Formal Constraints for Structured Document Retrieval

Tuomas Ketola
t.j.h.ketola@qmul.ac.uk
Queen Mary, University of London
United Kingdom

Thomas Roelleke
t.roelleke@qmul.ac.uk
Queen Mary, University of London
United Kingdom

ABSTRACT

The formalization of retrieval constraints for traditional (atomic) retrieval was a major milestone in information retrieval (IR) research. The aim of these constraints was to formalize IR heuristics which most retrieval models rely upon. In a similar fashion, this paper introduces constraints for structured document retrieval (SDR). Out of the many possible constraints, we focus on three that are shown to produce intuitive rankings in simple, but informative retrieval scenarios. It is shown that none of the widely used SDR models (BM25F, MLM, linear score aggregation) satisfy all three constraints. The underlying reason for this is shown to be the failure of existing models to balance between assuming independence of term occurrences across fields and considering the documents as atomic, rather than structured. The constraints introduced in this paper, together with the analysis of how they are satisfied by existing models, can be used to analytically reason about the behaviour of any SDR model in a variety of ranking scenarios.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; **Document structure**.

KEYWORDS

retrieval models, structured document retrieval

ACM Reference Format:
Tuomas Ketola and Thomas Roelleke. 2022. Formal Constraints for Structured Document Retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, Madrid, Spain, 6 pages. https://doi.org/10.1145/3539813.3545128

1 INTRODUCTION

Analytical retrieval models, such as the BM25 and Language Modelling (LM), are used widely in commercial and academic settings. The behaviour of these models is understood well due to extensive research over the last 20+ years. One important line of enquiry has been formal retrieval constraints / axioms [4, 5]. The aim of our paper is to develop such framework for structured document retrieval (SDR). This is accomplished by identifying three constraints that define optimal ranking behaviour in simple, but informative scenarios, and by analysing how existing models adhere to those constraints.

Tab. 1 summarises the intuition underlying the three chosen constraints for SDR. In order to better understand the intuition underlying the constraints, consider the retrieval scenario in Tab. 2. Given similar rareness of terms and similar document lengths, in an intuitive ranking, d1 should be first as it contains both query terms, d2 should be second as it contains only one query term occurring in different fields, and d3 should be last as it contains only one query term in one field.

Constraint	Intuition
Field importance	A model should be able to boost, or decrease the weight given to a field-based on some notion of field importance
Field distinctiveness	Adding a query term to a new field should increase the retrieval score more than adding it to a field where it already occurs
Term distinctiveness	Adding unseen query terms to a document should increase the retrieval score more than adding query terms already considered

Table 1: Intuition underlying formal constraints for SDR. Field refers to a field of a document; e.g. *abstract* or *author*.

field	plot		description		flattened doc
term	english	spy	english	spy	
d1	1	0	0	1	english spy
d2	1	0	1	0	english english
d3	0	0	2	0	english english

Table 2: Example with two fields (movie plot and description) and two query terms (english and spy) illustrating how rankings by existing SDR models are not always intuitive.

The **first contribution** of this paper is to formalize retrieval constraints that guarantee this ranking. It is not the intention of this paper to claim that the described intuitive ranking behaviour is always the correct one, as this is defined by the user. Instead, our intention is to formulate constraints that produce an intuitive ranking where no knowledge of user preferences is available.

The **second contribution** is to analyse why and how widely used SDR models satisfy, or fail to satisfy the constraints. It will be shown that the underlying reasons have to do with how they model term frequency across different fields. Models that aggregate over field-based scores consider term frequency to be independent across fields, an assumption which was shown to be harmful by Robertson et al. amongst others [9, 10] and results in the Term Distinctiveness constraint not being satisfied ($\text{score}(d_1) = \text{score}(d_2)$ rather than

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545128

score(d_1) > score(d_2) in Tab. 2). On the other hand, models that aggregate the fields at the term frequency level and calculate scores over a flattened document, such as BM25F and Mixture of Language Models (MLM), consider the document as atomic (rather than structured) after term frequency weighting, meaning they fail to fulfil the Field Distinctiveness constraint (score(d_2) = score(d_3) rather than score(d_2) > score(d_3) in Tab. 2). Finally, there are models such as PRMS that fail to consider field importance in any way, thus failing to satisfy the Field Importance constraint.

The **third contribution** is to discuss how SDR models could be developed in the future to better satisfy the constraints. Our findings suggest that in order for an SDR model to accomplish this, it should be able to balance between saturating term frequency across fields, whilst still explicitly considering the document structure.

