A Latent Model for Ad Hoc Table Retrieval

Ebrahim Bagheri[1(✉)] and Feras Al-Obeidat[2]

[1] Laboratory for Systems, Software and Semantics (LS3),
Ryerson University, Toronto, Canada
bagheri@ryerson.ca
[2] College of Technological Innovation,
Zayed University, Abu Dhabi, United Arab Emirates
Feras.Al-Obeidat@zu.ac.ae

Abstract. The ad hoc table retrieval task is concerned with satisfying a query with a ranked list of tables. While there are strong baselines in the literature that exploit learning to rank and semantic matching techniques, there are still a set of *hard queries* that are difficult for these baseline methods to address. We find that such hard queries are those whose constituting tokens (i.e., terms or entities) are not fully or partially observed in the relevant tables. We focus on proposing a latent factor model to address such hard queries. Our proposed model factorizes the token-table co-occurrence matrix into two low dimensional latent factor matrices that can be used for measuring table and query similarity even if no shared tokens exist between them. We find that the variation of our proposed model that considers keywords provides statistically significant improvement over three strong baselines in terms of NDCG and ERR.

1 Introduction

Tables provide a structured representation of data that can be quickly interpreted by the users. There have been important work that focus on the automated retrieval and interpretation of data in tabular format. These works range from mining tables from documents [3,8] to extracting information from within tables [11] and semantic analysis of table content [2,14], to name a few. Recently, Zhang and Balog have systematically introduced the task of *ad hoc table retrieval* from an information retrieval perspective [16]. The idea is to retrieve a ranked list of relevant tables for an input query given a corpus of existing tables. The task shows close resemblance to the traditional ad hoc document retrieval problem while distinguishing itself in that data in tables are quite short, often including a few terms or just numbers, and as such making it challenging to extract the context that is needed to draw *relevance* conclusions.

Methods for ad hoc table retrieval show strong performance on $nDCG@20$ metric where supervised methods based on the learning to rank approach report between 0.5206 [1] to 0.6031 [16] while semantic relevance methods report up to 0.6825 on a corpus with 1.6M tables. Despite this strong performance, we note

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 86–93, 2020.
https://doi.org/10.1007/978-3-030-45442-5_11

that their performance is not satisfactory over *hard* (difficult) queries. In other words, while these methods perform very well on a subset of queries, they are not equally effective on another subset. This is inline with findings within the ad hoc document retrieval literature that distinguishes the performance of a retrieval method on soft (easy) and hard (difficult) queries [7,15]. More specifically, when looking at the bottom 20% of the queries ranked based on AP, it is possible to see that in the top-10 results retrieved per query, there are 0 out of 12, 2 out of 12 and 7 out of 12 queries that had retrieved at least one relevant table by WikiTable [1], Learn To Rank (LTR) and Semantic Table Retrieval (STR) [16] methods, respectively. We observe that methods that are primarily based on keyword-based features for determining relevance do not perform well on hard queries. For instance, as we will show with more details later, for a query such as *'pain medication'*, the most relevant table based on relevance judgements is one that does not include any of the query terms. Therefore, methods such as STR [16] that leverage semantics perform better on hard queries.

However, while STR shows improved performance by considering semantic information, the employed semantics are derived from word and graph embeddings learnt on generic corpora such as Google News and DBpedia. In this paper, we will present a systematic approach for learning *low dimensional latent factor matrices* to represent queries and tables based on the co-occurrence of terms and entities within the table corpus. The learnt latent factor matrices allow us to efficiently compute query-table similarities for ranking. We show that the learnt latent representation allow us to, statistically speaking, significantly improve the performance of ad hoc table retrieval on *hard queries*, which would otherwise not receive appropriate treatment. Our method is specially suited for hard queries, as the learnt latent representations extract transitive relations through observed entity and term co-occurrences, which are appropriate for measuring relevance when query terms are not present in relevant tables.

2 Proposed Approach

The objective of our work is to learn low dimensional latent factor matrices to represent tables and queries, which can then be used to measure query-table relevance. We position our work within the context of factored item-item collaborative filtering [12] where item-item similarity is learnt as the product of two matrices $\mathbf{P}$ and $\mathbf{Q}$. Hence, similarity between items i and j can be simply computed as $\mathbf{p}_i \cdot \mathbf{q}_i$. Let us assume that the set of tokens observed in the table corpus is denoted as $\mathcal{V}$ and the set of observed tokens in Table t is $\mathcal{V}_t^o$. We denote the set of tokens not observed in t as $\mathcal{V}_t^u = \mathcal{V} - \mathcal{V}_t^o$. Let us assume that $\mathbf{P}$ and $\mathbf{Q}$ are already computed; on this basis, it is possible to estimate the relevance (R) of a token i to a table t as:

$$\hat{R}_{t,i} = b_t + b_i + \sum_{j \in \mathcal{V}_t^o} \mathbf{p}_j \cdot \mathbf{q}_i^\intercal \tag{1}$$

