Multi-Field Models in Neural Recipe Ranking

An Early Exploratory Study

Kentaro Takiguchi
gm19099@bristol.ac.uk
University of Bristol
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 documents 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 offers advantages over 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

Here, we define a document's fields as self-contained sub-documents with tight links to the document itself (e.g. title, main body). In information retrieval and recommendation, it is common to collapse complex, multi-field documents uniformly into embeddings [11]. This approach has been highly successful for many years across a multiplicity of applications [1, 9, 10, 15, 23, 25]. Yet, despite its broad success, 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 [14].

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 contaminate the optimisation objective resulting in more challenging loss surfaces and models more likely to converge to local optima [4, 21]. Thus, correlations between the document fields seem to be a critical factor that require consideration in modelling.

Taking a field-aware modelling approach offers several advantages. It de-correlates 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. Expanding on this, a recipe search application, for example, may learn that the title field is highly relevant for the query 'korma' and that the query 'slow cooker' aligns better with the procedure. In this manner, dynamic field weighting can reduce relevance dilution effects arising from the consideration of irrelevant fields [22].

We consider neural recipe re-ranking of search results in this work. Recipes are a canonical example of multi-field data, consisting of titles, ingredient lists, procedures, and images. 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 learnt on recipe, click, comment, and query data from Cookpad's search platform; we evaluated these as a suite of Non-Linear Multi-Field Interaction (NL-MFI) configurations. This early work is aimed at addressing these key hypotheses:

- H1. Redundant or partially overlapped field-to-field interactions may have a detrimental impact on the performance on neural recipe ranking tasks. Thus, low order interactions (1^{st} and 2^{nd}) are enough to obtain the best performance.
- H2. A selection of sparse field interactions based on field correlations may not result in additional benefits compared to using non-selected 1^{st} and 2^{nd} interactions (hidden inter-dependencies).

The remainder of this document is structured as follows. In Section 2 we review related work. Section 3 introduces our methodology, datasets, and evaluation procedures. Our experimental results and hypothesis evaluations are presented in Section 4, and we wrap up with discussions and conclusions in Sections 5 and 6.

2 RELATED WORK

While there are many works focusing on the application of neural models to information retrieval [11], most of them treat each document as a single instance of text (i.e., single field) disregarding semi-structured information in multiple fields.

Neural Ranking Models for Fields (NRM-F; its naming is inspired by BM25F [16]) is the seminal paper that discusses how neural models can deal with multiple document fields from an architectural perspective [24]. The work suggests that it is better to score the whole document jointly, rather than generate a per-field score and aggregate. Its formulation for document representation learning function Φ_D is as follows:

$$\Phi_D(F_d) = \Lambda_D\left(\Phi_{F_1}(F_1), \Phi_{F_2}(F_2), \cdots, \Phi_{F_k}(F_k)\right)$$

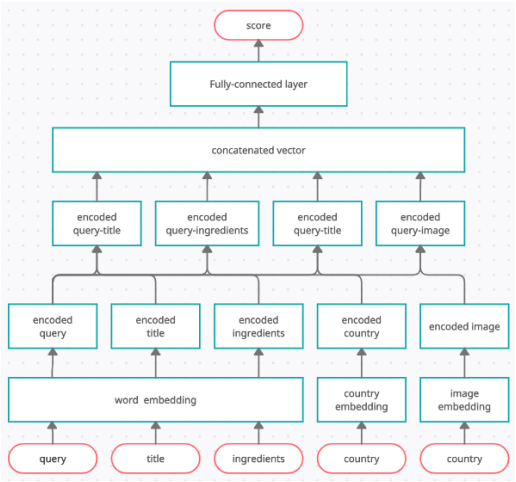
[1] https://medium.com/cookpadteam/the-changing-face-of-italian-cooking-during-lockdown-7b1bbbcb2b56

where Φ_{F_i} denotes the mapping function for the field F_i and Λ_D is an aggregation function which consolodates representations learned for all the fields. In their formulation, Λ_D is the concatenation function which combines the input vectors. A densely-connected a stack of layers outputs the final retrieval score.

