Inf Retrieval J
DOI 10.1007/s10791-017-9322-x

A study of untrained models for multimodal information retrieval

Melanie Imhof[1,2] · **Martin Braschler**[2]

Received: 22 December 2016 / Accepted: 23 October 2017

Abstract Operational multimodal information retrieval systems have to deal with increasingly complex document collections and queries that are composed of a large set of textual and non-textual modalities such as ratings, prices, timestamps, geographical coordinates, etc. The resulting combinatorial explosion of modality combinations makes it intractable to treat each modality individually and to obtain suitable training data. As a consequence, instead of finding and training new models for each individual modality or combination of modalities, it is crucial to establish unified models, and fuse their outputs in a robust way. Since the most popular weighting schemes for textual retrieval have in the past generalized well to many retrieval tasks, we demonstrate how they can be adapted to be used with non-textual modalities, which is a first step towards finding such a unified model. We demonstrate that the popular weighting scheme BM25 is suitable to be used for multimodal IR systems and analyze the underlying assumptions of the BM25 formula with respect to merging modalities under the so-called raw-score merging hypothesis, which requires no training. We establish a multimodal baseline for two multimodal test collections, show how modalities differ with respect to their contribution to relevance and the difficulty of treating modalities with overlapping information. Our experiments demonstrate that our multimodal baseline with no training achieves a significantly higher retrieval effectiveness than using just the textual modality for the social book search 2016 collection and lies in the range of a trained multimodal approach using the optimal linear combination of the modality scores.

✉ Melanie Imhof
imhf@zhaw.ch

Martin Braschler
bram@zhaw.ch

[1] Université de Neuchâtel, Neuchâtel, Switzerland

[2] Zurich University of Applied Sciences, Winterthur, Switzerland

Published online: 03 November 2017

Keywords Multimodal information retrieval · BM25 · Raw-score merging hypothesis

1 Introduction

The academic discipline that we term today “information retrieval” (IR) goes back, though opinions vary, to at least the seminal position paper by Bush (1945). In the ensuing roughly 70 years of work, some mechanisms have been introduced early on, but have persisted and proven versatile since then; e.g. the formulae that govern the ranking of retrieved documents. Amongst these are some of the most popular weighting schemes for (textual) retrieval, which can all be described in terms of how they combine three main components; the term frequency (*tf*); i.e. how often a term appears in a given document, the document frequency (*df*); i.e. in how many documents a term appears and a document length normalization component. Originally developed for retrieval on English language text, these weighting schemes have generalized well to many related tasks, such as multilingual retrieval (Peters et al. 2012), multimedia retrieval (Müller et al. 2010) and others.

Today, we have to deal with increasingly complex document collections and queries (Imhof and Braschler 2015) that no longer just consist of textual modalities but also of a large set of non-textual modalities such as visual words in image retrieval (Villegas et al. 2015), locations in geographical IR (Mandl et al. 2009) or timestamps in time-aware IR (Li and Croft 2003). This is particularly true in enterprise search, domain-specific IR and many real IR applications, where it is not an option to simply ignore or discard entire modalities. Therefore, we claim that it becomes crucial to treat the modalities with unified methods instead of finding new approaches for each new modality or train a new model for every combination of modaltities. In this paper, we discuss the underpinnings of weighting schemes for textual retrieval and show how they can be applied or adapted methodically to non-textual modalities, such as ratings of books and geographical coordinates, which we understand as the first step into finding a unified model.

As a contribution towards establishing best practices for the integration of many modalities into an IR application, we demonstrate that BM25 is a suitable weighting scheme outperforming its alternatives to be used on non-textual modalities and to merge them under the so-called raw-score merging hypothesis by checking the assumptions underlying the BM25 formula. Being able to merge the modalities under the raw-score merging hypothesis with little or no training is particularly important due to the limited generalizability of suitable test collections and training data.

We start by considering an “ideal” robust approach, which is based on term sampling in order to correct the differences in average document length, which is one of the most obvious collection statistics. Then, we prove that there are cases, where BM25 can be interpreted as being identical to this sampling based approach. Using the sampling approach, we can further correct the difference between the variance of the document lengths. Along the investigation of the sampling approach, we further analyze the *tf* saturation parameter k_1 of BM25 and explain its significance for non-textual modalities. Finally, we present experiments on the effectiveness of merging the results of the individual modalities into a unified multimodal result. We contrast our approach, which avoids learning, with an “optimized” baseline and find encouraging results.

The remainder of this paper is structured as follows. Section 2 outlines the anatomy of multimodal IR systems and describes the challenges faced when dealing with complex multimodal collections. We then demonstrate that BM25 is a suitable weighting scheme in multimodal IR systems w.r.t. document length normalization (Sect. 4). Section 5 describes how BM25 can be used for non-textual modalities by redefining the three main components of the weighting scheme. A sampling based BM25 approach is proposed in Sect. 6, which allows us to prove that BM25 fulfills the raw-score merging hypothesis w.r.t. the average document length and the variance of document lengths. In Sect. 7, we describe the multimodal test collections that we use for evaluation, followed by the experiments and the discussion of their results. Section 8 concludes this paper and discusses future work.

2 The anatomy of a multimodal IR system

2.1 Anatomy

In a multimodal IR system, both the documents as well as the queries consist of several modalities. Figure 1 shows an explanatory excerpt of four of the modalities of the documents in the social book search (SBS) collection used in the SBS lab at the CLEF evaluation forum (Koolen et al. 2016). The documents $(d_1, d_2, \ldots, d_D)$ consist of the modalities: book title, reviews, binding and ratings, each of which can be treated as a bag of features. Hereby, d_j^m is the bag of features of modality m of document d_j. The query both contains explicit and implicit modalities; i.e. the textual description of the request is explicit, while other information such as acceptable languages and ratings of the books are implicit. A more detailed description of the collection is given in Sect. 7.1.2. The queries in the SBS task are not particularly complex. In general, information needs embed several implicit and explicit modalities.

During retrieval, weighting schemes define the retrieval score (retrieval status value $\mathrm{RSV}(q, d_j^m)$) of modality m of document d_j w.r.t. query q. The retrieval scores allow producing a ranked list for each modality according to the estimated probabilities of relevance, although the retrieval scores are not necessarily probability values but are order-preserving w.r.t the probabilities of relevance Robertson and Zaragoza (2009). These ranked lists of all the modalities, similarly to multilingual retrieval, need to be merged into a single ranked list. Hence, a function f has to be found to compute the retrieval score for each document including the retrieval scores of all modalities

$$\mathrm{RSV}(q, d_j) = f(\mathrm{RSV}(q, d_j^1), \mathrm{RSV}(q, d_j^2), \ldots, \mathrm{RSV}(q, d_j^M)), \tag{1}$$

where M is the number of modalities.

Evaluation has a strong tradition in IR, since information is hard to be defined in general (Cleverdon 1967). A crucial part of an IR evaluation is the availability of a suitable test collection. However, most of the existing test collections are not representative for multimodal IR systems and it is clearly not practical to create a test collection that covers all possible modalities and their combinations (Imhof and Braschler 2015).

1:	Title	$d_j^1 = \{$Skylar, in, Yankeeland$\}$
2:	Reviews	$d_j^2 = \{$Delightful, The, is, the, best, McDonald, has, done, in, a, decade$\}$
3:	Ratings	$d_j^3 = \{5,1,3\}$
4:	Binding	$d_j^4 = \{$Hardcover$\}$

Fig. 1 Excerpt of four modalities of a sample document (denoted d_j) in the SBS collection

We are convinced that in order to improve and broaden the applicability of multimodal IR, a generalizable method to deal with complex collections with a large amount of very different modalities is crucial. Therefore, we claim that we need a unified weighting model for all types of modalities in order to avoid a lot of effort to come up with a new model for every modality type. Further, a merging strategy that works with little or no training is necessary, both because training can become very complex for a large amount of modalities and because in practical applications training data is not always available (Imhof and Braschler 2015).

2.2 Challenges

A multimodal IR system as described in this Section comes with several challenges that need to be solved in order to effectively use all the modalities. On the pursuit of a suitable weighting scheme for non-textual modalities, we can analyze the most popular textual weighting schemes. These can all be described in terms of how they combine three main components; the term frequency (*tf*), the document frequency (*df*) and the document length normalization component (Salton and Buckley 1988). Looking at these three components, we can understand their respective roles as follows: The first two components make sure that "characteristic" terms are weighed heavily. Hereby, a characteristic term is one that appears frequently in the document in consideration (term frequency) and rarely in the remainder of the collection (document frequency). These terms are suitable to distinguish a document from other documents in the collection. The third component, the document length normalization, was introduced to ensure no documents of a particular length are favored in an undue way, offsetting the increasing probability to observe terms frequently simply due to the verbosity of the document.

