This approach employs conditional optimal transport to streamline input tables, preserving essential items aligned with the query as a prior condition. The core assumption is that the input table can be streamlined by selecting a handful of key items, without substantial loss in semantic significance when matched with the provided query q. Next, we attribute specific meanings to the involved symbols, specifically making appropriate choices (for $\nu|q$, $\mu|q$, and $\gamma(x, y)$) in Eq. (3).

Specifically, let $T = \{t_1, \ldots, t_{|T|}\}$ be the linearized input table represented as a collection of $|T|$ items. These items encompass various components, such as the table caption, headers, cells, and *etc*. Furthermore, let $S \subseteq T$ represent the simplified table. Notably, the number of items in S is either less than or equal to the number of items in T, or $|S| \leq |T|$. Accordingly, the determination of S is to preserve crucial items from T, a task inherently connected with selection problems [9, 10]. The aim is to minimize the loss of semantic meaning while simultaneously simplifying the table and ensuring the alignment with q. Ideally, the items retained in S should capture the vital information within T, enabling effective "transport" of semantic meanings from one items to others. This is precisely the role of conditional optimal transport.

Towards this end, we pursue a binary vector $\boldsymbol{w} \in \{0, 1\}^{|T|}$, *i.e.*, a membership vector of length T where $w_i \in \boldsymbol{w}$ indicates whether the corresponding item t_i in T is selected ($w_i = 1$) or not ($w_i = 0$). Note that $\boldsymbol{w}$ can be interpreted as a realization of a multinomial distribution. This enables the end-to-end training by utilizing the concept of a concrete distribution [17]. In this approach, we introduce a trainable governing probability vector $\boldsymbol{h} \in \mathbb{R}^{|T|}$, which controls a Gumbel softmax distribution sampling that generates $\boldsymbol{w}$. The Gumbel softmax distribution allows for a differentiable approximation of the discrete sampling process, facilitating efficient and effective training.

Next, based on the COT, more precisely, conditional KR distance, we drive the choices for $\nu|q$, $\mu|q$, and $\gamma(x, y)$ in Eq. (3). To start, let $[*]_j$ represent the j-th component of a vector, and f_j^ν be the frequency of the j-th term in the union of the table T (including all its items) and the query q. Accordingly, we have:

$$[\nu|q]_j = \frac{f_j^\nu}{\sum_{j=1}^{|\nu|} f_j^\nu}. \tag{4}$$

Apparently, $\nu|q$ naturally fulfills the requirement of being a probability vector, effectively representing the semantic distribution associated with the given query q. The information of q seamlessly "transfuses" into the table by introducing or augmenting certain term frequencies within the table.

Regarding $\mu|q$ for the simplified table S, only a few selected items are retained. Consequently, the semantic distribution of $\mu|q$ will be supported solely by the remaining terms, and it can be defined in a manner similar to that of Eq. (4)

$$[\mu|q]_j = \frac{f_j^\mu}{\sum_{j=1}^{|\mu|} f_j^\mu}, \tag{5}$$

where f_j^μ is the frequency of the j-th term in S, *i.e.*, retained items from T. The transportation cost $\gamma(x, y)$ is computed based on the semantic dissimilarity between items. Specifically, we measure the dissimilarity as $1 - \frac{\boldsymbol{v}_j^\top \boldsymbol{v}_l}{(|\boldsymbol{v}_j||\boldsymbol{v}_l|)}$, where $\boldsymbol{v}_{j/l}$ represents the representation of the j/l-th item, derived from a pre-trained tokenizer and token embedding model (one single item may contain more than one token, we treat them as a cohesive unit and sum up all belonging token representations to produce one final embedding for that item).

The final component involves the computation of the conditional KR distance $W(\mu, \nu|q)$. Numerous solvers, such as the Sinkhorn approximator [1, 6], are available to perform this computation efficiently. Once the conditional KR distance is computed, the semantic meanings of items in the simplified table, S, are effectively transported to their corresponding counterparts in the original table, T, while considering the condition q.