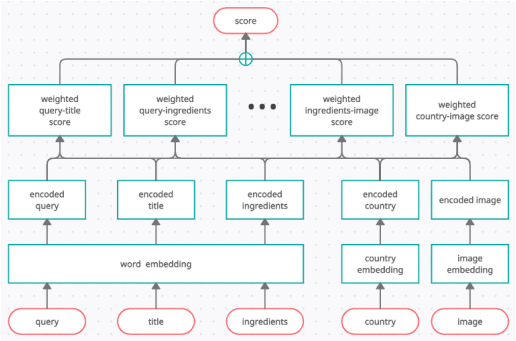
In NRM-F, both query text and text fields are represented using a character n-gram hashing vector [5], and a convolution layer is then employed to capture the dependency between terms. Although NRM-F explicitly learns query-field interactions, it does not distinctly consider field-to-field interactions. In this manner, it can learn about the relevance of each field, but not account for correlated field features. There is some existing work that investigates query-to-field interactions in these applications [8], but this work assumes that linear relationships between relevance models induced from each field.

As well as NRM-F, several other potential models can be extended to shine light on the value of complex field interaction modelling. Factorization Machine (FM) are one model type that are widely used supervised learning approach. These models effectively modeling of feature interactions, and interactions can be captured with arbitrary interaction functions, allowing for models that capture effect non-linear interactions between fields. Field-weighted Factorization Machine (FwFM) are state-of-the-art among the shallow models for click-through-rate prediction [14].

Many non model-based approaches for producing field-aware document representations exist. The classic BM25 heuristic, for example, has a field-aware varient called BM25F that is used information from multiple fields [18, 19]. Other approaches built on this idea without resorting to a linear combination of per-field scores: like, for instance, Bayesian networks [17], LambdaBM25 [20] (based on LambdaRank [2]), language modeling framework [13], probabilistic models [6], or feedback weighted field relevance [7].



(a) NRM-F architecture.



(b) FwFM architecture

Figure 1: Architecture of considered models.

3 METHODOLOGY

3.1 Datasets

Cookpad[2] is the world's largest recipe community web service where users can publish and search for recipes. For the experiements described later, we consider recipe and search log data from Cookpad.

Recipes consist of multiple fields and media types. We selected six fields for modelling: query, title, ingredients, description, country, and image. The title field is usually short and consists of 3.47 words on average. The ingredients field is a variable-sized set of texts since some recipes require more ingredients than others. The description is also a list of free text fields; some recipes have a surprisingly long description while there are recipes with virtually no description. The country field indicates the country in which the recipe was published. Recipes have a main image and several step images, but only the main image is used in the experiment since it captures the whole recipe, rather than capturing a specific step.

Search events are tracked when a user clicks a recipe in the search results. Each log contains session ID, event time, retrieved recipe IDs, clicked recipe ID, and clicked position. The average number of words in a query is 2.25. A variety of query types arre found in the logs, including question-like queries such as 'how to bake cake without oven' and 'what is buttermilk'. Over 99% of queries are 'clean' with users specifically searching for ingredients, dishes, regions, etc.

3.2 Data Processing and Modeling

Search logs are aggregated by session ID and query to form listwise data. Each result list is trimmed at the last clicked position in the list, and items above the position are treated as negatives, as they were examined by a user but not clicked. Regarding the text representation, we obtain fix-sized vectors under the assumption that terms are almost independent as observed earlier. We use the average of term vectors, which has been shown to perform similar or slightly better than recurrent units with significantly less training time [12]. Countries are treated as a category and embedded into a latent space. We employ an image embedding that is pre-trained on ImageNet [3].

In order to test if field interactions affect ranking performance in an architecture-dependent manner, we focus on two architectural models: NRM-F [12] and FwFM [14] to examine how the choice of architecture affects effectiveness along with the concatenation model that concatenates all encoded features as baseline. Figure 1

[2]Cookpad http://www.cookpad.com

show the diagrams of NRM-F (1a) and FwFM (1b) employed in our experiments. The characteristic differences of how fields are modelled, aggregated and distilled into scores are summarised in Table 1. Our experimentation will explore several novel configurations of both models that are described below.

Table 1: Model and field interaction configurations of the two architectures evaluated in this work.

	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

We employ Normalised Discounted Cumulative Gain (NDCG) at 20 to evaluate models. The cuttoff of 20 was chosen since this is the number of recipes served per page. The loss is computed in a pairwise fashion. We evaluate performance with pairwise T-tests with a (fairly stringent) threshold of significance set to $\alpha = 0.01$.

The entire dataset is divided into 10 sets by timestamp to obtain a sufficient number of individual datasets to evaluate the statistical significance of the obtained results. Each dataset is further divided by timestamp, with the first 75% used for training and the remaining 25% for validation.