The concept of "being characteristic", embodied through *tf* as well as *df*, is quite general and therefore applicable to other non-textual modalities (Robertson and Zaragoza 2009). One basically needs to check the assumption that an "unforeseen" local frequency of a feature hints at relevance. For non-textual modalities, the "term frequency" is usually referred to as "feature frequency" (*ff*). In the remainder of this paper, we will use the two expressions interchangeably. In Sect. 5, we show how we can define the *tf* and *df* for the two non-textual modalities ratings and geographical coordinates.

When analyzing the requirements of a weighting scheme for effective merging of ranked lists, usually the raw-score merging hypothesis is considered. The raw-score merging hypothesis describes that similarity values are directly comparable if they are produced from similar search engines and underlying collections with similar statistics (Braschler 2004; Kwok et al. 1995; Savoy 2003, 2005). In Appendix 1, we show that it is favorable to use the same weighting scheme for all modalities when using raw-score merging. However, already textual modalities often invalidate the raw-score merging hypothesis w.r.t. to the similar collection statistics. For non-textual modalities, this is usually even more severe, since they do not follow the language statistics. Therefore, we propose a sampling-based approach in Sect. 6 to eliminate the differences in average and variance of document lengths and show that BM25 satisfies the derived properties, which makes it a viable weighting scheme for raw-score merging.

We can summarize the challenges of building multimodal IR systems discussed in this paper as follows.

1. Adapt BM25 to non-textual modalities
   (a) Define *tf*, *df* and document length
   (b) Validate generalizability of document length normalization
2. Evaluate merging strategies (raw-score merging hypothesis)
3. Validate suitability of BM25 for raw-score merging
4. Evaluate effectiveness of the approach

3 Related work

Much work has been done using additional non-textual modalities in order to improve the retrieval effectiveness of textual IR systems. A famous example is the query-independent modality PageRank (Brin and Page 1998) and it is now an established practice to use modalities such as URL-type, anchor text and link indegree in retrieval of Web data (Craswell etal. 2005; Hashemi and Kamps 2014; Macdonald et al. 2015). A lot of other retrieval research sub-fields such as geographical IR (Mandl et al. 2009), image retrieval (Villegas et al. 2015), XML retrieval (Kamps et al. 2004) and living labs (Schuth et al. 2015) provide and use a large range of different modalities in order to optimize the retrieval results. Hereby, the additional modalities are often no longer query-independent, but also explicitly or implicitly (e.g. inside a user profile) part of the query. In contrast to this paper, most of these models have been developed for a specific modality and the generalization to other modalities was not a focus.

For non-textual modalities the document length normalization is particularly important, since items usually have large variances in the “length” of their content in terms of those modalities. Looking towards textual retrieval, a number of efforts investigating the role of document length in ranking textual documents exist. Generally, consensus is that including document length normalization in weighting schemes tends to improve the retrieval performance (Amati and Rijsbergen 2002; Chowdhury et al. 2002; Losada and Azzopardi 2008; Singhal et al. 1996). The weighting scheme Lnu.ltn (Singhal et al. 1996) is explicitly based on the idea of revisiting the cosine document length normalization of TF.IDF. Singhal et al. (1996) estimate the likelihood of relevance and the likelihood of retrieval for all document lengths and improve the document length normalization by tilting the slope of the likelihood of retrieval in order to better match the slope of the likelihood of relevance. This tilt of the slopes then results in the new improved “pivoted document length normalization scheme”. Investigations of the document length normalization of the BM25 weighting scheme have shown that it fails when documents are very long (Lv and Zhai 2011) and that choosing the right document length normalization parameter b in BM25 can increase the retrieval performance by 22% Chowdhury et al. (2002). In XML retrieval, document length normalization is particularly important, since the retrievable items (XML elements) have a great variety in length. Kamps et al. (2004) revisit the role of language model document length normalization in the context of XML retrieval. Amongst others, they found that a combination of restricting the minimal size of the XML elements and length priors results in a higher effectiveness.

Oftentimes multiple intermediate result lists, one per modality, are produced when matching on multimodal collections. The problem of merging multiple ranked lists into a single ranked list is known from multilingual, multimedia and distributed retrieval. Fox and Shaw (1994) propose different strategies to fuse the scores; e.g. the sum of the scores or the maximal score. However, as Callan et al. (1995) point out, the scores might not be directly comparable, due to the different ranges of the scores.

The merging problem is very prominently studied in the multimedia IR community. Depeursinge and Müller show that 62% of the ImageCLEF working notes deal with data fusion, their detailed analysis reveals that, similar to all the other domains, the most used fusion strategy is a linear combination of the scores (Depeursinge and Müller 2010). Mostly the weights of the linear combination are either found manually or based on training data. Wilkins et al. (2006) however describe a method to automatically determine query-dependent modality weights using the score distribution of visual and textual modalities used in the context of video retrieval. Another unsupervised method to fuse multiple ranked lists for medical IR is presented by Mourão et al. (2015). Their fusion method combines the inverse rank approach of reciprocal rank fusion (Cormack et al. 2009) with the number of times a document appears on a rank and achieves a high precision. The unsupervised methods proposed in this paper try to fuse the modality scores without any weights, which we claim, is possible when treating all modalities with the same model.

Robertson et al. (2004) show the problems that arise when using a linear combination of the scores obtained from scoring multiple textual fields individually using BM25. The most important reason why this leads to poor retrieval effectiveness is the non-linear treatment of the term frequencies. This non-linearity is desirable for individual fields, since the information gain on observing a term for the first time is greater than the information gained on subsequently seeing the term. However, when using a linear combination of scores this non-linearity breaks. Therefore, Robertson et al. (2004) propose a method that uses a linear combination of the term frequencies instead of using a linear combination of the scores, with which the problem can be solved. The term frequency is not the only point that has to be considered in a retrieval setup with multi-field documents, also the document length and the parameters of the weighting scheme have to be questioned. When computing a score for each individual field the weighting scheme parameters, in BM25 the *tf* saturation parameter k_1 and the document length normalization parameter b have to be optimized for each field individually, which results in a huge number of optimization parameters. With the method suggested by Robertson et al. (2004) only two weighting scheme parameters have to be optimized. The suggested method also leads to substantially different term frequencies, since they replicate the content of the fields with the weight, the authors therefore suggest to use an adapted k_1 that is a scaled version of the original k_1 by the ratio between the original and the resulting average term frequency. For our methods, we use the idea of scaling k_1 when sampling all modalities to the same length.

4 Validating the generalizability of document length normalizations

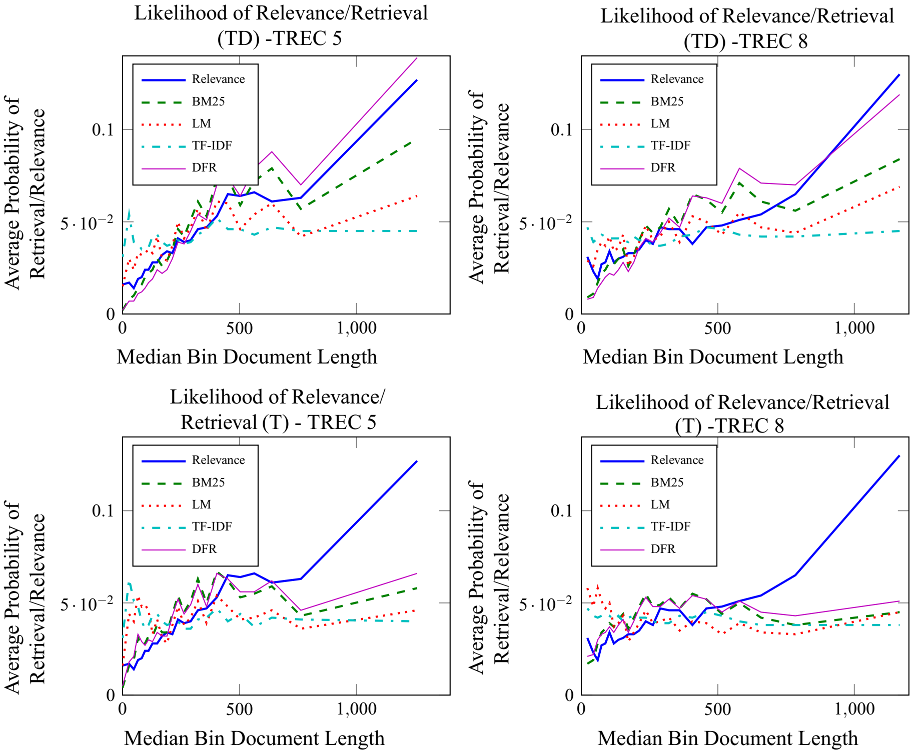
Similar to traditional textual retrieval, special care needs to be taken to handle varying document lengths for non-textual modalities as well. Non-textual modalities can have large variances in document lengths. In order to find a suitable weighting scheme for non-textual modalities, we analyze four of the most known weighting schemes with respect to their document length normalization robustness.