2 ATOMIC RETRIEVAL CONSTRAINTS

Fang et. al [4] introduced formal constraints (axioms) for (atomic) IR, to "capture retrieval heuristics, such as the TF-IDF, in a formal way, making it possible to apply them to any retrieval formula analytically". This paper does something similar, expect it is not heuristics we capture (what they consider heuristics has been formalized in many cases [1–3, 9]), but intuitive "rules", that SDR models should aim to follow. Out of the seven constraints by Fang et. al, the one that directly relates to our constraints is the second term frequency Constraint (TFC2) [4]. TFC2 consists of two constraints (here TFC2-1, TFC2-2). The first is written formally below (notation changed slightly to align with the one we apply).

Fang et. al TFC2-1: Let $q = \{t\}$ be a query with exactly one term t, S a scoring function, and $\mathrm{TF}(t, d)$ the term frequency of term t in document d. Assume $|d_1| = |d_2| = |d_3|$ and $\mathrm{TF}(t, d_1) > 0$.

$$\begin{aligned} &\text{if } \mathrm{TF}(t, d_2) - \mathrm{TF}(t, d_1) = 1 \text{ and } \mathrm{TF}(t, d_3) - \mathrm{TF}(t, d_2) = 1, \\ &\text{then } S(q, d_2) - S(q, d_1) > S(q, d_3) - S(q, d_2) \end{aligned} \tag{1}$$

This captures that the change of the retrieval score should be smaller if TF changes from 10 to 11 than from 1 to 2, i.e. the TF should be saturated. In other words, the second derivative of the scoring function with respect to term frequency should be negative: $\frac{\partial^2 S}{\partial \mathrm{TF}^2} < 0$.

Fang et. al point out that "the TFC2 constraint implies another desirable property — if two documents have the same total [number] of occurrences of query terms, a higher score will be given to the document covering more distinct query terms". However — as they point out — this is only true if the terms have the same IDF value.

Fang et. al TFC2-2: Let t_1 and t_2 occur in a query q. Assume $|d_1| = |d_2|$ and $\mathrm{IDF}(t_1, c) = \mathrm{IDF}(t_2, c)$.

$$\begin{aligned} &\text{if } \mathrm{TF}(t_1, d_1) = \mathrm{TF}(t_1, d_2) + \mathrm{TF}(t_2, d_2) \text{ and} \\ &(\mathrm{TF}(t_2, d_1) = 0, \mathrm{TF}(t_1, d_2) \neq 0, \mathrm{TF}(t_2, d_2) \neq 0), \\ &\text{then } S(q, d_2) > S(q, d_1) \end{aligned} \tag{2}$$

The term specificity ratio $\mathrm{IDF}(t_1)/\mathrm{IDF}(t_2)$ affects whether one wants the TFC2-2 constraint to hold. The marginal contribution of a term appearing again in a document depends on the way in which the TFC2-1 constraint by Fang et. al is satisfied, i.e. the degree to which term frequency is saturated. Different retrieval models saturate term frequency in different ways. This means their inherent specificity ratios for satisfying TFC2-2 are different.

3 STRUCTURED RETRIEVAL CONSTRAINTS

Regarding the example in Tab. 2, the following constraints lead to an intuitive ranking: document d1 should be ranked first because it contains both query terms, d2 should be second because the one query term appears twice and in different fields, and d3 should be last because the one query term occurs twice in the same field.

This "intuitive ranking" does not necessarily represent the "correct ranking", as this is ultimately judged by the user. For example the user might be more interested in the *description* field, in which case it might make sense to rank d3 higher than d2. However, lacking this kind of knowledge of user preferences, the ranking behaviour described above does correspond to two intuitive rules: 1. documents with many distinct query terms should rank higher than those with few, and 2. documents where a query term occurs in several fields should rank higher than if the term occurs only in few fields.

Constraint 1: Field Importance. Let Q denote a query, S a retrieval score, and d a document, where $d = \{t_{a,f1}, t_{a,f2}, \ldots, t_{b,f1}, \ldots\}$, and $t_{a,fi}$ term "a" that occurs in field f_i. $I(f_i)$ is the importance of f_i.

$$\begin{aligned} &\forall Q, d, f_i, t_a : \text{ if } t_a \in Q \text{ and } I(f_1) > I(f_2), \\ &\text{then } S(Q, d \cup \{t_{a,f1}\}) > S(Q, d \cup \{t_{a,f2}\}) \end{aligned} \tag{C1}$$

In other words, adding a query term to a field with a greater importance must increase the score more than adding one to a field with lower importance. This might seem trivial, but the point being made is that an SDR model should be able weight fields based on some notion of importance. This weighting can be done through learning field weights, or using heuristics for example.