3.3 Terminology

By 1^{st} order interactions, we refer to features constructed taking individual fields into consideration. Hence, 2^{nd} order interactions are composites of pairs of features. By all interactions, we refer to models that include 1^{st} and 2^{nd} order features altogether. Query field interactions are just a specific type of 2^{nd} order features

4 EXPERIMENTS

All the experiments presented in this section are available on GitHub.

4.1 Query-to-field vs Field-to-field Interactions

We gauge the effects of adding all feature interactions vs. focusing on query-field interactions only as follows.

- **Concatenation (all)** Concatenate all features into a vector (as a baseline implementation).
- **NRM-F (all) / FwFM (all)**: Consider all feature interactions without distinguishing between query and fields.
- **NRM-F (query-field) / FwFM (query-field)**: Consider query-field interactions.

Table 2 shows the mean of the scores of the mentioned models. The models that learned query-field interactions outperformed models trained using all interactions in both NRM-F and FwFM. The results of 'FwFM (query-field)' are the best, and its improvement over all others is statistically significant.

Figure 2 plots the NDCG scores on each validation data out of 10 splits. We can see that as well as having a higher median, the spread and inter-quartile range of this configuration is narrower. Given the statistical significance of the results, and the general improvements, these results add support for Hypothesis 1.

Table 2: Comparison of NRM-F-based and FM-based models on mean NDCG scores. † specifies statistically significant .

Model	NDCG@20
Concatenation model (all)	0.643
NRM-F-based model (all)	0.641
NRM-F-based model (query-field)	0.652
FM-based model (all)	0.661
FM-based model (query-field)	0.667 †

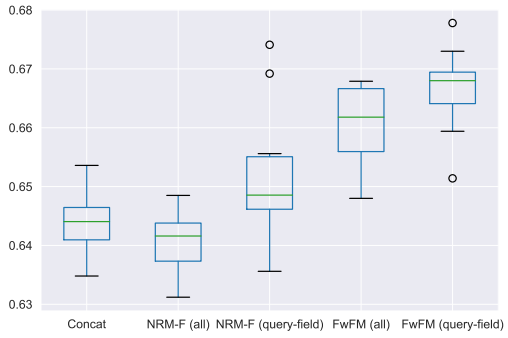


Figure 2: NDCG scores for various model configurations.

4.2 Beyond Query-Field Interactions

Subsection 4.1 suggests some naturally leads to wondering whether other field interactions beyond query-field interactions can be identified. In this set of experiments, we employed the FwFM model, trained using 1^{st} and 2^{nd} order interactions (6 and 15 feature interactions respectively). FMs compute the scores for each field independently and sum them up to produce the final score. We trained the model regularly and extracted the individual feature scores on validation data.

Since the sum of the individual scores is the final score, we sort the features by their correlation with label, assuming that correlation is a proxy indicator of field importance. Then, we compare the performance of the following three models.

Figure 3 shows the correlations of the activation of fields by label. We can see that several interaction pairs are highly correlated with the labels. The highest (title-country) suggests that regionality and the category (captured through the title) are highly predictive for search re-ranking. It is likely that the correlations between these two fields capture regional preferences which results in the high value of this pair. Image has a low weight. This is somewhat surprising and a useful outcome of the experiment since *a priori* image aesthetics were assumed to contribute highly to interactions due to their visual appeal. The low correlation on all image pairs indicates that image features (which are expensive to compute) may be dropped from models with limited consequences to metrics.

We can use the correlation figure to select a sparse subset of interactions to capture. We make selections by choosing the feature pairs that are most correlated with click labels in the validation set, and Table 3 presents the results. This table shows that using the correlation table to specify model architectures can result in

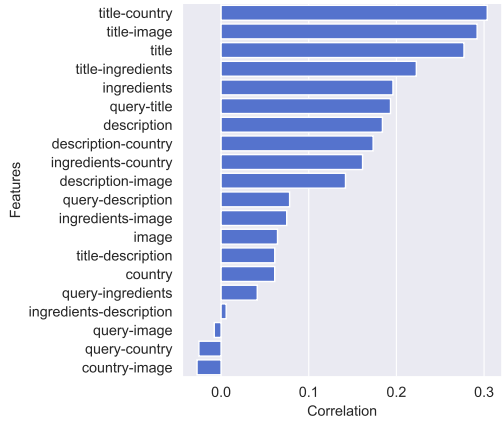


Figure 3: The correlation to the label of each feature.

Table 3: The results for the interaction-oriented experiments. † indicates that performance improvement is statistically significant over unmarked rows.

