Evaluation of Field-Aware Neural Ranking Models for Recipe Search

Kentaro Takiguchi
gm19099@bristol.ac.uk
University of Bristol
United Kingdom

Mikhail Fain
mikhail-fain@cookpad.com
Cookpad Ltd
United Kingdom

Niall Twomey
niall-twomey@cookpad.com
Cookpad Ltd
United Kingdom

Luis M Vaquero
luis.vaquero@bristol.ac.uk
University of Bristol
United Kingdom

ABSTRACT

Explicitly modelling field interactions and correlations in complex document structures has recently gained popularity in neural document embedding and retrieval tasks. Although this requires the specification of bespoke task-dependent models, encouraging empirical results are beginning to emerge. We present the first in-depth analyses of non-linear multi-field interaction (NL-MFI) ranking in the cooking domain in this work. Our results show that field-weighted factorisation machines models provide a statistically significant improvement over baselines in recipe retrieval tasks. Additionally, we show that sparsely capturing subsets of field interactions based on domain knowledge and feature selection heuristics offers significant advantages over baselines and exhaustive alternatives. Although field-interaction aware models are more elaborate from an architectural basis, they are often more data-efficient in optimisation and are better suited for explainability due to mirrored document and model factorisation.

CCS CONCEPTS

• **Applied computing** → **Document management and text processing**; • **Information systems** → **Learning to rank**.

KEYWORDS

recipes, field interactions, query-field, neural document ranking

1 INTRODUCTION

In information retrieval, single-field document ranking has been a central topic since the early days of the field. However, it is typical that documents to be retrieved have multiple fields in practical applications. Recipes are a canonical example of multi-field data, consisting of titles, ingredient lists, procedures, and images.

In such applications, it is common to collapse complex, multi-field documents uniformly into embeddings [8]. This approach has been highly successful [1, 6, 7, 12, 18, 20], but it is widely accepted that these data pipelines do not adequately capture how users view and interact with documents, nor do the models account for cross-field correlations [11].

Since the document's concept is the backbone that ties all fields together, correlated features will almost certainly be prevalent across fields. Feature duplication/overlap (resulting from correlated features) can result in more challenging loss surfaces and models are more likely to converge to local optima [3, 16].

Thus, de-correlating field representations empowers the model to learn field relevance from the training data and it further allows field weighting to be driven dynamically from context.

Many model-agnostic approaches for producing field-aware document representations exist such as BM25F, an heuristic extension of BM25 [14], Bayesian networks [13], LambdaBM25 [15], language modeling framework [10], probabilistic models [4], or feedback weighted field relevance [5].

There are also model-based approaches that learn query-field interactions such as Neural Ranking Models for Fields (NRM-F) [19], but fail to distinctly consider field-to-field interactions or account for correlated features. Other type of models could be extended to understand the value of complex field interaction modelling like field-weighted Factorization Machines (FwFM), which explicitly model feature interactions as arbitrary (non-linear) functions [11].

We explore a field-aware modelling approach in the domain of a recipe search application, including a breadth of options for model architecture and construction to determine whether dynamic field weighting can reduce relevance dilution effects arising from the consideration of irrelevant fields in recipe search [17].

Cooking has always been a vital daily routine for hundreds of millions of people, and under pandemic restrictions, it has brought

family cohesion and mental health support for many people globally [1]. Thus, improving retrieval tasks in the cooking domain has the potential to impart a positive impact on users of recipe recommendation services.

Our models are trained on recipe, click, and query data from Cookpad's search platform; we evaluated these as a suite of Non-Linear Multi-Field Interaction (NL-MFI) configurations.

The remainder of this document is structured as follows. Section 2 introduces our methodology, datasets, models, and evaluation protocol. Our experimental results and hypothesis evaluations are presented in Section 3, and we wrap up with discussions and conclusions in Section 4.

## 2 METHODOLOGY

2.1 Data and Preprocessing

This work focuses on recipe ranking. Recipes are complex multi-modal documents with several fields. We leverage existing search log data from the world's largest recipe community web service (www.cookpad.com) to learn and evaluate our models.