The experiments are conducted using the TREC 5 ad hoc collection (Voorhees and Harman 1996) and the TREC 8 ad hoc collection (Voorhees and Harman 1999). The choice of these rather classic test collections is motivated as follows: TREC 5 includes the Federal

Register sub-collection that contains very lengthy documents, resulting in a high variance w.r.t. the document lengths of the collection. TREC 8 has been chosen due to its use in earlier literature about document length normalization (Chowdhury et al. 2002; Losada and Azzopardi 2008; Lv and Zhai 2011), however has a smaller variance w.r.t. the document lengths than TREC 5 and we therefore expect that the effects of the document length component to be less pronounced. We used the full datasets and automatically generated queries from the topic title (T) and the description (D).

We examine the document length normalization and its impact on the retrieval effectiveness using the idea of Singhal et al. (1996). They calculate the likelihood of retrieval and relevance for each document length and employ these to adjust the document length normalization. We use these two likelihoods to visualize the effectiveness of the document length normalization of the four weighting schemes in study. To compute these likelihoods the documents are binned by their length. For each bin, the likelihood is defined as the ratio between the number of relevant/retrieved documents and the total number of documents in the bin. We then plot the likelihoods against the median document length in the bins.

Figure 2 shows the likelihood of relevance (bold line) and the likelihood of retrieval for all the weighting schemes for the TREC 5 and TREC 8 collections. The documents are divided in to 24 bins. As shown in this figure, longer documents have a higher probability of being relevant and retrieved. For both TREC 5 and TREC 8 as well as the long (TD) and short topics (T), BM25 and DFR match the likelihood of retrieval the best and we conclude



Likelihood of Relevance/Retrieval (TD) -TREC 5

Likelihood of Relevance/Retrieval (TD) -TREC 8

Likelihood of Relevance/ Retrieval (T) - TREC 5

Likelihood of Relevance/Retrieval (T) -TREC 8

Fig. 2 Likelihood of Retrieval/Relevance for the TREC5 / TREC 8 data using 24 bins and the original weighting schemes

that BM25 is able to handle large variances in document length. Since the document length normalization of BM25 is robust, it is suited to be used with non-textual modalities without any restriction regarding the variance of document lengths. Note that we did not include weighting scheme extensions, such as BM25L (Lv and Zhai (2011), that specifically target the robustness of the document length normalization, since they usually come with further assumptions regarding the statistics of the modalities.

5 BM25 model for non-textual modalities

5.1 BM25

Our experiments to validate the raw-score merging hypothesis and the generalizability of the document length normalization show that BM25 both works best for the raw-score merging and is amongst the most robust weighting schemes with highly varying document lengths. Therefore, we will focus our work with non-textual modalities on BM25.

Let us explore multimodal document collections such as used in GeoCLEF (Mandl et al. 2009) or in the social book search lab (Bogers et al. 2014). In these collections, documents are no longer just represented by only a set of terms (textual features) but also by geographical features or by book ratings that further describe the documents.

In this Section, we first re-capitulate BM25 for a textual modality and then show how its idea can be adapted to geographical coordinates and to book ratings. Table 1 shows the notations used for BM25 as well as for its non-textual adaptions.

The retrieval status value (RSV) of document d_j w.r.t. query q when using BM25 can be written as an inner product

$$w(\varphi_k, d_j) := \frac{\mathrm{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{l_j}{\Delta}) + \mathrm{ff}(\varphi_k, d_j)} \tag{2}$$

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) \cdot \log\left(\frac{0.5 + N - \mathrm{df}(\varphi_k)}{0.5 + \mathrm{df}(\varphi_k)}\right) \tag{3}$$

$$\mathrm{RSV}_{\mathrm{BM25}}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q), \tag{4}$$

where k_1 is the *tf* saturation parameter and b is the document length normalization parameter.

Table 1 Notation used for the BM25 for textual and non-textual modalities

D	Set of documents	$\Phi(d_j)$	Set of features representing document d_j
N	Number of documents	$\Phi(q)$	Set of features representing query q
d_j	Single document	$w(\varphi_k, d_j)$	Weight of feature φ_k for document d_j
q	Single query	$w(\varphi_k, q)$	Weight of feature φ_k for query q
Φ	Indexing vocabulary	$\mathrm{ff}(\varphi_k, d_j)$	Feature frequency of feature φ_k for document d_j
φ_k	Single indexing feature	$\mathrm{df}(\varphi_k)$	Document frequency of feature φ_k
l_j	Length of document d_j	$\mathrm{cf}(\varphi_k)$	Collection frequency of feature φ_k
Δ	Average document length	$\mathrm{RSV}(q, d_j)$	Retrieval status value of document d_j w.r.t. query q

For its document length normalization, BM25 (Robertson and Zaragoza 2009; Robertson et al. 1980) assumes a standard length of a document represented by the average document length. Hence, an author can decide to write a document longer or shorter than the standard length. Robertson and Zaragoza (2009) and Robertson et al. (1980) describe two cases why an author might decide to write a long document; either the author is more verbose than others or the author covers a larger scope. The verbosity assumption would lead to a division of the *tf* values by the document length. The scope assumption points to an opposite course of action, hence not dividing at all. Normally, the reason for a longer document is a combination of the two, thus Robertson's normalization balances the two using a tuning parameter b. Robertson proposed to use the number of tokens in a document as the document length, although he pointed out that BM25 should lead to similar results with slightly different definitions of the document length such as the number of characters. When using BM25 for non-textual modalities, it needs to be considered if this assumption holds true for those as well.

Since BM25 was originally designed for textual modalities, the question arises if its concept depends on the Zipfian distribution of the modalities as it is the case for natural language features. In particular the heuristic definition of the inverse document frequency (*idf*) can be motivated by the Zipf's law. However, over the years people have come up with several other interpretations on why the *idf* works as well as it does. For example, the theories that the *idf* corresponds to the probability of a term appearing in a document or to Shannon's information theory as described by Robertson (2004). It therefore is unclear how much the performance of BM25 depends on the Zipfian distribution of the modalities. Although we will not further investigate this question in this paper, we however assume that BM25 is generalizable to non-textual modalities with any distribution as long as the *tf* and *idf* can be defined in a way that the characteristic features still emerge.

Apart from the open question how well BM25 generalizes to modalities with a non-Zipfian distribution, it has been shown that BM25 is indeed generalizable to modalities with a Zipfian distribution such as a bag-of-visual-words in multimedia retrieval (Yang et al. 2007). Also the distribution of the modalities we use in our experiments satisfy Zipf's law. In the case of the GeoCLEF collection, which we use for our experiments with geographical coordinates, the coordinates have a Zipfian distribution, since they are extracted from the locations mentioned in the textual representation. Further, we analyzed the distribution of the ratings in the social book search collection and realized that they also have an approximate Zipfian distribution. It seems that the distribution of the ratings in this collection is not an exception, but appears to be a general phenomenon (Dalvi et al. 2013; Rajaraman 2009; Woolf 2014).

The *tf* saturation is parametrized by k_1 and makes sure that an increase of a high *tf* will contribute less to the score than an increase of a smaller *tf*. The higher the k_1 value, the more will an increase of a high *tf* contribute to the score, i.e. the saturation is less pronounced with high k_1 values.

The optimal choice of k_1 is not simple to make and also depends on the collection (Chowdhury et al. 2002). Further, k_1 needs to be adjusted if documents are replicated (Robertson et al. 2004). When replicating the content of all the documents (concatenate each document with itself; all documents have twice the length), neither the informativeness of a single document is changed nor the relevance of the documents to a particular query changes. However, if k_1 is not adjusted the BM25 weighting scheme will not lead to the same ranked list as without the replication. The BM25 weight for document d_j' that are replicated x-times is

$$w(\varphi_k, d'_j, k_1) = \frac{x \cdot \mathrm{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{x \cdot l_j}{x \cdot \Delta}) + x \cdot \mathrm{ff}(\varphi_k, d'_j)}, \tag{5}$$

which is not order preserving. However, if we set $k'_1 = x \cdot k_1$ we get $w(\varphi_k, d'_j, k'_1) = x \cdot w(\varphi_k, d_j, k_1)$ with which we can maintain the original ordering.