Constraint 2: Field Distinctiveness. Let Q, d, f_i and t be defined they were for Constraint C1. Field importance is uniform across all fields.

$$\begin{aligned} &\forall Q, d, f_i, t_a : \text{ if } t_a \in Q \text{ and } \; m > n, \\ &\text{then } S(Q, d \cup \{t_{a,f1} \ldots t_{a,fm}\}) > S(Q, d \cup \{t_{a,f1} \ldots t_{a,fn}\}) \end{aligned} \tag{C2}$$

In other words, the more fields a query term appears in, the higher the ranking score of the document should be. This constraint also implies that adding a query term to a new field of a document should increase the ranking score more than adding a query term to field where it already appears. For the documents in Tab. 2, this would mean that document d2 ranks higher than d3.

The order of $\{t_{a,f1} \ldots t_{a,fm}\}$ does not refer to the order of the fields in the documents, but the order in which query term t_a occurs in them, meaning f_1 is not the first field of the document, but the first field in which t_a occurs.

Constraint 3: Term Distinctiveness. Let Q, d, f_i and I be defined as they were for Constraint C2 and term t_i can occur in any of the document fields.

$$\begin{aligned} &\forall Q, d, t_i : \text{ if } t_i \in Q \text{ and } \; z > y, \\ &\text{then } S(Q, d \cup \{t_1, \ldots, t_z\}) > S(Q, d \cup \{t_1, \ldots, t_y\}) \end{aligned} \tag{C3}$$

I.e. adding many distinct query terms to a document should increase the score more than adding a few, no matter in which fields they appear. For the documents in Tab. 2 this would mean that document d1 ranks higher than d2.

The satisfaction of this constraint is central to the BM25F retrieval model, which is discussed in detail later. By saturating term frequency across fields, the BM25F gives more importance to the first occurrence of a query term, compared to subsequent occurrences of the same term, wherever in the document they occur [10]. By doing so, it puts more emphasis on a document having many distinct query terms, rather than few. This logic is one of the central aspects of the BM25F, which has been shown to outperform FSA-based models [10, 13].

In essence Constraint C3 is communicating a similar issue as constraint TFC2 in [4]. However, it is worth re-formalizing it for SDR because 1. it will be shown that many SDR models do not satisfy it, whereas in atomic retrieval this is not common, and 2. its implications are more severe for SDR, as term frequencies are often inflated through field weights. Furthermore, instead of saying that the constraint implies a property where documents with more distinct query terms will get a higher score (as done in [4]), we analyse where and why this is the case for SDR.

4 EXISTING MODELS AND CONSTRAINTS

4.1 Existing SDR Models

The SDR models considered in this paper do not represent all models found in literature, instead we focus on widely used analytical, non-domain specific approaches. These models include field score aggregation, the BM25F, Mixture of Language Models (MLM), Fielded Sequential Dependence Model (FSDM) and the Probabilistic Retrieval Model for Semistructured Data (PRMS). These models have been chosen because they are analytical; meaning it is easy to see how they behave with respect to the constraints and they can be used without training data.

All the models here are characterized by one of two underlying aggregation functions: 1. **Field Score Aggregation** (FSA), where the weights are applied over field-based retrieval scores (e.g. meta-search), or 2. **Term Frequency Aggregation** (TFA), where the weights are applied to within-field term frequencies and the score is calculated over a flattened document representation (e.g. BM25F and MLM).

FSA-Models score documents based on a weighted sum of their field-based retrieval scores. These models are closely related to the field of meta-search where the scores of different search engines are aggregated to a single ranking [10].

$$\mathrm{RSV}_{\mathrm{FSA,M}}(d,q,c) := \sum_{i=1}^{m} w_{f_i} \sum_{t \in q \cap d} \mathrm{S}_M(t, f_i, F_i) \tag{3}$$

where d is a document, q a query, c a collection, f a document field (e.g. title of document), F_i a collection field (all titles in collection c), m the number of fields, M is any single field retrieval model, w_{f_i} is the field weight and S_M is the scoring function for M.