Since our setting is recipe search, a key field to consider is the user's search query. Once a query is executed, set of candidate recipes are selected and served to a user based on the query text. The objective of this paper is to produce field-aware ranking models that are capable of improving the search experience of users.

Supplementary to the query field, five additional recipe fields are used in our modelling: title, ingredients, description, country, and image. Most of these fields are textual and have variable size. The title field is usually short (consisting of ≈ 3.5 words on average), whereas ingredients and descriptions are typically much longer consisting of multiple free-text fields. Recipes are often accompanied by a set of images (one main image for the recipe, and at most 3 per step). We incorporate image embeddings from the main recipe image (when available) since it captures global visual recipe traits and aesthetics.

It is clear that these fields are correlated. For example, given a recipe titled 'pepperoni pizza', an experienced cook is likely to have prior expectations on the ingredients, steps and the look of the recipe. Users interact with search in a variety of ways. For users that search with simple queries, like 'chicken korma', one would expect that matches based on the recipe title is adequate. However, a wide variation of queries occur and users often with to filter by tool ('korma with a crock pot'), questions ('how can I bake cake without oven?'), specific sets of potentially negated ingredients ('cake without milk'), diets ('gluten free bread'), religious dietary restrictions ('halal chicken curry'), and other meta terms ('healthy/fast/easy soup') at different levels of specificity ('dinner' *vs.* 'sugar-free granola for breakfast'). This variation of query types demonstrates the organic interactions between users and search and the nuanced link between the query and user expectations. The opportunity to avoid prescribing actions based on query types but instead to learn behavioural patterns from interaction data is a key motivator for this research.

We can completely reconstruct search sessions from the data logs. Each session is identifiable by a unique session ID, and aligned
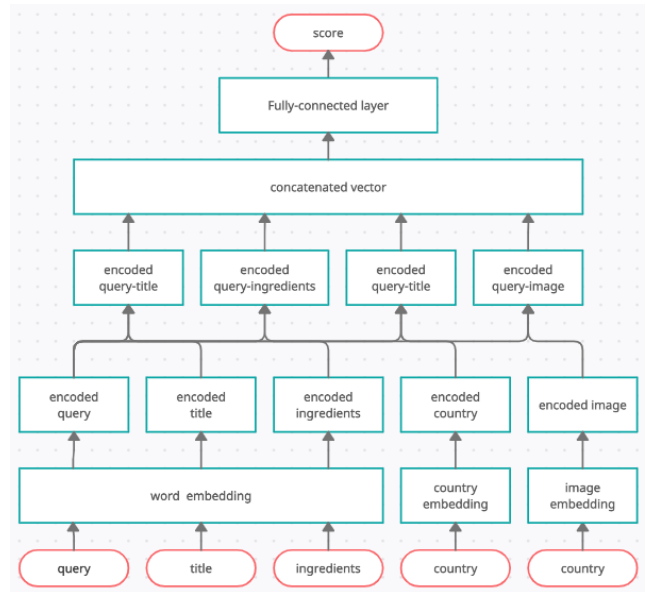
(a) NRM-F architecture.
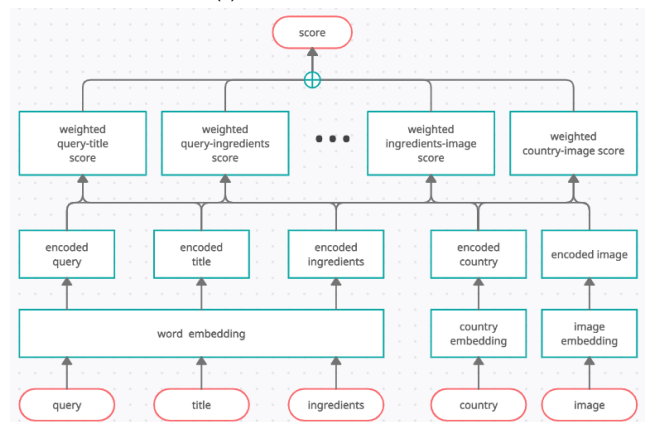
(b) FwFM architecture

Figure 1: Architecture of field-aware models considered.

metadata is also available including event time, retrieved recipe IDs, clicked recipe ID, and query string. Over 99% of queries are 'clean' in the sense that the query is specifically targeting ingredients, dishes, meals etc. Other queries (e.g. 'what is quinoa?') are removed when possible. We truncate the candidate recipe data at the last clicked position. Since we are aiming to optimise for click-through metrics, we implicitly treat the items which were examined by a user but not clicked as negatives. Only clicked items are positives and treated as the ground truth for training and evaluation.