5.2 Geographical coordinates

For our BM25 model for geographical coordinates, we consider documents that are enriched with a discrete set of geographical coordinates. Let us model the three main ingredients of our weighting scheme: *ff*, *df* and document length, as follows. The *ff* of a coordinate in a document is defined as the number of occurrences of that coordinate in the document. The *df* is the number of documents that contain this coordinate and the document length is the number of locations in a document. Hereby, we assume that a document annotated with many geographical coordinates, covers a larger scope than a document with less coordinates, thus the argument of the textual BM25 document length normalization holds. Further, we assume, that the queries ask for documents in a specific geographical area, therefore a query is described by a single bounding box that encloses this area. The feature set and the feature frequency of a geographical feature φ_k for a query q is defined as

$$\Phi(q) := \mathrm{boundingbox}(q) \tag{6}$$

$$\mathrm{ff}(\varphi_k, q) := 1. \tag{7}$$

5.3 Ratings of books

For the ratings, we consider documents, that describe books including ratings given by their readers. When searching for books with a textual query, we do not know any query specific preference for a rating. However, we assume that in general readers will prefer books with higher ratings. If the ratings are in the range between one and five, we define the query as

$$\Phi(q) := \{1, 2, 3, 4, 5\} \tag{8}$$

$$\mathrm{ff}(\varphi_k, q) := \varphi_k. \tag{9}$$

Hereby, all the possible ratings (1–5) are part of the query, while the weight of a rating is equal to the rating itself; i.e. the weight of the rating 5 is 5 times higher than the weight of the rating 1. The three main ingredients of our weighting scheme: feature frequencies $\mathrm{ff}(\varphi_k, d_j)$, document frequencies $\mathrm{df}(\varphi_k)$ and document lengths l_j, are defined analogously to their definition for textual modalities. The *ff* is the number of times a rating occurs in a given document, the *df* is the number of documents that contain a given rating and the document length is the number of ratings in a document. We assume that a document with many ratings covers a larger range of opinions, hence covering a larger scope and thus the argument of the textual BM25 document length normalization holds.

6 Sampling-based BM25 for modality merging

6.1 Sampling

The proposed BM25 adaption for non-textual modalities enables us to merge modalities using the same weighting scheme, i.e a similar search engine as requested by the raw-score merging hypothesis. However, the raw-score merging hypothesis not only demands that similar search engines are used but also that the collection statistics are similar. Note, that the raw-score merging hypothesis is a rather old concept that has been introduced when merging multiple, possibly distributed textual document collections. In retrieval tasks with multiple modalities, the "collections" are no longer a set of textual documents but the different modalities. We have seen that the non-textual modalities have vastly different collection statistics, which invalidates the raw-score merging hypothesis. Therefore, we suggest a sampling based approach that allows us to adjust some properties of the collection statistics in order to reduce the difference. In particular, we adjust the average document length and the variance of the document lengths.

Our proposed sampling approach is similar to what is done in image retrieval when using dense or random feature sampling, where the same number of features for each image regardless of the pixel density and the number of concepts shown in the image is used (Moulin et al. 2010). The idea is to sample all modalities in all documents to a fixed document length as illustrated in Figure 3 for a single modality before BM25 is applied. Hereby, we use the number of tokens as the document length, although different definitions can be used. This results in the same collection statistics for all the modalities with respect to the average document length and the variance of document lengths. Namely, the average document length is the sampling size and the variance is zero. Since all documents have the same length no BM25 document length normalization is necessary, thus we choose $b = 0$.

The randomized sampling, however, leads to data loss due to down sampling and non-deterministic results. Therefore, we idealize the sampling idea by not sampling the document but simply simulating the resulting term statistics. This can be done by scaling the feature frequencies by the relative change of the document length that would result from sampling. For a single document d_j and a single modality with length l_j and a token φ_k with the feature frequency $\mathrm{ff}(\varphi_k, d_j)$ the scaled term frequency $\mathrm{ff}'(\varphi_k, d_j)$ is

$$\mathrm{ff}'(\varphi_k, d_j) = \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}, \tag{10}$$

where s is the sampling size (the fixed length of all documents). For example, if s is $3l_j$, all term frequencies are multiplied by 3.

We denote our idealized sampling based BM25 adaption BM25*S, where S stands for the sampling and the asterisk shows that no traditional document length normalization is applied; i.e. $b = 0$. The resulting the BM25*S weight for document d_j with sampling size s is

Fig. 3 Visualization of sampling three documents to the sampling size 5

Original		Sampled
$d_1 : \varphi_1 \varphi_2$		$d_1' : \varphi_1 \varphi_1 \varphi_2 \varphi_2 \varphi_2$
$d_2 : \varphi_1 \varphi_2 \varphi_3$	$\Longrightarrow$	$d_2' : \varphi_1 \varphi_1 \varphi_2 \varphi_2 \varphi_3$
$d_3 : \varphi_1 \varphi_2 \varphi_3 \varphi_4 \varphi_5 \varphi_6$		$d_3' : \varphi_1 \varphi_2 \varphi_3 \varphi_4 \varphi_5$

$$w_{\mathrm{BM25*S}}(\varphi_k, d_j) = \frac{\mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 + \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}. \tag{11}$$

Our sampling approach is some form of document replication, and thus the *ff* saturation parameter k_1 is not optimal anymore as described in Section 5 and by Robertson et al. (2004). In order to achieve the same retrieval effectiveness as without the sampling, the k_1 parameter needs to be adjusted. Since not all documents are replicated with the same factor, the optimal adjustment of the k_1 parameter cannot simply be the replication factor as in Sect. 5. However, we observed an approximately linear dependency of the optimal k_1 parameter to the average document length. Therefore, we set

$$k_1' = \frac{\Delta'}{\Delta} \cdot k_1, \tag{12}$$

where Δ is the average document length of the original documents and Δ' is the average document length of the sampled documents. This adjustment is slightly different to the adjustment Robertson et al. (2004) suggested, who used the ratio between the average term frequencies rather than the average document lengths. However, with their setup the two ratios are equivalent. With the sampling, the two ratios are not exactly equal, although quite similar, therefore both options seem valid. Further, when sampling, calculating the ratio between the average document lengths is a lot simpler than between the average term frequencies since the average document length after the sampling is equal to the sampling size ($\Delta' = s$), while the new average term frequencies are only known after the sampling is performed.

The weight for a document d_j, when using the combination of the idealized sampling and the k_1 adjustment (BM25-sampled), is calculated as

$$w_{\mathrm{BM25-sampled}}(\varphi_k, d_j) = \frac{\mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 \cdot \frac{s}{\Delta} + \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}. \tag{13}$$

We now have a sampling method BM25-sampled that can be applied to all modalities. We suggest using the same sampling length for all modalities, which results in the same collection statistics for all modalities with respect to the average document length and variance in document lengths. Hence, the raw-score merging hypothesis is fulfilled with respect to these two properties.

We can prove that this sampling method results in exactly the same weights as for BM25 with the normalization parameter b set to one.

Proof

$$\begin{aligned} w_{\mathrm{BM25-sampled}}(\varphi_k, d_j) &= \frac{\mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 \cdot \frac{s}{\Delta} + \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}} \\ &= \frac{\mathrm{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{s}{\Delta} \cdot \frac{l_j}{s} + \mathrm{ff}(\varphi_k, d_j)} \\ &= \frac{\mathrm{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{l_j}{\Delta} + \mathrm{ff}(\varphi_k, d_j)} \\ &= w_{\mathrm{BM25}(b=1)}(\varphi_k, d_j). \end{aligned}$$

□

This proof shows, that BM25 with full document length normalization ($b = 1$) already guarantees that the raw-score merging hypothesis is fulfilled with respect to the average document length and variance in document lengths. Therefore, BM25 seems to be suited to be used in a multimodal retrieval task. It however has been shown, that using $b = 1$ for BM25 tends to underestimate the relevance of long documents and therefore usually a smaller b is used; e.g. $b = 0.75$. In the following, we show how the sampling idea can be extended to allow arbitrary document length normalization parameters b.

6.2 Scope-aware sampling

Sampling all documents to the same length, which is equal to using BM25 with full document length normalization ($b = 1$), assumes that all documents have the same scope. However, some documents might discuss more topics than other documents and thus indeed should be represented with more tokens as described in Sect. 5. Similarly to BM25, we assume that the original document lengths of the documents give an indication about their scope. Thus, we can account for different document scopes by sampling the documents to different lengths based on their original length.