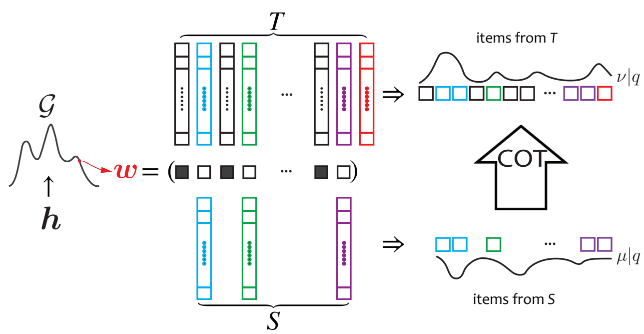


Figure 2: Illustration of the proposed conditional OT for table retrieval. We introduce $\boldsymbol{h}$ as a trainable parameter to control the Gumbel softmax distribution $\mathcal{G}$. $\boldsymbol{w}$ is a realization, *i.e.*, a sampled instance, from $\mathcal{G}$. $\boldsymbol{w}$ controls the selection from rows/cells and the selection quality is measured by the conditional KR distance.

The core of the implementation is to optimize the probabilities $\boldsymbol{h}$ in the Gumbel softmax distribution. That is, informative items from S are identified via the realization of $\boldsymbol{h}$, *i.e.*, $\boldsymbol{w}$, with the minimum conditional KR distance and the best possible semantic coverage. By effectively solving for $\boldsymbol{h}$, the proposed COTER identifies and retains the most meaningful and contextually relevant information from the original table as well as the given query q, thus significantly enhancing the overall retrieval performance and accuracy. Fig. 2 provides a pictorial explanation of the proposed COTER for table simplification based on the condition q, where $\boldsymbol{w}$, sampled from $\mathcal{G}$, yields the value of $W(\mu, \nu|q)$, and the optimization aims to find the optimal $\boldsymbol{h}$ for a minimum $W(\mu, \nu|q)$.

[2] The condition states that the candidate transport plan (positive measure) should not exceed the original optimal transport plan, *i.e.*, the one without this condition. With this constraint in place, the normalized candidate plan will become optimal for the corresponding conditional marginals.

It is crucial to emphasize that our primary objective is to simply tables. Thus, in COTER, the item selection is *solely performed on the table T* and excludes q. In other words, only specific table items are retained, while leaving q's items (*i.e.*, words) unselected. To incorporate semantic information, we thus embed q's words to form the support of ν and subsequently $\nu|q$. This results in the dependence of μ on q, ensuring the integration of query-specific semantic information into table simplification process.

Alternatively, there is also another option to enable the selection of items from q or to include q on S. In this case, the words in q would also appear in μ during the computation of conditional KR distance. Accordingly, the final selection of table items can be achieved by removing q's word(s), retaining only T's item(s). This alternative approach is referred to as `naive conditional OT`. However, it is not as effective as COTER due to the fact that q's words could also participate in the transportation and hence the removal of them could cause information loss. This leads to performance degradation and it is confirmed by our ablation studies presented in a later section.

REMARK 1 (THE SIZE OF S). *Given that $S \subseteq T$, and $|S| \leq |T|$, the process of table simplification is fundamentally framed as a subset selection problem, with the objective of minimizing the total conditional Optimal Transport cost. As a result, the optimal subset size or $|S|$ becomes a significant indicator. This problem is highly non-trivial due to its NP-complete nature, and various solutions exist, ranging from brute force to sparse solvers. A comprehensive comparison in the context of linear regression can be found in [8].*

For simplicity, we propose to utilize a selection budget denoted as $b_s \in (0, 1]$, representing the proportion of items retained from the entire table T. The implementation allows us to discard $\mathbf{w}$'s such that $\sum w_i > b_s|T|$ by assigning a total conditional OT cost of positive infinity. Although automating the determination of the size of S is desirable, we leave this aspect to be explored in future work. In this paper, we present an empirical study, in which we observe sparsity in the solutions, underscoring the effectiveness of our approach.