We follow standard approaches for embedding images and text. We use average word embeddings for each field which, despite its simplicity, has been shown to achieve similar accuracy than recurrent units with significantly less training time [9]. Image embeddings are obtained from pre-trained on ImageNet [2].

2.2 Models

We focus on the two field-aware models: NRM-F [9] and FwFM [11]. Our main objective is to assess their performance on recipe retrieval tasks. We will also investigate whether incorporating small architectural adjustments can improve retrieval performance.

Table 1: Model and field interaction configurations of the two architectures as specified in original publications.

	NRM-F	FwFM
First-order features	Not used	Used
Interaction selection	Query-field	All
Interaction representation	Hadamard product	Dot product
Interaction aggregation	Concatenation	Weighted sum

The architecture of these two models is shown in Figure 1 and the key differences are summarised by interactions, representations and aggregations (see Table 1). Field interactions are broadly categorised 2-nd order interactions (consisting of $k(k-1)/2$ pairwise combinations across fields) and 1-st order fields (consisting of k individual fields). The original NRM-F model specifies 'query-field' interactions, where the 'query' field is paired with all other 1-st order interactions (yielding $k-1$ interaction pairs). We also use the term 'all' interactions, which refers to the union of 1-st and 2-nd order interactions, and this was used by FwFM (giving $k(k+1)/2$ interaction pairs).

2.3 Evaluation

We use Normalised Discounted Cumulative Gain (NDCG) at 20 to evaluate models, which measures the quality of the first page of results. We evaluate performance with multiple pairwise t-tests with the Bonferroni correction with a (fairly stringent) threshold of significance set to $\alpha = 0.01$. Performance is evaluated with 10-fold cross validation for parameter selection. Each train fold contains about 33,000 search sessions on average.

All code for the proposed work is available on GitHub [2].

3 EXPERIMENTS

Our experiments explore variants of the FwFM and NRM-F models on recipe retrieval based on interaction pairings ('1-st', '2-nd', 'query-field', and 'all' which are defined in Section 2.2).

3.1 Effect of Interaction Selection

In this section, we investigate the effect of various interaction selection mechanisms on both models. Specifically we consider 'query-field' and 'all' interactions (defined in Section 2.2) on FwFM and NRM-F models. Intuitively, the 'query-field' interactions result in simpler models since fewer cross-field interactions are derived, but the 'all' configuration may be better able to learn complex interactions. As a baseline we consider a basic model which is the concatenation of all 1-st order features, and we also report performance of random models.

---

[2]https://github.com/rejasupotaro/field-interactions-in-document-ranking

Table 2: Comparison of NRM-F and FwFM models on mean NDCG scores. The FwFM query-field results are statistically significant better than all other configurations.

Model	NDCG@20
Random	0.523
Baseline	0.643
NRM-F model (all)	0.641
NRM-F model (query-field)	0.652
FwFM model (all)	0.661
FwFM model (query-field)	0.667 [†]

NDCG@20 scores over the five model configurations are shown in Table 2. We can note that the query-field interactions tended to consistently out-perform the others on both NRM-F and FwFM. Noting that this configuration is a subset of the 'all' configuration, it is interesting to see that the 'simpler' model out-performed the more complete interaction specification. The best performing model was 'FwFM (query-field)', whose results were significantly better than all others considered. We note also that we explored this model configuration to understand which configuration is most optimal for recipe re-ranking tasks. This novel configuration was not explored in the original FwFM paper. Random performance is relatively high because models are evaluated on search sessions that are cut-off at the last clicked position.