Many different definitions of a scope-aware sampling length using a document length normalization parameter bs are possible. We can however choose a definition so that the sampling based approach is identical to the traditional BM25 with parameter $b = bs$. We therefore define the adjusted number of sampled tokens s' for a document d_j as

$$s'(d_j) = l_j \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}. \tag{14}$$

All documents are now sampled to their corresponding sampling size $s'(d_j)$ rather than the same sampling size s for all documents. The adjusted feature frequencies therefore are

$$\begin{aligned} \mathrm{ff}'(\varphi_k, d_j) &= \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s'(d_j)}{l_j} \\ &= \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}. \end{aligned} \tag{15}$$

Unfortunately, this non-linear transformation of the document lengths does not exactly result in the same average document length for each modality, which would be necessary to fulfill the raw-score merging hypothesis. However, we found that the new sampled average document lengths of the modalities are close to each other and it is in practice a valid assumption that they are equal.

Further, we have found, that the optimal k_1 has no longer a linear dependency on the new average document length Δ' as we found for the sampling with a fixed sampling size s (BM25-sampled) as described in Sect. 6. It rather has a linear dependency to the sampling length s. Thus, for the scope-aware sampling we adjust the k_1 parameter as

$$k_1' = \frac{s}{\Delta} \cdot k_1. \tag{16}$$

We denote this scope-aware sampling with the k_1 adjustment and the non-normalized BM25 as BM25-scope. Its weight for a document d_j is calculated as

$$w_{\text{BM25-scope}} = \frac{\text{ff}(\varphi_k, d_j) \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}}{k_1 \cdot \frac{s}{\Delta} + \text{ff}(\varphi_k, d_j) \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}}. \tag{17}$$

With the scope-aware sampling it is possible to achieve approximately the same average document length for all modalities in all documents, while documents with a large scope are still represented by more tokens, by using the same sampling size parameter s for all modalities.

We can show that this scope-aware sampling is identical to the traditional BM25 for any document length parameter bs.

Proof

$$\begin{aligned} w_{\text{BM25-scope}} &= \frac{\text{ff}(\varphi_k, d_j) \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}}{k_1 \cdot \frac{s}{\Delta} + \text{ff}(\varphi_k, d_j) \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l_j}{\Delta}\right) \cdot \Delta}} \\ &= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{s}{\Delta} \cdot \frac{(1 - bs + bs \cdot \frac{l_j}{\Delta}) \cdot \Delta}{s} + \text{ff}(\varphi_k, d_j)} \\ &= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot (1 - bs + bs \cdot \frac{l_j}{\Delta}) + \text{ff}(\varphi_k, d_j)} \\ &= w_{\text{BM25}(b=bs)}(\varphi_k, d_j). \end{aligned} \tag{18}$$

□

Since BM25 is identical to our sampling approach BM25-scope, also BM25 is fulfilling the raw-score merging hypothesis with respect to the average document length with any document length normalization parameter. We can therefore conclude, that differences between average document lengths can be ignored when using raw-score merging with BM25. Hence, we can use BM25 with the same document length normalization parameter b for all modalities. The sampling approach is not needed in practice, since we have shown that it is identical to BM25.

Unlike BM25 with full document length normalization ($b = 1$), the variances of the document lengths are however not necessarily the same. Using our sampling idea, we can further adjust the definition of the sampled number of tokens in order to compensate the different variances of document lengths. We first apply a transformation to the document lengths to adjust the variance and then adjust the average document lengths as in Eq. 14 using the transformed document lengths. Thus, we do not ensure that all variances in document length are the same, but we ensure that the ratio between the standard deviation and the average document length is the same for all modalities. The adjusted number of tokens s'' with the adjustment for the variance of document length is

$$l'_j = (l_j - \Delta) \cdot rs \cdot \frac{\Delta}{\sigma} + \Delta \tag{19}$$

$$s''(d_j) = l'_j \cdot \frac{s}{\left(1 - bs + bs \cdot \frac{l'_j}{\Delta}\right) \cdot \Delta}, \tag{20}$$

where σ is the standard deviation of the document lengths and rs is the variance parameter that defines the target ratio between the standard deviation and the mean. We denote this sampling variation as BM25-var.

7 Experiments

The focus of our evaluation lies on measuring the effectiveness of a multimodal IR system built according to our guidelines (consistent treatment of the modalities, little or no training). In the scenarios we are interested in, the system needs to incorporate *all* modalities; ignoring modalities is not an option.

Our test system is built on top of Lucene[1] and is using the built-in weighting schemes wherever possible. For the scaled feature frequency and the k_1 adjustment, we adapted the built-in BM25 implementation. The merging of the modalities is performed using a raw-score merging (".raw") or a linear combination of the scores (".opt"). By using the latter, we violate our goal of using no training phase. Indeed, we use the opt-variant only for comparison purposes as a benchmark. In line with this role as a sort of "upper bound" on performance, we train the optimal weights using the same collection as used for testing. In essence for the opt-variant, we are only interested in showing that the effectiveness can be improved using BM25 on multiple textual as well as non-textual modalities.

Our experiments use two multimodal test collections, GeoCLEF and SBS.

7.1 Test collections

7.1.1 GeoCLEF

For the experiments with the geographical modality, we use the topics and collection of the GeoCLEF 2008 (Mandl et al. 2009) monolingual English search task. The collection is composed of the news articles from the British newspaper *The Glasgow Herald* (1995) and the American newspaper *The Los Angeles Times* (1994). In this task, 24 geographically challenging topics have been defined; e.g. "*Nobel prize winners from Northern European countries*". Here, we can differentiate between the textual information "Nobel price winners" and the geographical information "from Northern European countries". One of the challenges of geographical IR is that relevant documents not only contain the textual representation of geographical information "Northern European countries", but also concepts such as unions, countries or cities inside the geographical region.

Overell et al. (2008) and Buscaldi and Rosso (2008) proposed to separate the geographical information from the textual information, so that the two modalities (geographical and textual) can be treated differently. This allows that the additional information about geographical regions can be considered. Buscaldi and Rosso (2008) extracted location names from the documents and topics and mapped them to their geographical coordinates (longitude, latitude) using GeoWordNet. D. Buscaldi provided us a preprocessed geotagged version of the GeoCLEF 2008 collection. Further, we preprocessed the title fields of the topics by manually extracting a geographical bounding box for each topic. This could also be done automatically using the convex hull of the locations found with GeoWordNet (Buscaldi and Rosso 2008).

An important characteristic of the collection and task described above is the overlap of the textual and geographical modalities, since the geographical modality is extracted from

[1] https://lucene.apache.org/core/.

the text. Therefore, we also created a second modified version of the GeoCLEF 2008 test collection, which separates the geographical and textual information. For this, we removed the textual description of the geographical region from the queries; e.g. the query "*Nobel prize winners from Northern European countries*" becomes "*Nobel prize winners*" with the geographical bounding box that includes all Northern European countries. In the experiments, we refer to this task as "geoCLEFmod".

7.1.2 Social book search

For the experiments using the ratings as an additional modality, we use the social book search (SBS) 2016 lab task (Koolen et al. 2016). The collection consists of 2.8 million books from Amazon, extended with social meta-data from LibraryThing. For each book the fields ISBN, title, review, summary, ratings and tags are given. Each query is constructed from a real user request on LibraryThing. The query not only includes the title of the request and the description of the request itself but also example books mentioned by the user. Additionally, the personal catalog of each topic creator is available, which includes a list of the books the user has archived on LibraryThing along with his personal ratings. The relevance assessments are based on the actual suggestions to the original query on the LibraryThing forum. Forum suggestions normally get a relevance value of 1, however if the suggested book is already in the personal catalog of the topic creator the relevance value is 0. When the topic creator actually adds a suggested book to his library it is considered highly relevant and receives a relevance value of 4.

For the textual modality, we use the textual baseline established in our SBS participation (Imhof 2016; Imhof et al. 2015). We combine all textual fields of the documents into a single textual index field. The queries are constructed from the two textual topic fields title and request that are analogously combined into a single textual representation. Further, we expand the query text with the 35 most characteristic terms (determined by BM25) from the textual representation of the content of the example books given by the topic creator. All books already read by the topic creator are filtered from the result list.