BM25F was introduced in [10]. It allows for TF saturation across fields by applying the weights to TFs and calculating the retrieval score over a flattened document representation. This makes the underlying aggregation function TFA. This paper considers the version of BM25F introduced in [13], where the document length normalization is performed separately for each field.

$$\mathrm{RSV}_{\mathrm{BM25F},k_1,b}(q,d,c,r,\bar{r}) := \sum_{t \in q} \frac{n_{\vec{w}}(t,d)}{k_1 + n_{\vec{w}}(t,d)} w_{\mathrm{RSJ}}(t,c,r,\bar{r}) \tag{4}$$

Where $n_{\vec{w}}(t,d)$ is the weighted sum of document length normalized term frequencies and k_1 is the term frequency saturation hyperparameter, usually set between 1.2 and 2.0 [11]. w_{RSJ} is the Robertson-Sparck-Jones weight which in the absence of relevance information is the IDF [8, 9]: $w_{\mathrm{RSJ}}(t,c,r,\bar{r}) = \mathrm{IDF}(t,c)$.

Mixture of Language Models (MLM) calculates the retrieval score by applying weights over field-based language models, summing the resulting probabilities together and taking their product over the query terms [7]. MLM is closer to the BM25F than it is to FSA. This is because the field weights are incorporated into the model explicitly, meaning the weights are applied over the term frequency rather than a field-based score. This means that the underlying aggregation method for MLM is the same as BM25F, i.e. TFA. **The Fielded Sequential Dependence Model** (FSDM) by [14] incorporates MLM and therefore behaves exactly the same as MLM in terms of the constraints. **Probabilistic Retrieval Model for Semistructured Data** (PRMS) uses the probability of query terms appearing in fields for better mapping the two [6]. The underlying aggregation function of PRMS is FSA. Other approaches to SDR include [12], where context-specific frequencies (e.g. InvTitleFreq, InvSectionFreq) have been explored, and this will be considered in future research for the constraints.

4.2 Constraints

Tab. 3 shows which SDR model satisfies which constraints. **Conditional** satisfaction of a constraint denotes a case in which underlying collection statistics need to be considered, e.g. rareness of the query terms. This means that in order for a model to satisfy (or not) a constraint unconditionally, the query term rareness (IDF-values) should be assumed to be the same across fields and terms.

	Constraint 1 Field Import.	Constraint 2 Field Distinct.	Constraint 3 Term Distinct.
FSA	YES	Conditional	NO
PRMS	NO	Conditional	NO
BM25F	YES	NO	Conditional
MLM	YES	NO	Conditional
FSDM	YES	NO	Conditional

Table 3: Constraint satisfaction of SDR models: Conditional means that collection statistics need to be considered.

All of the retrieval models, apart from PRMS, have field weights that need to be estimated in order for the model to perform optimally. Usually this is done either heuristically, or using supervised learning. In terms of constraint satisfaction we do not consider field weighting. So for example, even though we could get BM25F to rank d_2 higher than d_3 by boosting the *plot* field in Tab. 2, we do not consider this as an aspect of the model itself, but rather user intervention. In order for a model to analytically satisfy a constraint it needs to do so without interference. This means that we must use "default", or analytically defined values for the hyperparameters and the field weights should be assumed to be uniform.

4.2.1 ***Constraint 1.*** Constraint 1 is the easiest to satisfy. As long as the model is able to give weight to fields based on their importance this constraint is satisfied. For FSA, BM25F, MLM and FSDM this can be done through field weighting. However, for PRMS this is not possible as the weight is not based on the importance of a field, but on how each query term is mapped to it.

4.2.2 ***Constraint 2.*** TFA models (MLM, BM25F, FSDM) do not satisfy Constraint C2. After applying the field weights at the term level, they consider the document as atomic. This issue is obvious in the retrieval scenario in Tab. 2 for the ranking of documents d2 and d3: Assuming equal IDF-values, it does not matter whether *english* appears twice in *description*, or once in *plot* AND once *description*, the documents get the same rank-score.

Satisfying Constraint C2 is conditional for the FSA-based models. The following will explain this conditionality in the general case for FSA-BM25 (Eqn. 3 with M = BM25), after which we will discuss how the general case can be simplified to capture the conditionality of satisfying Constraint C2 in a more intuitive way.