The distribution of results are shown in Figure 2. We can see that as well as having a higher median, the spread and inter-quartile range of this configuration is narrower.
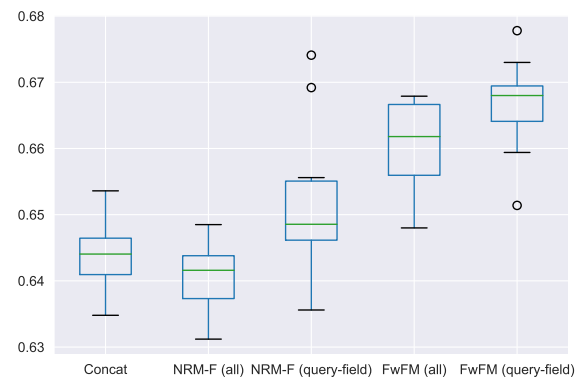

Figure 2: This image shows the spread of NDCG scores for various model configurations. We can see that FW query-field results are more tightly packed.

3.2 Non-Linear Field Interactions

We continue the exploration on new interaction pair configurations in this section, specifically to understand the value given by 1-st order interactions. Table 3 shows the average NDCG scores for the above models. The results for the FwFM have not significantly changed, though they dipped slightly. However, the average results for the NRM-F based model have improved significantly over

Table 3: Comparison of models with different feature sets. FwFM with 1-st and 2-nd interactions are statistically significant better than all other configurations.

Model	1-st	2-nd	NDCG@20
NRM-F		✓	0.652
NRM-F	✓	✓	0.650
FwFM		✓	0.645
FwFM	✓	✓	0.661 [†]

Table 4: The results for the interaction-oriented experiments. The performance of 'FwFM (all)' is significantly worse than the other two models.

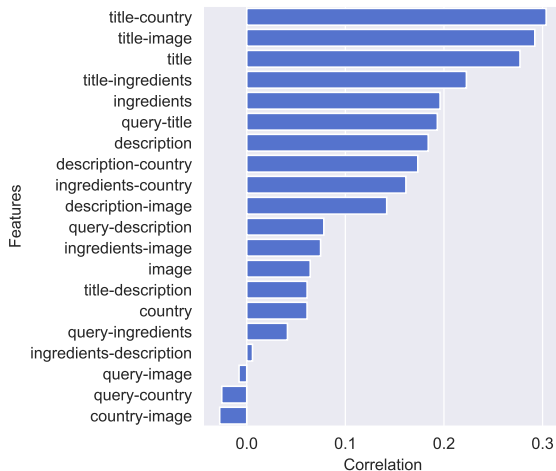| Model	NDCG@20
FwFM (all)	0.661
FwFM (selected)	0.663 [†]
FwFM (query-field)	0.667 [†]

Figure 3: This figures shows the calculated correlation between interaction terms and the click labels. Larger values indicate closer connections.

the baselines presented in Section 3.1. FwFM models are still the top-performing on this dataset, and 'FwFM 1-st & 2-nd' remains significantly better than the rest.

3.3 Exploring High-Yield Interactions

In this section we take an ablation approach to understand whether the inclusion of novel interaction pairs can improve on the query-field results. The correlation between the individual interaction scores and ground truth click labels from the FwFM model trained on all 1-st and 2-nd order interactions is calculated. FwFM was chosen because it is the strongest performing model on our dataset. This correlation score is assumed to act as a proxy indicator to field importance, and the interactions are sorted by their correlation score in order to help us understand the pairs which are deemed important.

Figure 3 shows the correlations between interaction and click labels. *A priori*, one may have expected query-field interactions to be the most highly correlated to labels in search ranking tasks. However, a simple inspection of Figure 3 shows that 'query' fields are not highly ranked. To understand this, we must remember that our ranking task follows an initial candidate generation phase which we do not influence (see Section 2.1). As a result the items that we score are implicitly dependent on and correlated to the query even

when not directly including the query. A complementary perspective is that the query-field interactions do not add significant *new* information to the scoring model, implying that the interactions that are most highly correlated to the labels are those which contribute new information. We note too that the query-field set of interactions would almost certainly be the most highly scored fields in candidate generation tasks.

Interactions on other field pairs build on the implicit query dependency of candidates and introduce new features which lead to them being highly ranked. The highest pair is 'title-country.' We understand the importance of this arising due to regional preferences being strong indicators of interest in and of itself. We were surprised to see that image features have relatively low weight. This is a valuable insight of the experiment since *a priori* image aesthetics were assumed would contribute highly to clicks due to their visual appeal. The low correlation on all image pairs indicates that image features (which are expensive to compute) may be a candidate for removal.