7.2 Results

Following our own guidelines on how to build a multimodal IR system, we sample the non-textual modalities to the same length as the textual modality. For the GeoCLEF 2008 collection, we therefore sample the geographical modality from an average document length of 7.4 to the sampling length of 357.7. Analogously, the ratings in the SBS collection with an average document length of 5.05 are sampled to the sampling length of 674.7. The target standard deviation ratio parameter *rs* is chosen based on the textual modality as well. For GeoCLEF 2008 this is 1.01 and 2.75 for SBS. This results in a reduction of the standard deviation for the non-textual modalities to 83% respectively 93%. For the runs using the scope-aware sampling (BM25-scope and BM25-var) the normalization parameter *bs* is 0.75. Note that the scope-aware sampling BM25-scope is identical to BM25 and BM25-sampled is identical to BM25 with document length normalization parameter $b = 1$.

As mentioned, the goal of this paper is to establish a baseline for a multimodal IR system that involves all the given modalities and merges the scores generated by a unified model under the raw-score merging hypothesis. Hereby, we require all the modalities to be considered in the result list. We argue that in practice, it is not possible, for many reasons, including e.g. regulatory ones, to simply ignore or discard entire modalities, or parts of the document collection. For example, a book selling company might find that good ratings of books positively influences the purchase behavior of their customers and thus the ratings have to be included in the search engine.

Building an effective multimodal IR system that integrates all modalities with little or no training remains a hard challenge. Wildly different characteristics, and wildly different degrees of informativeness across the modalities means that the average retrieval effectiveness may *drop* when integrating all modalities, such as evaluated through popular measures like MAP. We advise caution in overinterpreting such a result. Firstly, the average hides many meaningful changes in system behavior and secondly, user perception will likely be different from the measured average improvement if a user realizes that parts of his query or of the documents are ignored. For the time being, a lower retrieval effectiveness of an experiment integrating all modalities versus an experiment discarding some modalities thus mainly serves to highlight how far we still are from finding the perfect recipe for multimodal retrieval, but not to point to a reduced system as a viable, practical alternative.

In the following experiments, we show the effectiveness of our multimodal baseline using the three derived versions of BM25 as the unified weighting scheme for all the modalities merged under the raw-score merging hypothesis. In each of the following Tables 2, 3, 4 and 5, we compare two runs with the same collection. We underline any statistically significant differences in performance according to the MAP to the first run resulting from a paired randomization test (Smucker et al. 2007) (significance level $\alpha = 5\%$). For the GeoCLEF 2008 collection, we removed the outlier query 79-GC to calculate the significance. In Appendix 2 we additionally show the same runs evaluated using the nDCG@10 measure. The following conclusions drawn from the results using the MAP are all supported by the results using the nDCG@10.

7.2.1 *Base performance of systems integrating non-overlapping modalities*

We start our experiments by establishing the base performance of multimodal systems that integrate all non-overlapping modalities as built according to our guidelines.

To this end, Table 2 shows the MAP for the SBS 2016 and the GeoCLEFmod 2008 collection both for the multimodal baseline (denoted as ".raw") and the runs with the textual modalities alone (denoted as ".text"). As a consequence of our discussion above, the ".text"-run can only serve as a yardstick: it violates the rule that we want to integrate all modalities. Effectively, it gives us a "lower bound" of performance to compare to. For the SBS collection, the multimodal baseline achieves a significantly higher MAP than the textual run. For the GeoCLEFmod 2008 collection the run with BM25 with no document length normalization (BM25 ($b = 1$)), which is identical to BM25-sampled, achieves a

Table 2 Retrieval results (MAP) for the runs with the textual modalities and the raw-score merging of both modalities for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text	0.0320	0.0396	0.0396
SBS.text+ratings.raw	0.0390	0.0448	0.0447
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678

Table 3 Retrieval results (MAP) for the runs with the textual modalities and the non-textual modalities (geographical coordinates and ratings) for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text	0.0320	0.0396	0.0396
SBS.ratings	0.0089	0.0121	0.0121
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.geo	0.0540	0.0589	0.0588

Table 4 Retrieval results (MAP) for the runs with the textual modalities and the raw-score merging of both modalities for the GeoCLEFmod 2008 and the GeoCLEF 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678
geoCLEF.text	0.2509	0.2566	0.2566
geoCLEF.text+geo.raw	0.1548	0.0705	0.0703

Table 5 Retrieval results (MAP) for the runs with the raw-score merging of the modalities and the optimized linear combination of the modality scores for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text+ratings.raw	0.0390	0.0448	0.0447
SBS.text+ratings.opt	0.0398	0.0450	0.0450
SBS.text+ratings.rcpr	0.0104	0.0139	0.0139
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678
geoCLEFmod.text+geo.opt	0.2351	0.2442	0.2446
geoCLEFmod.text+geo.rcpr	0.1292	0.1393	0.1393

MAP in the range of the textual run. The BM25-scope and BM25-var runs with raw-score merging achieve a lower MAP than the run with text only.

7.2.2 Analysis of individual modalities

It is helpful to further look into the contributions of individual modalities to the overall result. Table 3 shows the retrieval effectiveness of each modality individually. Both the geographical modality and the ratings do not achieve the same retrieval effectiveness as the textual modality. This was expected for both, since intuitively the textual description of a book is more important than its ratings and the textual content of a newspaper article is more important than the mentioned geographical locations.

Merging under the raw-score hypothesis suggests adding the scores of the different modalities into a single score without any weights. However, as shown in Table 2 even though we proved that the raw-score merging hypothesis is fulfilled w.r.t. the average document length as well as for the variance of the document lengths (for BM25-var) the merged result list is only better than the textual run for the SBS task and not for the GeoCLEF task. We claim that this is since the method so far cannot properly capture the difference in informativeness of the modalities.

7.2.3 Dealing with overlapping modalities

We next want to explore to what extent the overlapping of content in modalities has an impact on the overall effectiveness. Table 4 shows the MAP of the textual run and the multimodal baseline using the GeoCLEFmod 2008 task as well as GeoCLEF 2008 task.

As expected the textual modality in the GeoCLEFmod task achieves a lower MAP than the textual modality in the original GeoCLEF task. This is due to the deletion of the geographical information in the textual modality as described in Sect. 7.1.1. The modalities in the GeoCLEFmod 2008 task therefore do not have an information overlap, while the modalities in the GeoCLEF 2008 task do contain overlapping information, namely all the information present in the geographical modality is also present in the textual modality. The experiments that merge the two modalities under the raw-score merging hypothesis show that without the information overlap between the modalities the MAP of the merged run ("geoCLEFmod.text+geo.raw") is within the range of the textual modality alone. However, when merging modalities with an information overlap ("geoCLEF.text+-geo.raw") the MAP drops significantly—it is much harder to merge the modalities so that only the "additional" contribution makes a beneficial impact.

7.2.4 Optimal merging potential due to training

We argue that a lot of the drop in retrieval effectiveness from the ".text" to the ".text+geo.raw" experiment is due to the inherent difficulty of appropriately merging the contributions of the individual modalities into the overall result. The closest method to raw-score merging that allows us to weight the contributions of the individual modalities is a linear combination of the scores. Therefore, we try to verify this assumption through comparing the multimodal baseline (".raw") with an approximate upper bound using a linear combination of the scores with trained weights (".opt") (see Table 5). The optimal weights are trained on the information available in the relevance assessments of the test collection. Clearly, this information is not available in practice. Furthermore, training the optimal weights on the same queries as were tested turns this in a retrospective evaluation. As the obtained result is merely a data point to compare our results to, we accept these limitations. For SBS there is no significant difference between merging the modality scores under the raw-score hypothesis and merging using the optimal linear combination. However, for the GeoCLEFmod 2008 collection merging the scores of the textual and the non-textual modalities using optimal linear combination has a significantly higher MAP then the merging under the raw-score merging hypothesis. Consider, however, that the opt-variants only serve as a yardstick: They can only be used when training data is available which is often missing in practical applications and which was not the goal of this paper. The optimal run also shows that the usage of BM5 for the non-textual modalities not only leads to good results when merging under the raw-score merging hypothesis but also when training optimal weights. The traditional BM25, which is identical to BM25-scope, already seems to be a good choice, since the variance adjustment does not lead to a significantly better result neither for raw-score merging nor for the optimal linear combination of the scores.

To get more context in order to judge the performance of our ".raw" runs, we have also explored the use of reciprocal rank fusion (Cormack et al. 2009), another well known unsupervised fusion method. These runs are denoted with ".rcpr" in Table 5, where we underline the runs that are significantly different to the ".raw" runs. For the SBS collection, reciprocal rank fusion leads to a significantly lower MAP for all BM25 variants. However, for the GeoCLEFmod 2008 collection the MAP is in the same range as the raw-score merging run with BM25-sampled but significantly better with BM25-scope and BM25-var, although still significantly lower than the optimal linear combination (".opt").

