BM25-FIC: Information Content-based Field Weighting for BM25F

Tuomas Ketola and Thomas Roelleke

Queen Mary, University of London, UK
{t.j.h.ketola,t.roelleke}@qmul.ac.uk

Abstract. BM25F has been shown to perform well on many multi-field and multi-modal retrieval tasks. However, one of its key challenges is finding appropriate field weights. This paper tackles the challenge by introducing a new analytical method for the automatic estimation of these weights. The method — denoted BM25-FIC — is based on field information content (FIC), calculated from term, collection and field statistics. The field weights are applied to each document separately rather than to the entire field, as normally done by BM25F where the field weights are constant across documents. The BM25-FIC outperforms the BM25F in terms of P@10, MAP and NDCG on a small test collection. Then the paper introduces an interactive information discovery model based on the field weights. The weights are used to compute a similarity score between a seed document and the retrieved documents. Overall, the BM25-FIC approach is an enhanced BM25F method that combines information-oriented search and parameter estimation.

1 Introduction

Formal retrieval models for multi-modal and heterogeneous data are becoming more necessary, as the complexity of data-collections and information needs grow. Formality is required to keep the models interpretable, a quality often expected in fields such as law and medicine. Most of the data-collections searched these days — whether it is websites, product catalogues or multi-media data — consist of objects with more than one feature and more than one feature type.

BM25F has been shown to be effective for multi-modal and multi-field retrieval [**?**]. One of its main challenges is the choice of field weights. The main contribution of this paper is to introduce a new method for automatically determining these weights; the BM25-FIC (BM25 Field Information Content).

The proposed method calculates the field weights based on field information content, estimated from term, collection and field statistics. As the weights are calculated directly, no learning or heuristics are needed to determine appropriate field weights, as is the case with BM25F. This makes BM25-FIC much easier to implement. Furthermore, the field weights are determined for each document

field separately, rather than for the entire field of the collection, as is done by the normal BM25F. This means that the BM25-FIC is able to capture more complicated relationships between the query and the different fields.

Our experiments confirm that BM25-FIC outperforms BM25F. However, it needs to be noted that this result is obtained on a small test-collection and without training for the BM25F weights. Training was not performed on the benchmarks as BM25-FIC itself requires no training.

The second contribution of the paper is to introduce an interactive information discovery model that benefits from the obtained field weights. It uses a seed document as a reference point to help the user better define their query intent.

2 Related Work

Multi-modal retrieval has received much attention in the IR community. Multi-modal approaches closely relate to multi-field and multi-model approaches. Here the terms multi-modal and multi-field are used interchangeably as the fields are assumed to represent different data types. However, our approach is not a multi-model one, as the BM25 is used for all fields. Multi-modal data can be fully text based, rather than audio and text for example, as different feature types can be represented in text, e.g. author lists, abstracts or geographic information [**?**].

It has been shown that the BM25F generalizes well to multi-modal data [**?**]. As with normal textual data, this involves setting the field weights. He and Ounis have examined the setting of the field weights and other field level hyperparameters extensively [**?**]. Outside the BM25F, various probabilistic and learning based models have been considered for multi-field ad-hoc-retrieval. These models will not be explained in detail here as the focus is on the BM25F. Instead, the reader is advised to see [**?**] for a summary of the different approaches.

The principle of polyrepresentation relates to the concept of relevance dimensionality discussed in this paper. According to the principle, relevance consists of multiple overlapping cognitive representations of documents. The most relevant documents are most likely found where these representations overlap [**?**,**?**]. The different dimensions of relevance, represented by the different documents fields, can be seen as forms of cognitive representations when they communicate different types of information.

3 BM25-FIC - Information Content based BM25F

There are two common ways in which multiple fields are considered in the BM25 context. The first option is to get the BM25 scores for each field and calculate a weighted sum over them. In this paper this approach is denoted BM25F-macro:

$$\mathrm{RSV}_{\text{BM25F-macro},b,k_1}(q,d,c) := \sum_{f \in F_c} w_f \ \mathrm{RSV}_{\text{BM25},b,k_1}(q, f_d, c) \qquad (1)$$

where q is a query, d a document, c a collection, f a document field, and w_f is the field weight. F_c is the set of fields. Note that f is the type (e.g. title, abstract,

body) whereas f_d denotes an instance (e.g. the title of document d). b and k_1 are BM25 hyper-parameters.