4 EXPERIMENTS

4.1 Setup

Datasets. Two highly-competitive benchmarks are employed, *i.e.*, **WikiTables** [33] and **WebQueryTable** [21]. These datasets consist of query-table pairs sourced from various domains, such as Wikipedia and commercial Web search engines. The statistics of employed benchmarks are summarized in Table 1, including the numbers/sizes of tables/queries, and their averaged length.

Implementation Details. We compare the proposed COTER with following state-of-the-arts, *i.e.*, `RGTR` [24], `MTR` [19], `TRUKTQ` [34], `GTR` [28], `StruBERT` [25], and `QMTR` [14]. Those methods are reviewed in Section 2, and results are either directly sourced from original papers (if reported) or re-implemented using provided source codes (if not explicitly reported). In addition, the `Base` model is implemented (described in Section 3.1) via simply linearizing input tables and concatenating all items. We also compare with `TabFact` [29] as an linearizition alternative, where the i-th row is flattened as a single sentence as "for row i, h_1 is $c_{i,1}$; $\cdots$; h_m is $c_{i,m}$" with h_j being the head of the j-th column and $c_{i,j}$ the value in the (i, j)-th cell.

Table 1: The statistics of adopted table retrieval benchmarks.

Dataset	WikiTables	WebQueryTable
# of tables	2,879	273,816
Avg # of columns	4.69	4.55
Max # of columns	40	52
Min # of columns	1	1
Avg # of rows	14.63	9.15
Max # of rows	486	1,517
Min # of rows	1	2
# of questions	60	21,113
Avg len. of questions	2.72	4.61

Similar to all baselines, BERT(base) is adopted as the primary encoder for COTER. The dropout rate is set as 0.1 for each layer. An Adam optimizer is adopted with weight decay and an initial learning rate of $1e^{-4}$. For WikiTables, we use batch size of 16 and warmup steps of 100. On WebQueryTable, we use batch size of 4 and warmup steps of 1000. The maximal number of training iteration is set to 200. Additionally, for COTER, tables are linearized row by row and cell by cell, with the "$\cdot$" symbol to separate rows or cells. Linearized tables and queries are concatenated and then fed to the encoder, with `[SEP]` as the delimiter to separate tables and queries. Then the representation of `[CLS]` from the final layer is utilized as the hidden state for the table-query matching. At last, the proposed model is trained using a machine of the NVIDIA A100 GPU server.

Evaluation Metrics. Following previous works [14, 25, 28], the Mean Average Precision (**MAP**) and Normalized Discounted Cumulative Gain (**NDCG@**) with cut-off points {5, 10, 20} are employed to evaluate models; accordingly, higher values indicates the better model performance.

4.2 Main results

The proposed COTER is applied on both the row and cell level, cast as **COTER-R** and **COTER-C** respectively, to select the most informative rows and cells (we decided not to include columns in our analysis since they can be regarded as a special case of cells). Additionally, the selection budget (for the fraction of retained rows/cells) is 60% and 40% for COTER-R and COTER-C, respectively, while the impact of selection percentage is investigated later in the ablation study. The experiment is run for five trials (with randomly initialized seeds) and averaged results are shown in Table 2.

The comparison results show that the proposed COTER (from both row and cell perspectives) consistently outperforms state-of-the-arts across employed benchmarks. For instance, with the WikiTables dataset, the strongest baseline, `QMTR`, achieves an absolute-point improvement of 0.058, 0.060, 0.057, and 0.053, with regarding to NDCG@5, 10, 20, and MAP, respectively, compared to `Base`, while COTER-R achieves a further 0.030, 0.028, 0.015, and 0.017 absolute point over `QMTR`. Empirically, it amounts to a comparable improvement and demonstrates the superiority of the proposed approach. We also notice that COTER-C achieves even better performance than that of COTER-R. The main reason is that COTER-C targets on individual cells, compared to COTER-R focusing on the entire row,

Table 2: Performance comparison between the proposed COTER with previous best reported results. Statistically significant gains achieved by the proposed method at p-values < 0.01 are marked with †.