7.2.5 Summary of results

We can summarize the results of our experiments with the following questions.

1. Can we produce a multimodal baseline with an effectiveness in the range of the textual run? **Yes**, we find better retrieval effectiveness for the SBS collection and retrieval effectiveness in the same range (within statistical significance) for the GeoCLEF collection without overlapping modalities.
2. Do modalities differ with respect to their contribution to relevance? **Yes**, in both collections the contribution by the textual modality is by far the greatest, thus turning the ".text" yardstick into a challenging lower bound.
3. Does it matter that modalities have overlapping information? **Yes**, it is much harder to merge individual contributions by modalities in case they are overlapping.
4. Is it possible to get competitive performance without training? **Yes and no**. We have found competitive performance in the case of the SBS collection, where we have no overlapping modalities. We are still a long way from matching the performance of the opt-variant on the GeoCLEF collection, however.

8 Conclusions

In this paper, we demonstrate best practices for the integration of many modalities into an IR application without the use of training data. We claimed that in complex multimodal collections with a large number of diverse modalities, it becomes crucial to treat the modalities with a unified model, due to the quickly increasing complexity. We started by analyzing the requirements for such a unified model and showed that BM25 is a suitable weighting scheme to be used and to merge the modalities under the raw-score merging hypothesis. We proposed an adaptation of the BM25 weighting scheme for the two non-textual modalities ratings and geographical coordinates and established a multimodal baseline that uses all the modalities and merges them under the raw-score merging hypothesis without any training.

In order to show the suitability of BM25 scores to be merged under the raw-score merging hypothesis, a sampling based approach for BM25 was introduced to deal with the different collection statistics, in particular the average document length and the variance of the document lengths of the modalities. We proved that applying BM25 with full document length normalization $b = 1$ to all modalities already ensures that the raw-score merging hypothesis w.r.t. the average document lengths and the variance of document lengths is fulfilled, since it is identical to the sampling approach. Analogously, we proved that the raw-score merging hypothesis w.r.t. the average document length also holds for BM25 with a general document length normalization parameter $b \neq 1$, however not w.r.t. the variance of document length. Our experiments show that adhering to the raw-score merging hypothesis is indeed beneficial.

In our experiments, we established a multimodal baseline that involves all the given modalities and merges the scores generated by a unified model under the raw-score merging hypothesis. We showed that by following our approach the multimodal baseline reaches a significantly better retrieval effectiveness than the textual run for the SBS collection and lies within the same range (within statistical significance) for the GeoCLEF 2008 collection without overlapping modalities. Further, we analyzed the contribution of the individual modalities to relevance and found that the contribution of the textual modalities is the greatest. Also, we saw in the experiments that dealing with modalities with overlapping information is a hard problem. Finally, we found similar performance of our multimodal baseline when comparing it to a trained linear combination of the scores in case of the SBS collection, which we consider to be very encouraging.

The multimodal baseline presented in this paper merges the modality scores under the raw-score merging hypothesis and therefore assumes that each modality is equally important for the overall relevance of a document. However, in the experiments we saw that there are wildly different degrees of informativeness across the modalities. As a next step towards best practices for multimodal IR systems, we will investigate to further extend the proposed methods but incorporate the informativeness of the different modalities without the usage of any training data.

Appendix 1: Validating the raw-score hypothesis

As shown in Sect. 2, we propose to handle each modality separately. This fundamental approach models that a unified "merged" result list needs to be synthesized. As pointed out in Sect. 3, the raw-score merging hypothesis states that merging scores from multiple ranked lists is more effective when the scores are produced from the same underlying weighting scheme with the same collection statistics. As the hypothesis is an important stepping-stone to the definition of a consistent, "best-practice" way of treating each modality, we present an attempt to verify it experimentally. Similar to Savoy (2005), we investigate this hypothesis on the multilingual document collection used in the CLEF 2004 AdHoc-News task. It consists of four document collections in four languages: English, Finnish, French and Russian. The queries are also provided in all the four languages; however the goal is to present a single ranked list with all the relevant documents from all languages. Hence, the result lists resulting from the monolingual retrieval have to be merged.

In contrast to the work by Savoy, we are interested in how different commonly used weighting schemes behave in respect to the raw-score merging hypothesis. In the following experiments, we therefore use the four weighting schemes: BM25, divergence from randomness (DFR), language models (LM) and TF-IDF for the retrieval and show the resulting retrieval effectiveness when merging them into a single ranked list. Since the goal of the experiments is not to get the highest effectiveness possible but to show the validity of the raw-score merging hypothesis, we did not optimize any parameters of the weighting schemes and used the default analyzers for each language provided by Lucene. The saturation parameter k_1 of BM25 is set to 1.2 and the document length normalization parameter $b = 0.75$. The basic model of the DFR weighting scheme is the limiting form of Bose-Einstein, for the first normalization the Laplaces law of succession is used and the second normalization is based on the second hypothesis. For the LM we use Dirichlet smoothing with a smoothing parameter μ of 2000. We constructed short queries using the title and the description given for each topic. Table 6 shows the mean average precision (MAP) for each language individually using the four different weighting schemes. For English the MAP is highest when using BM25, for Finnish the highest effectiveness is reached using LM and for French and Russian DFR leads to the highest MAP, when merging the scores resulting from these four runs we call it "Best".

We underline any statistically significant differences with respect to the run with the highest per-language MAP, which is printed in bold letters. Hereby, the significance is calculated using a paired randomization test (Smucker et al. 2007) (significance level $\alpha = 5\%$). Looking closely at the difference in MAP between the BM25 and LM for the English collection, we can observe that for 10 queries over 50, LM offers a higher performance while for 25 requests BM25 performs better than LM. For the remaining 15 queries, the MAP difference between the two runs is smaller than 0.02. Thus, in average, BM25 depicts a higher MAP than LM. From a statistical point of view however, the difference cannot be viewed as significant because for several queries, LM presents a higher performance. A similar reasoning applies to the LM and the TF-IDF run for the Russian collection.

We merge the ranked lists produced from four languages into a single ranked list using four different well-known merging strategies. In general when using raw-score merging all the scores of a document in all ranked lists are added to a single score, which is then used to produce the merged ranked list. However, in this multilingual setup each document is only available in a single language and therefore only gets a single score. In this case, the raw-score merging just results in ordering the documents from all languages with respect to the scores in the per language ranked list. In the round robin merging approach, we take one document in turn from each individual ranked list (Voorhees et al. 1995). The third merging strategy is "Norm(max)" were we normalize the scores before merging by dividing them by the maximal document score of the corresponding query. The last strategy is a linear combination of the scores of the ranked lists. This strategy requires a

Table 6 Monolingual retrieval results (MAP) for CLEF 2004 using short queries (TD)

Run	English	Finnish	French	Russian
BM25	**0.4320**	0.3728	0.1618	0.2686
DFR	0.4228	0.3748	**0.1642**	**0.2760**
LM	0.4075	**0.3809**	0.1606	0.2360
TFIDF	0.4121	0.3580	0.1583	0.2577

Bold denotes the best result per language

Table 7 Multilingual retrieval results (MAP) for CLEF 2004 using different merging strategies

Run	Raw-score	Round robin	Norm(max)	lin.comb.
BM25.all	**0.2494**	0.1874	0.2018	0.2606
DFR.all	0.2471	0.1892	0.2018	0.2589
LM.all	0.2400	0.1825	0.1940	0.2430
TFIDF.all	0.2412	0.1807	0.1866	0.2494
Best	0.1812	**0.1926**	**0.2046**	**0.2656**

Bold denotes the best result per language

training set to find the optimal weights for each language. We used the same collection for the training and testing since we are not interested in optimizing the effectiveness but to show the difference of the individual weighting schemes. Table 7 shows the MAP of four runs in which ranked lists produced by a single weighting scheme are merged into a ranked list (BM25.all, DFR.all, LM.all, TFIDF.all) and the MAP of the run were the ranked lists that produced the best MAP for each language individually are merged into a ranked list ("Best"). We underlined any statistically significant differences in performance according to the MAP of the runs using a single weighting scheme with respect to the "Best" run. Hereby, the significance is calculated using a paired randomization test (Smucker et al. 2007) (significance level $\alpha = 5\%$).