Definition 1 (Cross-Field IDF Ratio). *The ratio of the IDF-values between fields j and i for term t is denoted* IDF-CF-Rat(t, F_j, F_i).

$$\text{IDF-CF-Rat}(t, F_j, F_i) := \frac{\text{IDF}(t, F_j)}{\text{IDF}(t, F_i)} \tag{5}$$

Definition 2 (Cross-Field IDF Ratio Threshold). *Let $q = \{t_1, \ldots, t_n\}$ be a query, d a document with T occurrences of term t in field f_i and z occurrences of term t in another field f_j. Let $\overline{d}$ be an amended version of d, where the occurrences of term t in f_j have been moved to f_i and z occurrences of non-query terms have removed from f_i and added to f_j. These non-query terms ensure that* IDF-CF-Rat *is only concerned with query term occurrences, rather than document lengths.*

$$\text{IDF-CF-Rat}_{\text{th}}(t, F_i, F_j, k_1) := \frac{w_i}{w_j} \frac{\frac{T+z}{k_1+T+z} - \frac{T}{T+k_1}}{\frac{z}{z+k_1}} \tag{6}$$

IDF-CF-Rat$_{\text{th}}(t, F_i, F_j, k_1)$ defines the threshold for IDF-CF-Rat(t, F_j, F_i) above which Constraint C2 is satisfied, meaning $\text{RSV}_{\text{FSA,M}}(d, q, c) > \text{RSV}_{\text{FSA,M}}(\overline{d}, q, c)$. See Appendix A for formal theorem and proof.

In order to understand how FSA-BM25 satisfies the constraints more intuitively and in terms of Tab. 2, we assume uniform field weights and set $T = 1$ and $z = 1$ (see documents d2 and d3 in the example). This simplifies Eqn. (6) to:

$$\text{IDF-CF-Rat}_{\text{th}}(t, F_i, F_j, k_1) = \frac{2k_1 + 2}{2 + k_1} - 1 \tag{7}$$

Meaning Constraint C2 is satisfied by the FSA models as long as the ratio of the the highest and lowest IDF-value for all terms is greater than $\frac{2k+2}{2+k_1} - 1$. This would be likely if the two fields are correlated in their content, as we would expect similar IDFs for a given term in both fields. The above analysis has focused on the BM25, however FSA models can be used with any retrieval function. A similar analysis on LM would focus on the hyperparameter μ and the background model.

4.2.3 ***Constraint 3.*** FSA models do not satisfy Constraint C3. This is because the field-based scores are summed together, with no regard to whether both query terms (*english* AND *spy*) occur, or only one of them. The issue is evident from Tab. 2. Assuming equal specificity weights (e.g. IDF), d_1 and d_2 are rank equal. Intuitively we want documents with more query terms to rank higher. The problem comes from the fact that FSA assumes term frequency to be independent across fields for a given term, thus "double accounting" the occurrence of *english*. TFA solves this by saturating term frequencies across the fields, i.e. it assumes a constant dependency of term occurrences between fields for a given term. It has been shown that this significantly increases the robustness of the models and makes them less noisy [10, 13].

The satisfaction of Constraint C3 is conditional for the TFA-based models. They suffer from the same issue as atomic models where it comes to specificity ratio of query terms as discussed in Section 2. There exists a threshold for the ratio of IDF-values between query terms at which a second occurrences of a query term can dominate over the first occurrence of another query term.

The following will explain this conditionality in the general case for BM25F, after which we will discuss how the general case can be simplified to capture the conditionality of satisfying Constraint C3 in a more intuitive way.

Definition 3 (Cross-Term IDF Ratio). *The ratio of the IDF values between terms b and a is denoted* IDF-CT-Rat(t_a, t_b, c).

$$\text{IDF-CT-Rat}(t_a, t_b, c) := \frac{\text{IDF}(t_b, c)}{\text{IDF}(t_a, c)} \tag{8}$$

Definition 4 (Cross-Term IDF Ratio Threshold). *Let $q = \{t_1, \ldots, t_n\}$ be a query, d a document with T occurrences of term t_a in field f_i and z occurrences of term t_b in another field f_j. Let $\overline{d}$ be an amended version of d, where the occurrences of term t_b in f_j have been replaced by occurrences of t_a.*

$$\text{IDF-CT-Rat}_{\text{th}}(t_a, t_b, c, k_1) := \frac{\frac{w_i T + w_j z}{k_1 + w_i T + w_j z} - \frac{w_i T}{k_1 + w_i T}}{\frac{w_j z}{k_1 + w_j z}} \tag{9}$$

IDF-CT-Rat$_{\text{th}}(t_a, t_b, c, k_1)$ defines the threshold for IDF-CT-Rat(t_a, t_b, c) above which score(d) > score$(\overline{d})$. See Appendix B for formal theorem and proof

In order to understand how BM25F satisfies the constraints more intuitively and in terms of Tab. 2, we assume uniform field weights and set $T = 1$ and $z = 1$ (see documents d1 and d2 in the example). This simplifies Eqn 9 to:

$$\text{IDF-CT-Rat}_{\text{th}}(t_a, t_b, c, k_1) = \frac{2k_1 + 2}{k_1 + 2} - 1 \tag{10}$$