WikiTables	NDCG@5	NDCG@10	NDCG@20	MAP
Base	0.6183	0.6213	0.6764	0.6358
TabFact[29]	0.6127	0.6151	0.6778	0.6309
RGTR[24]	0.6246	0.6244	0.6404	0.6242
MTR [19]	0.6631	0.6813	0.7370	0.6058
TRUKTQ[34]	0.6172	0.6267	0.6644	0.6273
GTR[28]	0.6671	0.6856	0.7272	0.6859
StruBERT[25]	0.6393	0.6453	0.6844	0.6378
QMTR[14]	0.6763	0.6810	0.7330	0.6886
COTER-R	0.7058†	0.7091†	0.7478†	0.7051†
COTER-C	0.7127†	0.7150†	0.7564†	0.7157†

WebQueryTable	NDCG@5	NDCG@10	NDCG@20	MAP
Base	0.7387	0.7528	0.7585	0.7104
TabFact[29]	0.7367	0.7475	0.7601	0.7097
RGTR[24]	0.6438	0.6610	0.6691	0.6200
MTR [19]	0.7624	0.7817	0.7993	0.6694
TRUKTQ[34]	0.7096	0.7190	0.7278	0.6824
GTR[28]	0.7670	0.7866	0.7963	0.7457
StruBERT[25]	0.7351	0.7404	0.7494	0.6934
QMTR[14]	0.7776	0.7813	0.8026	0.7486
COTER-R	0.8179†	0.8333†	0.8400†	0.7917†
COTER-C	0.8249†	0.8390†	0.8447†	0.7966†

thereby offering more flexibility. In addition, the significance test (*i.e.*, the one-sample T-test) is also considered. Specifically, p-values of our MAP results being greater than relevant strongest baseline (QMTR), from WikiTables and WebQueryTable, are 3.2×10^{-6} and 8.7×10^{-7}, respectively, which verifies the effectiveness and stability of COTER.

In regard to computational complexity, the proposed algorithm presents a highly competitive training process. As an illustration, we consider the two top-performing baseline models from Table 2, namely GTR and QMTR. Let n be the number of items from the linearized table (including headers and cells), and d is the hidden dimension (d=768 for BERT-base specifically). GTR relies on constructing a hypergraph with all items as nodes and linking adjacent items (as edges while the number of edges $e \approx 2 \times n$), so its time complexity is approximately $O((n^2 + 2 \times n) \times d)$. QMTR extends the original table via retrieving relevant items (with the length of $\widehat{n}$ and $\widehat{n} > n$) from a large corpus, while extended information also needs to be encoded for verification purposes. Accordingly, its time complexity is $O((\widehat{n}^2 + n^2) \times d)$. By contrast, COTER only retains the informative items (with the length of $\bar{n}$), so its time complexity is $O(2 \times n^2 \times \log(n)) + O(\bar{n}^2 \times d)$ where $\bar{n} < n$ and the first term is the computation of the Wasserstein distance [1]. As such, the proposed method is computationally affordable.

4.3 Ablation study

To better understand the effectiveness of the proposed method, a series of careful studies are carried out. The following experiments are considered using WikiTables, and results are again reported as the averaged accuracy over five runs.

On the selection budget (b_s). This experiment is to evaluate the impact of the selection budget (b_s) on COTER. Obviously, with a higher value of b_s, more rows/cells will be retained for the subsequent retreival. Specifically, experiments are conduced by varying b_s within the range of 20%, 40%, 60%, and 80%. Notably, with b_s=100%, COTER is equivalent to the `Base` method with the entire table.