Here, b_t and b_i are table and token biases. Now, the objective will be to efficiently learn matrices $\mathbf{P}$ and $\mathbf{Q}$ through a regularized optimization problem. Since our

problem focuses on the effective ranking of tables, we are interested in minimizing *ranking error*, and so, we adopt the ranking-based loss function proposed in [13], which minimizes overall rank loss:

$$\sum_{t \in T} \sum_{i \in \mathcal{V}_t^o, j \in \mathcal{V}_t^u} \Big((R_{t,i} - R_{t,j}) - (\hat{R}_{t,i} - \hat{R}_{t,j}) \Big)^2 . \tag{2}$$

where T is the set of tables in the table corpus, $R_{t,i}$ is the relevance of token i to table t and $\hat{R}_{t,i}$ is the predicted estimation for $R_{t,i}$.

With this loss function, we learn $\mathbf{P}$ and $\mathbf{Q}$ by minimizing a regularized optimization function that considers three factors within its optimization function:

$$\begin{aligned} minimize_{\mathbf{P},\mathbf{Q}} & \underbrace{\frac{1}{2} \sum_{t \in T} \sum_{i \in \mathcal{V}_t^o, j \in \mathcal{V}_t^u} \| (R_{t,i} - R_{t,j}) - (\hat{R}_{t,i} - \hat{R}_{t,j}) \|_F^2}_{\text{loss function}} \\ & + \underbrace{\frac{\beta}{2}(\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2) + \frac{\beta}{2}(\|b_i\|_2^2)}_{\text{regularization terms}} \\ & + \sum_{t_i, t_j \in T} \underbrace{\frac{\gamma}{2} \|\mathbf{p}_i \cdot \mathbf{q}_j - Sim(t_i, t_j)\|_F^2}_{\text{regularizing with baseline similarity}} . \end{aligned} \tag{3}$$

The first part minimizes the loss function, which is based on the Bayesian Parameterized Ranking loss function [13]. The second part includes regularizers based on norms of the two matrices and the biases for tokens. The third part is included to ensure that additional feature-based similarity information are considered when optimizing $\mathbf{P}$ and $\mathbf{Q}$. Here the estimated similarity between two tables t_i and t_j, which is to be estimated based on $\mathbf{p}_i \cdot \mathbf{q}_j$, is compared to the value of a similarity function, $sim(t_i, t_j)$. We consider $sim(t_i, t_j)$ to be a similarity function derived from the baseline methods. This minimization function can be optimized using Stochastic Gradient Descent.

As indicated earlier, the tokens in a table can either be terms or entities. The tables in the corpus [16] are from Wikipedia and hence have links to other Wikipedia pages. We use these links as a representation of entities and the terms used in the table as representation of terms. Hence, we propose three variations, namely (1) Keyword, (2) Entity, and (3) Keyword + Entity. Within the Keyword variation, the relevance of token i to table t, i.e., $R_{t,i}$, is defined as: $R_{t,i} = 1$ if $i \in t$ and 0 otherwise. In the Entity model, $R_{t,i}$ is defined slightly differently because it is possible to determine the relevance of an entity to a table even if the entity is not observed in the table. We define $R_{t,i}$ for the Entity variation as $R_{t,i} = 1$ if $i \in t$ and $wmd(i,t)$ otherwise. In this case, if the entity is not present in the table, we compute the word mover's distance [9] to compute the similarity of entity i to table t. The Keyword + Entity variation is a combination of these two relevance functions.

The intuition for our approach is that tables are modelled as a collection of tokens (i.e., terms or entities), which can be distributed across multiple tables.

Table 1. Comparison with STR. † indicates statistical sig with paired t-test at 0.05.

	STR	Entity	Keyword	Keyword + Entity
NDCG@10	0.1603	0.1541	**0.1782**	0.1491
Δ		*−3.88%*	*+11.14%*†	*−7.02%*
NDCG@20	0.1919	0.1928	**0.2343**	0.1758
Δ		*+0.44%*	*+22.09%*†	*−8.39%*
ERR@10	0.1507	0.148	**0.187**	0.1522
Δ		*−1.76%*	*+24.14%*†	*+1.01%*
ERR@20	0.1757	0.1862	**0.2474**	0.1547
Δ		*+6.00%*	*+40.83%*†	*−11.95%*

Table 2. Comparison with LTR. † indicates statistical sig with paired t-test at 0.05.