Eqn. (10) shows that whether the BM25F satisfies Constraint C3 depends on the ratio of the IDFs and the term frequency saturation parameter k_1. If $k_1 = 2.0$ the ratio of IDF values below which BM25F would fail to satisfy Constraint C3 equals 0.5. So if the rarest term of the query has an IDF twice the size of the most common term, the constraint is not satisfied. There are cases where it makes sense for a model to not satisfy Constraint C3, for example if the common term is a stopword. The IDF value for stopwords tends to be very close to 0, so the constraint is obviously not satisfied, nor should it be. However, a term can easily have half the IDF of another and

still be important, so the conditionality of Constraint C3 should be considered analytically. This issue is present in both SDR and atomic retrieval. The following discusses how it might be more severe for SDR, due to field weighting.

Consider a scenario where $k_1 = 2.0$ and the occurrences of t_a for $\overline{d}$ in f_j occur in third field f_k. The field weights are $w_{f_i} = 1$, $w_{f_j} = 1$ and $w_k = 3$ Maybe f_k is a *title* of the book and we wish to boost it compared to the *abstract* and *body* for example. In such a situation an occurrence of the new term t_b in field f_j would have the same effect on score, as a second occurrence of t_a in f_k, even if $\text{IDF}(t_a) = \text{IDF}(t_b)$, i.e. Constraint C3 would not be satisfied even if the terms had the same IDF.

The key take away here is that when heuristically boosting fields because they are important — say the title of a book — other hyperparameters should be considered as well. In order for the field boosting to work, it is therefore likely that all the parameters have to be optimised using supervised learning of some form.

5 DISCUSSION

The key points in this section are 1. the trade-offs between constraints C2 and C3, 2. the relationship between $\text{IDF-CF-Rat}_{\text{th}}(t, F_i, F_j, k_1)$ and $\text{IDF-CT-Rat}_{\text{th}}(t, F_i, F_j, k_1)$, 3. query-type and domain considerations, and 4. what an SDR model that satisfies all three constraints would look like.

Tab. 3 illustrates the trade-off between Field Distinctiveness and Term Distinctiveness. Models that satisfy Constraint C2 do not satisfy Constraint C3, and vice versa. As discussed by Robertson et al. FSA-based models assume independence of term frequencies across fields [8]. Regarding the example in Tab. 2, this means that it does not matter whether a document has both the query words "english" AND "spy", or just "english" spread over two fields, meaning the Term Distinctiveness constraint is not satisfied. TFA-based models solve this problem by saturation term frequency across fields. They assume a constant level of dependence between term occurrences in different fields, defined by their underlying scoring-functions. However, in doing so they have to consider the document as atomic, rather than structured. In terms of the example in Table 2, this means that it does not matter whether a document has occurrences of "english" in the *plot* AND *description* fields, or just *description*, meaning the Field Distinctiveness Constraint is not satisfied.

Appendices A and B analyse the conditions for FSA-based models satisfying Constraint C2 and TFA-based models satisfying Constraint C3. For FSA, the key metric to knowing whether a constraint is satisfied is the cross-field IDF ratio (Def. 1) and for TFA the cross-term IDF ratio (Def. 3). For each of these, there exists a threshold above which the Constraints C2 and C3 are satisfied, respectively. If we simplify Eqns.(6) and (9) assuming the term frequencies from Table 2 and uniform field weights we get the simplified threshold values presented in Eqns. (7) and (10). Interestingly we observe that

$$\text{IDF-CF-Rat}_{\text{th}}(t, F_i, F_j, k_1) = \text{IDF-CT-Rat}_{\text{th}}(t_a, t_b, c, k_1) \qquad (11)$$

meaning the cross-field IDF ratio threshold for satisfying Constraint C2 for FSA is equal to the cross-term IDF ratio threshold for satisfying Constraint C3 for TFA. For FSA the ratio is defined for a given term and for TFA between different terms. Whether each of the models satisfy their respective constraints depends on underlying collection statistics and the query. For example, if the query includes terms that have very different IDF-values across fields (e.g. terms with different meanings in different fields), FSA models might not satisfy Constraint C2. Or, if the query terms have very different IDF-values (some very rare and some common), TFA models might not satisfy Constraint C3. Which one is more likely, depends on the nature of the retrieval scenario. For example, in a QA retrieval scenario, it is likely that the query contains stopword-like terms. In such cases, not satisfying Constraint C3 fully could be desirable. For keyword-like queries the opposite is likely to be true. Not satisfying Constraint C2 is more harmful in scenarios where there are many fields that carry different kinds of information, rather than scenarios with redundant, or very similar fields.