Table 3 presents a comparison between different selection budgets (b_s) and their impact on model performance. Surprisingly, even with a modest selection budget of only b_s=20%, both COTER-R/C

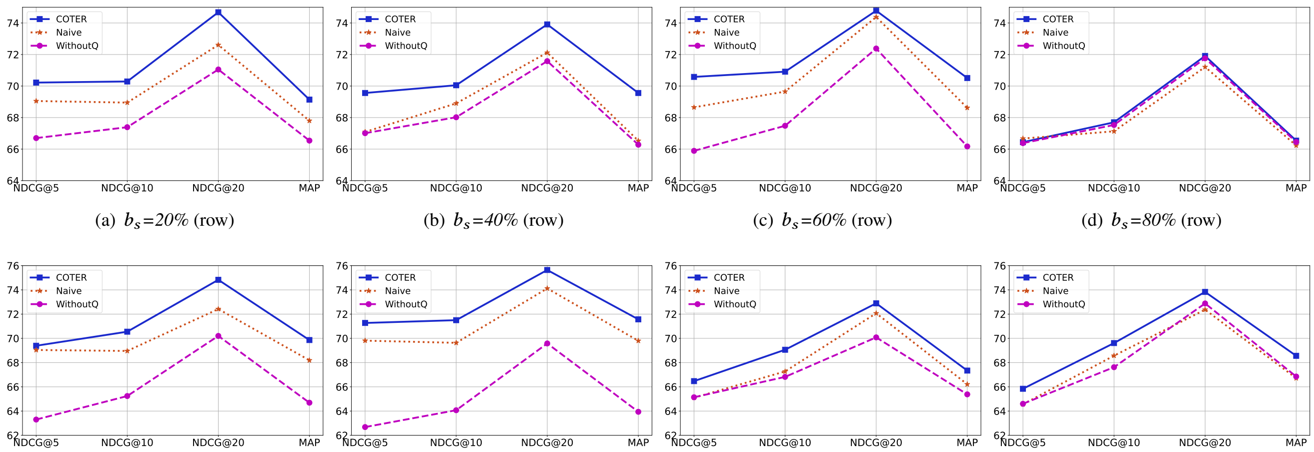
Table 3: Performance comparison as a function of the selection budgets (b_s). As b_s increases, the retrieval performance demonstrates a decline, primarily attributed to the inclusion of less-important item.

COTER-R	NDCG@5	NDCG@10	NDCG@20	MAP
20%	0.7022	0.7029	0.7468	0.6914
40%	0.6956	0.7005	0.7391	0.6956
60%	0.7058	0.7091	0.7478	0.7051
80%	0.6644	0.6770	0.7191	0.6654
COTER-C	NDCG@5	NDCG@10	NDCG@20	MAP
20%	0.6939	0.7055	0.7483	0.6986
40%	0.7127	0.7150	0.7564	0.7157
60%	0.6647	0.6906	0.7289	0.6734
80%	0.6584	0.6961	0.7383	0.6856

outperform the strongest baseline (QMTR from Table 2) in terms of retrieval.

On the other hand, by increasing the selection budget (*i.e.*, 20% → 60% for COTER-R and 20% → 40% for COTER-C), the performance of both models can be further enhanced as more relevant table items are incorporated. However, it is worth noting that COTER-R reaches a plateau at around b_s=60%, beyond which it drastically declines. Similarly, COTER-C exhibits the same trend with a plateau at b_s=40%. Not surprisingly, with a selection budget of b_s=80%, both variants yield the worst results. The comparison reveals clearly that increasing the selection budget does not necessarily lead to improved performance. The presence of redundant information, which may be attributed to the existence of similar contents/items across multiple tables, will reduce the retrieval performance. This finding emphasizes the significance of employing a conditional OT-based method to identify the most relevant items for subsequent retrieval, as opposed to existing methods treating the entire table as a singular entity.

On the OT strategy. The following ablation is performed to evaluate the effectiveness of the proposed conditional OT strategy. As previously mentioned, COTER only retains significant table items via taking into account the given query. Yet, there are another two alternatives for comparison purposes:



(a) b_s=20% (row) (b) b_s=40% (row) (c) b_s=60% (row) (d) b_s=80% (row)

(e) b_s=20% (cell) (f) b_s=40% (cell) (g) b_s=60% (cell) (h) b_s=80% (cell)

Figure 3: Comparison of COTER with two OT variants (in terms of the retrieval MAP), where the variety of selection budgets are considered.