	LTR	Entity	Keyword	Keyword + Entity
NDCG@10	0.0197	0.0466	**0.0872**	0.0417
Δ		*+136.90%*†	*+343.31%*†	*+111.99%*†
NDCG@20	0.0894	0.0952	**0.1195**	0.097
Δ		+6.49%	+33.66%	+8.50%†
ERR@10	0.0486	0.0569	**0.0965**	0.066
Δ		*+17.08*†	*+98.56%*†	*+35.80%*†
ERR@20	0.0710	0.0723	**0.1117**	0.0835
Δ		*+1.83%*	*+57.32%*	*+17.61%*†

The occurrence of tokens in multiple tables delivers indirect semantics on the relationship between the different tables. The two derived low dimensional latent factor matrices capture the semantics and allow us to compute the similarity between any two tables. For ranking tables for a query, we model the query similar to tables and use the same similarity function as a score of relevance.

3 Experimental Setup

Corpora: We used the corpus from [16] that includes 1.6M tables from Wikipedia.
Topics: We employ the 60 topics that accompany the table corpus to perform our experiments. We annotate the queries using TagMe as done in [4–6].
Baselines: The state of the art include the methods in [16]. Zhang and Balog report that WikiTables [2] is also a strong baseline. We define hard queries for each baseline as the bottom 20% of queries based on Average Precision (AP).
Neural Embeddings: The neural embeddings used for $R_{t,i}$ in Eq. 5 were based on the Hierarchical Category Embedding (HCE) model proposed in [10].
Metrics: Retrieval effectiveness was evaluated with NDCG and ERR.

4 Findings

Given the set of hard queries is different for each baseline, we report the performance improvements obtained through our proposed model separately for each baseline. Furthermore, given the $Sim(.,.)$ function employed in the third regularization term of Eq. 3 is dependent on the baseline, it is necessary to report the findings separately for each baseline. The results are reported in Tables 1, 2 and 3 for STR, LTR and WikiTable methods, respectively.

Table 3. Comparison with WikiTables. † indicates statistical sig with paired t-test at 0.05. * not possible to calculate delta improvement given divide by zero.

	WT	Entity	Keyword	Keyword + Entity
NDCG@10	0	0.0684	**0.0921**	0.0885
Δ		*	*†	*†
NDCG@20	0.09074	0.1053	**0.1195**	0.1107
Δ		*+16.06%*†	*+31.69%*†	*+21.99%*†
ERR@10	0	0.0521	0.1287	**0.1334**
Δ		*	*†	*†
ERR@20	0.0616	0.0749	**0.147**	0.1414
Δ		*+21.55%*†	*+138.68%*†	*+129.58%*†

When considering the three baselines and based on the three variations of our proposed approach, it is possible to see that the Keyword variation shows stronger performance compared to both Entity and Keyword+Entity approaches. The reason for the lower performance of the Entity-based variations was related to the sparsity of entities in tables compared to terms. It is more difficult for the latent factor model to identify table similarities based on sparse entity occurrence information. Furthermore, we found that the use of a semantic similarity score in Eq. 5 for cases when the entity was not explicitly observed in the table leads to undesirable derived table similarities. Hence, the performance of the variations that included entity information in our proposed approach is weaker than the variation that only uses terms.

We further observe that regardless of the variation, our approach improves over the baselines except for STR, which was only significantly outperformed using the Keyword variation. We postulate that this is due to the fact that STR already benefits from entity information in their semantic formulation and as such the use of sparse entity information in our model does not lead to observable improvements, while our variation based on Keywords outperforms STR. The other noticeable improvement is observed over ERR and NDCG at 10 for WikiTable. As seen in Table 3, there are no relevant retrievals at rank 10 by this baseline and hence both ERR@10 and NDCG@10 are reporting zero. However, all three variations of our approach have been able to improve WikiTable by identifying relevant tables for the hard queries at rank 10.

Table 4. Top-2 queries with most improvement by the *Keyword* model over baselines. †denotes tables not containing query terms, ‡indicates partial query presence in table.