The analysis in this paper suggests that in order for an SDR model to satisfy the constraints defined in Section 3 (even conditionally), the model would need to facilitate term frequency saturation across fields (unlike the FSA), but should not revert to considering documents atomic (unlike the TFA). Furthermore, the model should consider the findings in Theorems 1 and 2, i.e. analytically asses the term specificity ratios at which the constraints are satisfied.

6 CONCLUSION

Analytical retrieval models for atomic data, such as BM25 and LM, are used widely across business and academia. Their behaviour is well understood due to extensive research. The objective of this paper has been to expand this kind of understanding to SDR, through formal retrieval constraints and an analysis of how existing models satisfy them.

The first contribution of this paper was to identify essential retrieval constraints for SDR, and to demonstrate that they are sufficient for achieving ranking behaviour that can be considered intuitive (in the absence of user preferences). The work has been inspired by the formal constraints for atomic document retrieval [4, 5]. Their aim was to formalise constraints to capture "information retrieval heuristics". Our aim is to capture "intuitive rules" that an SDR model should follow. The intuition underlying constraints is presented in Table 1, and the formal definitions are in Section 3.

Section 4 demonstrated how widely used analytical SDR models (BM25F, MLM, FSDM, field score aggregation) satisfy, or fail to satisfy the constraints. It is shown that there exists a trade-off between two of the constraints, one which focuses on the number of distinctive query terms in a document (Term Distinctiveness) and one which focuses on the number of fields a given query term appears in (Field Distinctiveness). The theorems in Appendices A and B demonstrate how collection and query statistics affect the constraint satisfaction of existing models. This allows us to analytically discuss model behaviour in various contexts, including term rarity measures, hyperparmeters and the nature of the retrieval scenario.

Our findings suggest that in order for an SDR model to satisfy all three constraints, one of the main challenges is to balance between saturating term frequency across fields, whilst still considering the document structure explicitly throughout.

The formal constraints introduced in this paper and the analysis that has followed, can be used as a framework for analytical reasoning of ranking behaviour in future SDR research.

ACKNOWLEDGEMENTS

We would like to acknowledge the suggestions of the ICTIR reviewers regarding notation and the consideration of user preferences.

REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (Jan. 2003), 45–65.
[2] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems* 20, 4 (Oct. 2002), 357–389.
[3] Ronan Cummins, Jiaul H. Paik, and Yuanhua Lv. 2015. A Polya Urn Document Language Model for Improved Information Retrieval. *ACM Transactions on Information Systems* 33, 4 (May 2015), 21:1–21:34.
[4] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics *(SIGIR '04)*. ACM, New York, NY, USA, 49–56.
[5] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *SIGIR 2005*. 480–487.
[6] Jinyoung Kim, Xiaobing Xue, and W. Bruce Croft. 2009. A Probabilistic Retrieval Model for Semistructured Data. In *Advances in Information Retrieval*. Springer, Berlin, Heidelberg, 228–239.
[7] Paul Ogilvie and Jamie Callan. 2003. Combining document representations for known-item search *(SIGIR '03)*. ACM, New York, NY, USA, 143–150.
[8] Stephen Robertson. 2004. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation* 60, 5 (Jan. 2004), 503–520. Publisher: Emerald Group Publishing Limited.
[9] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (Dec. 2009), 333–389.
[10] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields *(CIKM '04)*. ACM, Washington, D.C., USA, 42–49.
[11] Thomas Roelleke. 2013. *Information Retrieval Models: Foundations and Relationships*. Morgan & Claypool Publishers.
[12] Jun Wang and Thomas Roelleke. 2006. Context-Specific Frequencies and Discriminativeness for the Retrieval of Structured Documents. In *ECIR, London*.
[13] Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. 2004. Microsoft Cambridge at TREC–13: Web and HARD tracks. (2004).
[14] Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. 2015. Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data *(SIGIR '15)*. ACM, New York, NY, USA, 253–262.

A CROSS-FIELD IDF RATIO THRESHOLD

The underlying idea of the cross field idf-ratio theorem is that there exists a threshold for IDF-CF-Rat(t, F_i, F_j) below which Constraint C2 is not satisfied by FSA-BM25, meaning the documents d_2 and d_3 in Table 2 would be ranked incorrectly.