- `Naive` conditional OT: this represents the selection process that is conducted on both the table and query. That is, words from q will be selected and contribute to the computation of conditional KR distance in μ.
- `WithoutQ` OT: this only considers retaining table items, regardless the given query. Apparently this approach is no use of the query as the condition at all.

Results from Fig. 3 evidently state the effectiveness of the conditional OT. For instance, COTER clearly outperforms the other two variants across all cases, demonstrating the superiority of taking the query as the condition. In particular, when the selection budget is constrained, such as at b_s=20% and 40%, COTER consistently demonstrates significantly superior scores (in relation to MAP), compared to `Naive` and `WithoutQ`-based OT methods. At a higher budget, say b_s=80%, COTER and the other two variants exhibit similar performance, indicating the lower bound of the proposed method.

On the other hand, the `Naive` strategy performance worse than the proposed COTER. The reason is q's words becomes support for μ, and the removal of them brings information leakage such that the transport plan is only sub-optimal. Additionally, the `WithoutQ` strategy yields the worst results. This can be attributed to its reliance on a large searching space without the condition q, making it more likely for the model to get trapped in solutions irrelevant to q.

The findings highlight the importance of striking the right balance in utilizing the query. Using too few (the `WithoutQ` strategy) or too many (the `Naive` strategy) query words can lead to the performance degradation. On the contrary, COTER optimizes its approach through meticulous control over the query's influence on the selection process, guaranteeing the achievement of optimal outcomes. This also emphasizes the significance of properly integrating query information to achieve the table simplification.

Comparison with existing selection strategies. In this experiment, we evaluate COTER's ability via adapting other selection methods. In particular, for comparison purposes, we implement `Rand-R` and `Rand-C`, which utilize a random selection of rows and cells, respectively. `TSDM-R` and `TSDM-C` [4] is the approach of choosing items using the word-embedding similarity (among the query and table words). `Clu` [22] clusters cells from flattened tables (treated as documents) and then selects words that are close to the cluster centroids.

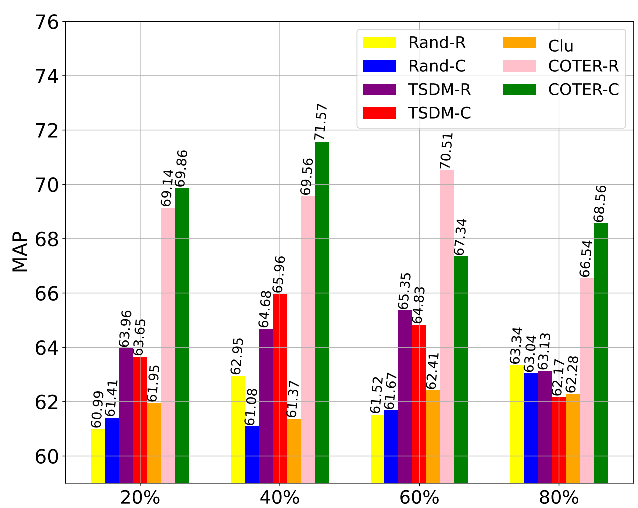


Figure 4: Comparison of COTER against existing selection based methods (in terms of MAP) to determine the most informative rows and cells.

The MAP comparison from Fig. 4 provides compelling evidence for the superiority of COTER, in identifying the most informative table items (rows and cells), as a function of the selection budget. Notably, both `Rand-R/C` and `Clu` do not reply on queries to identify/filter key items; accordingly they demonstrate relatively stable performance. For instance, `Clu` yields results ranging from approximately 61.95% to 62.28%, irrespective of the selection budget (b_s). However, `TSDM`, which incorporates query-based selection, exhibits performance variation across different b_s and outperforms `Rand-R/C` and `Clu`.

In contrast, COTER consistently outperforms other methods in all scenarios and achieves remarkable MAP improvement of 3.7%-4.3%. With the increase of b_s, both `TSDM` and COTER are observed with declined performance, which might be attributed to the inclusion of irrelevant items. Yet, the proposed method still maintains the best retrieval performance. Again, the comparison underscores the efficacy of the proposed approach in determining relevant and informative table items with varying selection budgets.