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

improved performance. Both 'query-field' and 'selected' offer a statistically significant improvement over 'all' but with many fewer interactions modelled. Although the 'query-field' results are higher, the 'selected' option may be useful in scenarios without a query term (e.g. for social network feeds). These results offer evidence for accepting Hypothesis 2 on the value of the value of sparse interaction pairs is provided here.

4.3 Non-Linear Field Interactions

The original implementation of NRM-F does not use 1^{st} order field interactions. In this section, we explore the impact of 1^{st} order field interactions in performance. We compare NRM-F-based and FM-based models with different sets of features as follows:

- NRM-F (2): Use 2^{nd} order query-field interactions only (as the original implementation).
- NRM-F (1 & 2): Use 1^{st} order features along with 2^{nd} order query-field interactions.
- FwFM (2): Use 2^{nd} order interactions only.
- FwFM (1 & 2) : Use 1^{st} and 2^{nd} order interactions (as the original implementation).

Table 4 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 the baselines presented in Section 4.1. Since including field interactions improved NRM-F, we believe evidednce for Hypothesis 2 is provided here. FwFM models are still the top-performing on this dataset, and 'FwFM 1^{st} & 2^{nd}' is significantly better than the rest.

Table 4: Comparison of models with different feature sets.

Model	Interactions	NDCG@20
NRM-F model	2^{nd}	0.652
NRM-F model	1^{st} & 2^{nd}	0.650
FwFM model	2^{nd}	0.645
FwFM model	1^{st} & 2^{nd}	0.661 †

4.4 Discussion

Our experiments investigate the value of modelling field-interactions in the recipe retrieval domain. We have repeatedly shown statistically significant performance gains that are attributed to factorising document fields and interactions in the model architecture. The main advantage gained by this architectural decision is that between-field correlations are reduced when compared to baseline models that we considered, which are exhaustive on interactions.

The FwFM model architecture is consistently shown to be superior to baseline and NRM-F models for in the recipe retrieval domain across several novel architectures. Nonetheless, we believe that incorporating field interaction in model architectures should offer improvements. Experiments with NRM-F architecture with first and second order interactions have shown this to be the case, and make significant improvements over the original architecture.

Reducing feature correlation and duplication in field representations effects can lead to simpler optimisation objectives. We believe this to be the main contributing factor to the improvements we have reported. Indeed, we have shown that it is easy to capture the field-wise correlations associated with click labels in validation sets, and to use these to 'sparsify' model architectures. We show that the performance increases using this technique, the model has fewer parameters to optimise, the analysis of the results offer new insights to practitioners about their domain. Our experiments and results are highly focused in search retrieval tasks. For this reason the 'query'-related interactions are of high value. Outside of search, this field is unavailable, however, but the correlation-based field selection technique is flexible and still a highly suited to a broad set of interfaces (e.g. social network feed).

5 CONCLUSIONS

Strong evidence is provided in this paper for the broad advantages provided by field-interaction models in the domain of recipe retrieval. By reducing between-field feature correlations, providing models that are more data efficient, repeatedly calculating significant improvement over baselines, and demonstrating 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 data sources (e.g. feed).

REFERENCES

[1] Marko Balabanović and Yoav Shoham. 1997. Fab: content-based, collaborative recommendation. *Commun. ACM* 40, 3 (1997), 66–72.
[2] Christopher J Burges, Robert Ragno, and Quoc V Le. 2007. Learning to rank with nonsmooth cost functions. In *Advances in neural information processing systems*. 193–200.

[3] 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

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

[5] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2333–2338.

[6] 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.

[7] 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.

[8] Binsheng Liu, Xiaolu Lu, Oren Kurland, and J. Shane Culpepper. 2018. Improving Search Effectiveness with Field-Based Relevance Modeling. In *Proceedings of the 23rd Australasian Document Computing Symposium* (Dunedin, New Zealand) *(ADCS '18)*. Association for Computing Machinery, New York, NY, USA, Article 11, 4 pages. https://doi.org/10.1145/3291992.3292005

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

[10] 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.

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

[12] 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

[13] 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

[14] 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.

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

[16] José R Pérez-Agüera, Javier Arroyo, Jane Greenberg, Joaquin Perez Iglesias, and Victor Fresno. 2010. Using BM25F for semantic search. In *Proceedings of the 3rd international semantic search workshop*. 1–8.

[17] 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.

[18] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)* (overview of the third text retrieval conference (trec–3) ed.). Gaithersburg, MD: NIST, 109–126. https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/

[19] 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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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

[25] 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.