We leverage these results and construct a correlation-driven interaction model on a sparse subset of interactions. Interaction pairs are selected from the validation set. In Table 4 we can see that we can improve over baseline performance with this approach. Both 'query-field' and 'selected' offer a statistically significant improvement over 'all' but with many fewer interactions modelled. The performance of 'query-field' and 'selected' configurations is similar, with no significant differences between them. Although we have not beaten the query-field results with correlation selection, we believe this approach is a valuable contribution since it extends the scope for these models with improved performance. For example, query-field interactions cannot be used in re-ranking recipe feeds since no 'query' field exists, and performance in these situations may benefit from restricted interactions achieved by the correlation-driven interaction selection.

4 CONCLUSIONS

Strong early evidence is provided in this paper for the broad advantages provided by non-linear multi-interaction field models in the domain of recipe ranking. By reducing between-field feature correlations and methodically calculating significant improvement over baselines, we demonstrate how new insights can be gleaned for practitioners about their domain. We believe that we have shown that non-linear multi-field interaction models investigated are strong candidates for the domain. The next steps for this research are to explore the value of these models in online experiments and across different domains and tasks.

REFERENCES

[1] Marko Balabanović and Yoav Shoham. 1997. Fab: content-based, collaborative recommendation. *Commun. ACM* 40, 3 (1997), 66–72.

[2] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[3] Donald E Farrar and Robert R Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* (1967), 92–107.

[4] Jinyoung Kim, Xiaobing Xue, and W. Bruce Croft. 2009. A Probabilistic Retrieval Model for Semistructured Data. In *Advances in Information Retrieval*, Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 228–239.

[5] Jin Young Kim and W Bruce Croft. 2012. A field relevance model for structured document retrieval. In *European Conference on Information Retrieval*. Springer, 97–108.

[6] R Logesh and V Subramaniyaswamy. 2019. Exploring hybrid recommender systems for personalized travel applications. In *Cognitive informatics and soft computing*. Springer, 535–544.

[7] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.

[8] B. Mitra and N. Craswell. 2018. . https://doi.org/10.1561/1500000061

[9] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019. Semantic Product Search. *CoRR* abs/1907.00937 (2019). arXiv:1907.00937 http://arxiv.org/abs/1907.00937

[10] Paul Ogilvie and Jamie Callan. 2003. Combining Document Representations for Known-Item Search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 143–150. https://doi.org/10.1145/860435.860463

[11] Junwei Pan, Jian Xu, Alfonso Lobos Ruiz, Wenliang Zhao, Shengjun Pan, Yu Sun, and Quan Lu. 2018. Field-weighted factorization machines for click-through rate prediction in display advertising. In *Proceedings of the 2018 World Wide Web Conference*. 1349–1357.

[12] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.

[13] Benjamin Piwowarski and Patrick Gallinari. 2003. A machine learning model for information retrieval with structured documents. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 425–438.

[14] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (Washington, D.C., USA) *(CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 42–49. https://doi.org/10.1145/1031171.1031181

[15] Krysta M Svore and Christopher JC Burges. 2009. A machine learning approach for improved BM25 retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 1811–1814.

[16] Laura Toloşi and Thomas Lengauer. 2011. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* 27, 14 (2011), 1986–1994.

[17] Niall Twomey, Mikhail Fain, Andrey Ponikar, and Nadine Sarraf. 2020. Towards Multi-Language Recipe Personalisation and Recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 708–713.

[18] Vaishali S Vairale and Samiksha Shukla. 2021. Recommendation of Food Items for Thyroid Patients Using Content-Based KNN Method. In *Data Science and Security*. Springer, 71–77.

[19] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2017. Neural Ranking Models with Multiple Document Fields. *CoRR* abs/1711.09174 (2017). arXiv:1711.09174 http://arxiv.org/abs/1711.09174

[20] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. 2008. Large-scale parallel collaborative filtering for the netflix prize. In *International conference on algorithmic applications in management*. Springer, 337–348.