Qualitative study. We further investigate COTER via visualizing selected table items. Specifically, Fig. 5 compares retained items obtained from COTER and TSDM, in relation to the query "countries capital" (also shown in Fig. 1). Both methods are capable of identifying the most relevant items, namely the table headers containing the keywords "country" and "capital" (located in the first and last cells of the first row).

Country or territory	List of countries and dependencies by population Population.	List of sovereign states and dependent territories by population density Population density.	Languages (official in bold)	Capital
Anguilla	15000	164.8	Demographics of Anguilla English	The Valley, Anguilla The Valley
Canada	33573000	3.4	Languages of Canada English	Ottawa
Mexico	112322757	57.1	Languages of Mexico Spanish	Mexico_City
Aruba Aruba	107000	594.4	Languages of Aruba Papiamentu	Oranjestad, ruba Oranjestad
United States	311630000	32.7	Languages of the United States English	Washington, D.C

Figure 5: Comparison of COTER and TSDM for table item filtering, in response to the query of "countries capital". COTER based items are highlighted in yellow colour, while TSDM based ones are underlined in blue colour. Those with purple are identified by both two methods.

However, even though TSDM successfully captures those primary items, it also selects certain irrelevant items, including cells with the keyword of "languages" and even incorporating two numerical values. In contrast, COTER excels in performance by pinpointing items with higher semantic relevance (*i.e.*, cells closely tied to countries and capitals), effectively excluding insignificant items such as those numerical. This once more highlights the advantage of leveraging COTER for determining contextually relevant items to the given query.

5 CONCLUSION

We introduced a novel table retrieval method, termed **C**onditional **O**ptimal **T**ransport based table retriev**ER** (COTER), that achieves a remarkable advancement surpassing previous state-of-the-art models. The proposed method simplifies tables by leveraging the concept of conditional optimal transport, with the given query acting as the conditional factor. Specifically, COTER minimizes semantic loss during table simplification by effectively "transporting" the semantic meanings of one or several words (from the original table) to individual words (from the simplified table). Moreover, the provided query plays a crucial role as a prior condition, ensuring that retained table items align effectively with the query. This allows COTER to dynamically adjust chosen items according to the provided query.

Our extensive experimental results provide compelling evidence of the superiority of our proposed algorithm over existing methods. In future work, we plan to enrich COTER by incorporating layout information derived from tables (such as the structural correlation inherent in table items). Additionally, the proposed conditional optimal transport method remains agnostic to downstream tasks, *i.e.*, we could seamlessly incorporate it into other applications.

ETHICAL STATEMENT

Table retrieval represents a fundamental task in the field of Web search and Natural Language Processing, and it is generally considered ethically unproblematic. Furthermore, to ensure transparency and reproducibility, all datasets employed in this study are openly accessible to the public. Nevertheless, we acknowledge the need to address the possibility of subtle biases that might arise from the use of Pre-trained Language Models (PLMs) as encoders for generating the latent representation of tables and queries. These PLMs might inherit biases present in the data they were trained on. However, it is important to note that during our analysis, we did not identify any concerning outcomes related to bias.

REFERENCES

[1] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. 2017. Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., 1961–1971.

[2] Ebrahim Bagheri and Feras Al-Obeidat. 2020. A Latent Model for Ad Hoc Table Retrieval. In *Advances in Information Retrieval*. Springer International Publishing, 86–93.

[3] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1026–1036.

[4] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D Davison. 2020. Table search using a deep contextualized language model. (2020), 589–598.

[5] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Dawei Yin, and Brian D. Davison. 2021. MGNETS: Multi-Graph Neural Networks for Table Search. (2021), 2945–2949.

[6] Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., 2292–2300.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.

[8] Yi Guo and Mark Berman. 2012. A comparison between subset selection and L1 regularisation with an application in spectroscopy. *Chemometrics and Intelligent Laboratory Systems* 118 (Aug. 2012), 127–138.

[9] Yi Guo, Mark Berman, and Junbin Gao. 2014. Group Subset Selection for Linear Regression. *Comput. Stat. Data Anal.* 75 (jul 2014), 39–52.