$$\mathrm{RSV}_{\mathrm{BM25},b,k_1}(q, f_d, c) := \sum_{t \in f_d} \mathrm{TF}_{\mathrm{BM25},b,k_1}(t, f_d, c) \cdot w_{\mathrm{RSJ}}(t, c) \tag{2}$$

where the TF component is the BM25 term frequency quantification:

$$\mathrm{TF}_{\mathrm{BM25},b,k_1}(t, f_d, c) := \frac{(k_1 + 1)\, n(t, f_d)}{n(t, f_d) + k_1 \left(b \frac{\mathrm{len}(f)}{\mathrm{avgfl}(c)} + (1 - b)\right)} \tag{3}$$

where $n(t, f)$ is the raw term frequency, and avgfl is determined globally for the collection. $w_{\mathrm{RSJ}}(t, c)$ can be defined based on documents, or can consider field-based frequencies. For example, $w_{\mathrm{RSJ}}(t, c) := \frac{N_F(c)+0.5}{\mathrm{df}(t,F_c)+0.5}$ is a field-based rather than document-based weight, as described by Robertson et. al [**?**].

The second option for multi-field BM25 retrieval was introduced by Robertson et. al as they noted that approaches which use weighted sums of field based retrieval scores give too much weight to some query terms [**?**]. Their approach is different from BM25F-macro in that the constant field weights are applied to the raw term frequencies $n(t, f)$ and the BM25 score is calculated over the summed terms frequencies from all the fields. This model is commonly known as BM25F, here it is denoted BM25F-micro for clarity.

In both BM25F-macro and BM25F-micro the field weights are set as constant and are applied in the same manner to each document in the corpus. Therefore, the BM25F-macro and BM25F-micro models assume that a given field always affects relevance in the same manner. Furthermore, field weights are defined through learning, or heuristics — both costly tasks.

To counteract these two issues, we propose the BM25-FIC which does not assume the field weights to be constant. The ranking score between a query and a document is defined as the weighted sum of the document field BM25 scores, where the weight is calculated from the information content of a document field.

Definition 1 (BM25-FIC Ranking Score).

$$\mathrm{RSV}_{\mathrm{BM25\text{-}FIC},\mathrm{Inf},b,k_1}(q, d, c) := \sum_{f \in F_c} w_f(q, c, \mathrm{Inf})\, \mathrm{RSV}_{\mathrm{BM25},b,k_1}(q, f_d, c)$$

where Inf is the chosen information content model.

Comparing Definition 1 to BM25F-macro (or micro), it is clear that they are closely related. The difference is that instead of having constant field weights, in BM25-FIC the weight is dependent on the query q, the document field f, the collection field F, the collection c and the information content model Inf.

3.1 Rationale for the BM25-FIC Score and the Field Weights

The main research question in this paper is the rationale and estimation of the field weights $w_f(q, c, \mathrm{Inf})$. Before we propose the estimates, we consider the wider

picture of probability and information theory for creating the rationale for the estimation of w_f.

An aggregation of values (scores) as for BM25F, is inherently related to the 1st moment (expected value):

$$\text{Mean:} \quad E[X] = \sum_{x \in X} x \; P(x)$$

Regarding BM25F, x is a score for a field, and $P(x)$ is a probability associated with the field.

Regarding an information-theoretic approach, the entropy is the expected value (EV) of the negated logarithm of the probability:

$$\text{Entropy:} \quad E[-\log(P(X))] = -\sum_{x \in X} P(x) \; \log(P(x))$$

Entropy or related concepts such as log-likelihood are commonly used for justifying estimates. These probabilistic and information-theoretic rationales justify the field weights.

3.2 Estimates for Field Weights w_f(Inf)

Following the framework of probabilistic and information-theoretic expectation values, the candidates for w_f are derived from the concept of information content:

$$w_f(q, c, \text{Inf}) = \text{Inf}(q, f, c) := -\sum_{t \in q \cap f} \log(P(t|f, c)) \tag{4}$$

$P(t|f, c)$ is defined via the max-likelihood method as the number of document fields where term t occurs ($\text{df}(t, f, c)$), divided by the number of potential document fields where t could appear: $P(t|f, c) := \frac{\text{df}(t,f,c)}{N_P}$.