Theorem 1 (FSA and the Field Distinctiveness Constraint). *Let $q = \{t_1, \ldots, t_n\}$ be a query, d a document with T occurrences of term t in field f_i and z occurrences of term t in another field f_j. Let $\overline{d}$ be an amended version of d, where the occurrences of term t in f_j have been moved to f_i and z occurrences of non-query terms have removed from f_i and added to f_j.*

$$\begin{aligned}&\forall\ t \text{ and } (F_i, F_j) \in q \cap d :\\ &\text{IDF-CF-Rat}(t, F_i, F_j) > \text{IDF-CF-Rat}_{\text{th}}(t, F_i, F_j, k_1)\\ &\implies \text{RSV}_{\text{FSA,M}}(q, d, c) > \text{RSV}_{\text{FSA,M}}(q, \overline{d}, c)\end{aligned} \tag{12}$$

Proof. Following Definition 1 the threshold for satisfying Constraint C2 becomes

$$\frac{\text{IDF}(t, F_j)}{\text{IDF}(t, F_i)} > \frac{w_i}{w_j} \frac{\frac{T+z}{k_1+T+z} - \frac{T}{T+k_1}}{\frac{z}{z+k_1}} \tag{13}$$

$$w_i \frac{T}{T+k_1} \text{IDF}(t, F_i) + w_j \frac{z}{z+k_1} \text{IDF}(t, F_j) > w_i \frac{T+z}{k_1+T+z} \text{IDF}(t, F_i) \tag{14}$$

The BM25 retrieval status value of field f is calculated as

$$\text{RSV}_{\text{BM25},k_1,b}(q, f, F) := \sum_{t \in q} \frac{n_{\text{norm}}(t, f, b_f)}{k_1 + n_{\text{norm}}(t, f, b_f)} \text{IDF}(t, F) \tag{15}$$

Since $|d| = |\overline{d}|$, the ranking of the documents, i.e. the inequality of the scores is not affected by document length normalisation. Therefore we can set $n_{\text{norm}}(t, f, d) = n(t, f, d)$ in Eqn. (15) without changing the analysis. Assuming the term frequencies from the theorem, and following Eqn. (15) we can re-write Eqn. (14) as

$$\text{RSV}_{\text{FSA,M}}(q, d, c) > \text{RSV}_{\text{FSA,M}}(q, \overline{d}, c) \tag{16}$$

□

B CROSS-TERM IDF RATIO THRESHOLD

The underlying idea of the cross term idf-ratio theorem is that there exists a threshold for IDF-CT-Rat(t_a, t_b, c) below which Constraint C3 is not satisfied by BM25F, meaning the documents d_1 and d_2 in Table 2 would be ranked incorrectly.

Theorem 2 (BM25F and the Term Distinct. Constraint). *Let $q = \{t_1, \ldots, t_n\}$ be a query, d a document with T occurrences of term t_a in field f_i and z occurrences of term t_b in another field f_j. Let $\overline{d}$ be an amended version of d, where the occurrences of term t_b in f_j have been replaced by occurrences of t_a.*

$$\begin{aligned}&\forall (t_a, t_b) \in q \cap d :\\ &\text{IDF-CT-Rat}(t_a, t_b, c) > \text{IDF-CT-Rat}_{\text{th}}(t_a, t_b, c, k_1)\\ &\implies \text{RSV}_{\text{BM25F}}(q, d, c) > \text{RSV}_{\text{BM25F}}(q, \overline{d}, c)\end{aligned} \tag{17}$$

Proof. Following Definition 3 the threshold for satisfying Constraint C3 becomes

$$\frac{\text{IDF}(t_b, c)}{\text{IDF}(t_a, c)} > \frac{\frac{w_i T + w_j z}{k_1 + w_i T + w_j z} - \frac{w_i T}{k_1 + w_i T}}{\frac{w_j z}{k_1 + w_j z}} \tag{18}$$

$$\frac{w_i T}{k_1 + w_i T} \text{IDF}(t_a, c) + \frac{w_j z}{k_1 + w_j z} \text{IDF}(t_b, c) > \frac{w_i T + w_j z}{k_1 + w_i T + w_j z} \text{IDF}(t_a, c) \tag{19}$$

Since $|d| = |\overline{d}|$, the ranking of the documents, i.e. the inequality of the scores is not affected by document length normalisation. Therefore we can set $n_{\text{norm}}(t, f, d) = n(t, f, d)$ in Eqn. (4) without changing the analysis. Assuming the term frequencies from the theorem and following Eqn. (4) we can re-write Eqn (19) as

$$\text{RSV}_{\text{BM25F},k_1,b}(q, d, c) > \text{RSV}_{\text{BM25F},k_1,b}(q, \overline{d}, c) \tag{20}$$

□