[10] Yi Guo, Junbin Gao, and Feng Li. 2013. Dimensionality Reduction with Dimension Selection. In *Advances in Knowledge Discovery and Data Mining*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 508–519.

[11] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4320–4333.

[12] Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving Cross-Lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. Association for Computing Machinery, 1048–1056.

[13] Madelon Hulsebos, Xiang Deng, Huan Sun, and Paolo Papotti. 2023. Models and Practice of Neural Table Representations. In *Companion of the 2023 International*

Conference on Management of Data (SIGMOD '23). Association for Computing Machinery, 83–89.
[14] Maryam Khodabakhsh and Ebrahim Bagheri. 2022. Qualitative Measures for Ad Hoc Table Retrieval. *Inf. Sci.* 607, C (aug 2022), 1–26.
[15] Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 5969–5979.
[16] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *International Conference on Learning Representations*.
[17] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*.
[18] Nasim Nouri. 2022. Text Style Transfer via Optimal Transport. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2532–2541.
[19] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Cannim. 2020. Web Table Retrieval Using Multimodal Deep Learning. (2020), 1399–1408.
[20] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards Table-to-Text Generation with Numerical Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 1451–1465.
[21] Yibo Sun, Zhao Yan, Duyu Tang, Nan Duan, and Bing Qin. 2019. Content-based table retrieval for web queries. *Neurocomputing* 349 (2019), 183–189.
[22] Hongyin Tang, Xingwu Sun, Beihong Jin, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Improving Document Representations by Generating Pseudo Query Embeddings for Dense Retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5054–5064.
[23] Mohamed Trabelsi, Zhiyu Chen, Brian D. Davison, and Jeff Heflin. 2020. A Hybrid Deep Model for Learning to Rank Data Tables. In *2020 IEEE International Conference on Big Data (Big Data)*. 979–986.
[24] Mohamed Trabelsi, Zhiyu Chen, Brian D. Davison, and Jeff Heflin. 2020. Relational Graph Embeddings for Table Retrieval. In *2020 IEEE International Conference on Big Data (Big Data)*. 3005–3014.
[25] Mohamed Trabelsi, Zhiyu Chen, Shuo Zhang, Brian D. Davison, and Jeff Heflin. 2022. StruBERT: Structure-Aware BERT for Table Search and Matching. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, 442–451.
[26] Mohamed Trabelsi, Brian D. Davison, and Jeff Heflin. 2019. Improved Table Retrieval Using Multiple Context Embeddings for Attributes. In *2019 IEEE International Conference on Big Data (Big Data)*. 1238–1244.
[27] Cédric. Villani. 2009. *Optimal Transport Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg.
[28] Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. 2021. Retrieving Complex Tables with Multi-Granular Graph Representation Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1472–1482.
[29] Jianshu Chen Yunkai Zhang Hong Wang Shiyang Li Xiyou Zhou Wenhu Chen, Hongmin Wang and William Yang Wang. 2020. TabFact : A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia.
[30] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8413–8426.
[31] Hai-Tao Yu, Adam Jatowt, Hideo Joho, Joemon M. Jose, Xiao Yang, and Long Chen. 2019. WassRank: Listwise Document Ranking Using Optimal Transport Theory. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, 24–32.
[32] Li Zhang, Shuo Zhang, and Krisztian Balog. 2019. Table2Vec: Neural Word and Entity Embeddings for Table Population and Retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1029–1032.
[33] Shuo Zhang and Krisztian Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 World Wide Web Conference*. 1553–1562.
[34] Shuo Zhang and Krisztian Balog. 2021. Semantic Table Retrieval using Keyword and Table Queries. *ACM Trans. Web* 15 (2021), 11:1–11:33.
[35] Yichao Zhou, Ying Sheng, Nguyen Vo, Nick Edmonds, and Sandeep Tata. 2022. Learning Transferable Node Representations for Attribute Extraction from Web Documents. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)*. Association for Computing Machinery, 1479–1487.