Three different definitions of the number of potential documents ($N_P(c)$) are used to create the three candidate models for information content Inf (Inf_1, Inf_2 and Inf_3).

Estimate P1 The first model defines N_P as the total number of documents in the collection: $N_{P1}(c) := N_D(c)$.

Estimate P2 The second model defines N_P as the number of documents in the collection for which the field in question is not empty, that is contains at least one term. $N_{P2}(c) := |\{d|f \in F_c \wedge \exists t, f_d : n(t, f_d) > 0\}|$, where f_d is the instance of a field in document d.

P2 ensures that fields which are empty for many documents are given less weight than they would otherwise. This makes sense as often fields are empty for reasons, such as data redundancies.

Estimate P3 The third model normalizes N for each field according to their average field lengths (avgfl): $N_{P3}(c) := N_{P2}(c) \frac{\text{avgfl}(c)}{\text{avgfl}(f)}$. where avgfl($c$) is the avg field length over all fields, and avgfl(f) is the average for a specific field (e.g. title).

P3 ensures that short fields get more weight. Adding weight to shorter fields has been shown to be beneficial in previous research [**?**].

3.3 Evaluation Data and Results

For evaluation we use the Kaggle Home Depot product catalogue data set[1], also used by [**?**]. Their results on the data were similar to ones obtained on more established test-collections such as TREC-GOV2. The data set contains 55k products with name, description and attribute fields. The attribute field contains additional information, such as notes and can also be empty.

We considered 1000 queries with the most relevance judgements available. The documents were judged by humans on a scale between 1 and 3. We defined 3 as relevant and anything under as non-relevant. All together there are 12093 judgements, 10260 relevant and 1833 non-relevant.

The table below shows the results of the experimentation, with a clear indication that the BM25-FIC is outperforming BM25F. As mentioned earlier, the field weights for the benchmarks are uniform, as no learning or heuristics are used on the BM25-FIC either. The relative improvements of in the table are calculated based on the better performing baseline; BM25F-micro.

	MAP	Δ MAP	P@10	Δ P@10	NDCG	Δ NDCG
Baseline $\text{BM25F}_{\text{macro}}$	0.218	-	0.219	-	0.496	-
Baseline $\text{BM25F}_{\text{micro}}$	0.232	-	0.220	-	0.499	-
BM25-FIC_{P1}	0.288	+24%	0.278	+27%	0.546	+10%
BM25-FIC_{P2}	0.290	+25%	0.281	+28%	0.547	+10%
BM25-FIC_{P3}	0.300	+29%	0.291	+33%	0.554	+11%

Out of the three candidate models BM25-FIC_{P3} is most accurate, consistently outperforming BM25F-micro and BM25F-macro for all metrics.

4 Field Weights for Defining User Query Intent

To better reflect the query intent we are enhancing the BM25-FIC model by using a seed document as reference point. The user picks a document from the initial results that corresponds to a type of query intent. This is a simplification of usual approaches to relevance feedback [**?**,**?**,**?**,**?**]. Specific to our model is the usage of the field weights.

The field weights reflect query intent, as each field contributes to relevance in a different way. By analysing the similarity between the seed document field weights and those of other documents, the model allows the user to prioritize or de-prioritize documents with similar query intent to that of the seed document.

[1] https://www.kaggle.com/c/home-depot-product-search-relevance/data

Definition 2 (Interactive Model).

$$\mathrm{RSV}_{\text{BM25-FIC,Inf,inter},a}(q,d,c,d_{\text{sd}}) := \mathrm{RSV}_{\text{BM25-FIC}}(q,d,c) + a \cdot S(q,d,c,d_{\text{sd}},\text{Inf})$$

where S is a similarity measure. This could be a retrieval model or the Euclidean distance: $S := 1 - ||\bar{w}_d(q,c,\text{Inf}) - \bar{w}_{d_{\text{sd}}}(q,c,\text{Inf})||_2$ and the normalized field weight vector for document d is defined as $\bar{w}_d(q,c,\text{Inf}) := (\frac{w_{f1}(q,c,\text{Inf})}{\sum_f w_f(q,c,\text{Inf})} \dots \frac{w_{fn}(q,c,\text{Inf})}{\sum_f w_f(q,c,\text{Inf})})$, where $f_i \in F_c$

The parameter a can be adjusted by the user. Positive values result in higher scores for documents that are relevant in a similar way to d_{sd}. Negative values do the opposite.