STR				
Query 7: Prime Ministers of England				
Table ID	Table Caption	Rel	Base	Ours
0406-281	Labour prime ministers‡	2	7	3
Query 51: Cereals nutrition value				
1573-730	Sesame seed kernels, toasted‡	2	9	2
LTR				
Query 7: Prime ministers of England				
0406-281	Labour prime ministers‡	2	17	7
Query 57: Board games number of players				
1098-540	List of Japanese board games‡	1	13	3
WikiTable				
Query 30: Pain medication				
1444-126	Threshold of pain‡	1	17	1
0520-188	Diseases and conditions†	1	12	2
Query 59: Constellations closest constellation				
0177-367	Deep space rendezvous†	1	17	3
1437-680	Constellations	1	20	6
1264-76	Symbols for zodiac constellations‡	1	16	9
1113-680	Solar encounter‡	1	19	16

Now, in order to analyze the performance of our approach in contrast with the baselines, we select two queries with the highest improvement over NDCG@10 for each baseline and report them in Table 4. The third column of the table shows the relevance score given to the table for the query in the relevance judgements, the fourth and fifth columns are the table rank produced by the baseline and the Keyword variation of our approach, respectively. For all the relevant tables for each query, our approach has been able to improve the ranking of the relevant table in these queries. We further looked into the characteristics of these tables for possible explanation of the better performance of our approach. We classify the tables into two types, denoted by † and ‡, which represent those tables that either have only a subset of the query tokens mentioned in them, or none of the query tokens mentioned in them, respectively. As seen in Table 4, all tables, except for Table 1437–680, are classified as either † or ‡. Given our latent factor model is able to identify implicit relations between tables based on transitively shared tokens, it is able to identify query and table relevance even if the same tokens do not appear in them. For instance, for Query 30: 'Pain Medication', our approach has been able to identify Table 0520-188: 'Diseases and conditions'

despite the fact that none of the tokens in the query appear in the table. A similar pattern can be observed in the other examples as well.

We observe that hard queries for the baselines are those queries which have relevant tables that are classified as † or ‡. However, soft queries are those whose tokens are explicitly observed in the relevant tables. We believe that for such soft queries employing baseline methods that check for explicit query term occurrence would be a better strategy compared to our latent factor model that derive relevance based on transitive token-table occurrence patterns.

5 Concluding Remarks

In summary, we find that:

1. The Keyword variation of the proposed latent factor model is able to improve the performance of all the baseline in a statistically significant way.
2. While the variations that consider Entity information are able to also improve the performance of the baselines, they fail to do so when tested against the STR model, which already incorporates the semantics of entity information in its retrieval model.
3. Our model is suitable for satisfying those queries whose tokens do not appear or only partially appear in the relevant tables. This is due to the factorization that captures the implicit relationship between tables and queries.

References

1. Bhagavatula, C.S., Noraset, T., Downey, D.: Methods for exploring and mining tables on wikipedia. In: Interactive Data Exploration and Analytics, IDEA 2013, pp. 18–26 (2013)
2. Bhagavatula, C.S., Noraset, T., Downey, D. Tabel: entity linking in web tables. In: International Semantic Web Conference, pp. 425–441 (2015)
3. Clark, C.A., Divvala, S.: Looking beyond text: extracting figures, tables and captions from computer science papers. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
4. Dargahi Nobari, A., Askari, A., Hasibi, F., Neshati, M.: Query understanding via entity attribute identification. In: The 27th ACM International Conference on Information and Knowledge Management, pp. 1759–1762 (2018)
5. Ensan, F., Bagheri, E.: Document retrieval model through semantic linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, pp. 181–190 (2017)
6. Hasibi, F., Balog, K., Bratsberg, S.E.: Dynamic factual summaries for entity cards. In: The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017, pp. 773–782 (2017)
7. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: The 17th ACM Conference on Information and Knowledge Management, pp. 1419–1420 (2008)
8. Khusro, S., Latif, A., Ullah, I.: On methods and tools of table detection, extraction and annotation in pdf documents. J. Inform. Sci. **41**(1), 41–57 (2015)

9. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning, pp. 957–966 (2015)
10. Li, Y., Zheng, R., Tian, T., Hu, Z., Iyer, R., Sycara, K.P.: Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Osaka, Japan, 11–16 December 2016, pp. 2678–2688 (2016)
11. Muñoz, E., Hogan, A., Mileo, A.: Using linked data to mine rdf from wikipedia's tables. In: The 7th ACM International Conference on Web Search and Data Mining, WSDM 2014, pp. 533–542 (2014)
12. Ning, X., Karypis, G.: Slim: sparse linear methods for top-n recommender systems. In: 2011 IEEE 11th International Conference on Data Mining, pp. 497–506. IEEE (2011)
13. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: The Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452–461 (2009)
14. Venetis, P., et al.: Recovering semantics of tables on the web. Proc. VLDB Endow. **4**(9), 528–538 (2011)
15. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural query performance prediction using weak supervision from multiple signals. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 105–114 (2018)
16. Zhang, S., Balog, K.: Ad hoc table retrieval using semantic similarity. In: The 2018 World Wide Web Conference, WWW 2018, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1553–1562 (2018)