Using the seed document as a reference point and by adjusting a, the user can define their query intent in an intuitive manner. The idea is to make use of the fact that when presented with the initial search results, a user can often understand why some documents are high in the ranking, even though they might not correspond to their query intent. By using these documents as reference points, the user can easily refine and fine-tune their query intent.

As a more concrete example, consider the query "Wizards and magic J.K. Rowling 1995" used to search a book catalogue with the document fields *plot*, *author* and *publication year*. The fact that the first Harry Potter book was only published in 1997 poses questions about the users query intent. Are they looking for Harry Potter books, but got the year wrong? Or books about wizards from 1995 similar to J.K. Rowlings work? There are many possible ways — such as the two above — in which documents can be relevant. What the model from Definition 2 does is to help navigate these different dimensions of relevance: Say the user is not looking for Harry Potter books, but some of them come up in the search results. They can choose one of them as the seed document and by making a negative, other Harry Potter books would disappear from the top results. This is because the remaining results would be those with higher weights for publication year and lower ones for author. With a high positive a, they would see all Harry Potter books in the top results.

5 Conclusion

The main contribution of this paper is to introduce a new method for automatic field weighting in the BM25F retrieval model. We denote the model BM25-FIC, as it uses field information content (FIC) to calculate the document level field weight. Compared to basic BM25F (macro or micro), there is a relative improvement between 10% (NDCG) and 30% (MAP and P@10) for the Kaggle Home Depot data set.

Moreover, the paper introduces a new re-ranking retrieval model based on the BM25-FIC weights, which is a good candidate for interactive retrieval. The aim of the model is to give a user better tools for defining their query intent. This is done using a seed document as a positive or negative reference point for finding the desired dimensionality of relevance.

Overall, the research confirms that there is unexpected potential in the refinement of BM25F field weights, and that the "weighted sum" is also useful for other multi-dimensional measures than just document fields.

References

1. Balaneshinkordan, S., Kotov, A., Nikolaev, F.: Attentive Neural Architecture for Ad-hoc Structured Document Retrieval. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1173–1182. CIKM '18, Association for Computing Machinery, Torino, Italy (Oct 2018). https://doi.org/10.1145/3269206.3271801, https://doi.org/10.1145/3269206.3271801
2. Buckley, C.: Automatic Query Expansion Using SMART : TREC 3. In: In Proceedings of The third Text REtrieval Conference (TREC-3. pp. 69–80 (1993)
3. Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., van Rijsbergen, K.: Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In: Proceedings of the third symposium on Information interaction in context. pp. 115–124. IIiX '10, Association for Computing Machinery, New Brunswick, New Jersey, USA (Aug 2010). https://doi.org/10.1145/1840784.1840802, https://doi.org/10.1145/1840784.1840802
4. He, B., Ounis, I.: Setting Per-field Normalisation Hyper-parameters for the Named-Page Finding Search Task. In: Amati, G., Carpineto, C., Romano, G. (eds.) Advances in Information Retrieval. pp. 468–480. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2007)
5. Imhof, M., Braschler, M.: A study of untrained models for multimodal information retrieval. Information Retrieval Journal **21**(1), 81–106 (Feb 2018). https://doi.org/10.1007/s10791-017-9322-x, https://doi.org/10.1007/s10791-017-9322-x
6. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval **3**, 333–389 (Jan 2009). https://doi.org/10.1561/1500000019
7. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. pp. 42–49. CIKM '04, Association for Computing Machinery, Washington, D.C., USA (Nov 2004)
8. Rocchio: Relevance feedback in information retrieval. The SMART retrieval System Expedriments in Automatic Document Processing (1971)
9. Roelleke, T.: Information Retrieval Models: Foundations and Relationships. Morgan & Claypool Publishers (2013), google-Books-ID: SFavNAEACAAJ
10. Zellhöfer, D., Schmitt, I.: A user interaction model based on the principle of polyrepresentation. In: Proceedings of the 4th workshop on Workshop for Ph.D. students in information & knowledge management. pp. 3–10. PIKM '11, Association for Computing Machinery, Glasgow, Scotland, UK (Oct 2011). https://doi.org/10.1145/2065003.2065007, https://doi.org/10.1145/2065003.2065007