As expected from the results by Savoy (2005), the runs using the same weighting scheme for all languages perform significantly better than the "Best" run using the raw-score merging, while BM25 performs slightly better than the other weighting schemes. Using the other merging strategies, the "Best" run performs the best, although not significantly. Also, the "Best" run requires that the best weighting scheme per-language is known, which usually is not the case in practical applications.

Appendix 2: Experimental results with nDCG@10

The following tables show the results of the experiments described in Sect. 7.2 using the nDCG@10 measure (Tables 8, 9, 10, 11).

Table 8 Retrieval results (nDCG@10) for the runs with the textual modalities and the raw-score merging of both modalities for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text	0.0467	0.0561	0.0561
SBS.text+ratings.raw	0.0561	0.0634	0.0633
geoCLEFmod.text	0.1709	0.1826	0.1826
geoCLEFmod.text+geo.raw	0.1500	0.0728	0.0646

Table 9 Retrieval results (nDCG@10) for the runs with the textual modalities and the non-textual modalities (geographical coordinates and ratings) for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text	0.0467	0.0561	0.0561
SBS.ratings	0.0122	0.0205	0.0207
geoCLEFmod.text	0.3573	0.3932	0.3932
geoCLEFmod.geo	0.2038	0.0733	0.0654

Table 10 Retrieval results (nDCG@10) for the runs with the textual modalities and the raw-score merging of both modalities for the GeoCLEFmod 2008 and the GeoCLEF 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
geoCLEFmod.text	0.0467	0.0561	0.0561
geoCLEFmod.text+geo.raw	0.0122	0.0205	0.0207
geoCLEF.text	0.1709	0.1826	0.1826
geoCLEF.text+geo.raw	0.0591	0.0630	0.0630

Table 11 Retrieval results (nDCG@10) for the runs with the raw-score merging of the modalities and the optimized linear combination of the modality scores for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions

Run	BM25 ($b = 1$)	BM25 ($b = 0.75$)	
	BM25-sampled	BM25-scope	BM25-var
SBS.text+ratings.raw	0.0561	0.0634	0.0633
SBS.text+ratings.opt	0.0584	0.0648	0.0648
SBS.text+ratings.rcpr	0.0145	0.0228	0.0230
geoCLEFmod.text+geo.raw	0.1500	0.0728	0.0646
geoCLEFmod.text+geo.opt	0.3425	0.3869	0.3894
geoCLEFmod.text+geo.rcpr	0.1614	0.1773	0.1773

References

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 357–389.

Bogers, T., Koolen, M., Jaap, K., Kazai, G., & Preminger, M. (2014). Overview of the INEX 2014 social book search track. In *Conference and labs of the evaluation forum* (pp. 462–479).

Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Information Retrieval*, *7*(1), 183–204.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Buscaldi, D., & Rosso, P. (2008). The UPV at GeoCLEF 2008: The GeoWorSE system. In working notes from the cross language evaluation forum.

Bush, V., et al. (1945). As we may think. *The Atlantic Monthly*, *176*(1), 101–108.

Callan, J. P., Lu, Z., & Croft, W. B. (1995). Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–28). ACM.

Chowdhury, A., McCabe, M. C., Grossman, D., & Frieder, O. (2002). Document normalization revisited. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 381–382). ACM.

Cleverdon, C. (1967). The cranfield tests on index language devices. In *Aslib proceedings* (Vol. 19, pp. 173–194). MCB UP Ltd.

Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 758–759). ACM.

Craswell, N., Robertson, S., Zaragoza, H., & Taylor, M. (2005). Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 416–423). ACM.

Dalvi, N. N., Kumar, R., & Pang, B. (2013). Para'normal'activity: On the distribution of average ratings. In *ICWSM*.

Depeursinge, A., & Müller, H. (2010). Fusion techniques for combining textual and visual information retrieval. *ImageCLEF* (pp. 95–114). Berlin: Springer.

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. NIST Special Publication SP, pp. 243–243.

Hashemi, S. H., & Kamps, J. (2014). *Venue recommendation and web search based on anchor text*. DTIC Document, Tech. rep.

Imhof, M. (2016). BM25 for non textual modalities in social book search. In *Seventh international conference of the CLEF Association, CLEF*.

Imhof, M., Badache, I., & Boughanem, M. (2015). Multimodal social book search. In *Sixth international conference of the CLEF Association, CLEF*.

Imhof, M., & Braschler, M. (2015). Are test collections real? Mirroring real-world complexity in IR test collections. In I. R. Experimental (Ed.), *Meets multilinguality, multimodality, and interaction* (pp. 241–247). Berlin: Springer.

Kamps, J., De Rijke, M., & Sigurbjörnsson, B. (2004). Length normalization in XML retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 80–87). ACM.

Koolen, M., Bogers, T., Gäde, M., Hall, M., Hendrickx, I., Huurdeman, H., et al. (2016). Overview of the CLEF 2016 social book search lab. In *International conference of the cross-language evaluation forum for European languages* (pp. 351–370). Berlin: Springer.

Kwok, K., Grunfeld, L., & Lewis, D. (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. NIST Special Publication SP, pp. 247–247.

Li, X., & Croft, W. B. (2003). Time-based language models. In *Proceedings of the twelfth international conference on information and knowledge management, CIKM '03* (pp. 469–475). New York, NY: ACM. https://doi.org/10.1145/956863.956951.

Losada, D. E., & Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, *11*(2), 109–138.

Lv, Y., & Zhai, C. (2011). When documents are very long, BM25 fails! In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 1103–1104). ACM.

Macdonald, C., Dinçer, B. T., & Ounis, I. (2015) Transferring learning to rank models for web search. In *Proceedings of the 2015 international conference on the theory of information retrieval* (pp. 41–50). ACM.

Mandl, T., Carvalho, P., Di Nunzio, G. M., Gey, F., Larson, R. R., Santos, D., & Womser-Hacker, C. (2009). GeoCLEF 2008: The CLEF 2008 cross-language geographic information retrieval track overview. In *Evaluating systems for multilingual and multimodal information access* (pp. 808–821). Berlin: Springer.

Moulin, C., Barat, C., & Ducottet, C. (2010). Fusion of tf. idf weighted bag of visual features for image classification. In *2010 International workshop on content-based multimedia indexing (CBMI)* (pp. 1–6). IEEE.

Mourão, A., Martins, F., & Magalhães, J. (2015). Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, *39*, 35–45.

Müller, H., Clough, P., Deselaers, T., & Caputo, B. (2010). Image-CLEF: Experimental evaluation in visual information retrieval series. The information retrieval series

Overell, S., Rae, A., & Rüger, S. (2008). MMIS at GeoCLEF 2008: Experiments in GIR. In Working notes from the cross language evaluation forum.

Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Berlin: Springer.

Rajaraman, S. (2009). Five stars dominate ratings. https://youtube.googleblog.com/2009/09/five-stars-dominate-ratings.html

Robertson, S. E., Van Rijsbergen, C., & Porter, M. F. (1980). Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual ACM conference on research and development in information retrieval* (pp. 35–56). Butterworth & Co.

Robertson, S., Zaragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on information and knowledge management* (pp. 42–49). ACM.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, *60*(5), 503–520.

Robertson, S., & Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond*. Breda: Now Publishers Inc.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

Savoy, J. (2003). Advances in cross-language information retrieval: Third workshop of the cross-language evaluation forum, CLEF 2002 Rome, Italy, September 19–20, 2002 Revised Papers, chap. Report on CLEF 2002 experiments: combining multiple sources of evidence, pp. 66–90. Berlin: Springer.

Savoy, J. (2005). *Data fusion for effective european monolingual information retrieval* (pp. 233–244). Berlin: Springer. https://doi.org/10.1007/11519645_24.

Schuth, A., Balog, K., & Kelly, L. (2015). Overview of the living labs for information retrieval evaluation (LL4IR) CLEF lab 2015. In *International Conference of the cross-language evaluation forum for European languages* (pp. 484–496). Berlin: Springer.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–29). ACM.

Smucker, M. D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management* (pp. 623–632). New York, NY: ACM. https://doi.org/10.1145/1321440.1321528

Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A. G. S., Bromuri, S., Amin, M. A., & Mohammed, M. K., et al. (2015). General overview of imageCLEF at the CLEF 2015 labs. In *International conference of the cross-language evaluation forum for European languages* (pp. 444–461). berlin: springer.

Voorhees, E. M., & Harman, D. (1999) Overview of the eighth text retrieval conference (TREC-8). In *TREC*.

Voorhees, E., Gupta, N. K., & Johnson-Laird, B. (1995). The collection fusion problem. NIST Special Publication SP, pp. 95–95.

Voorhees, E. M., & Harman, D. (1996). Overview of the fifth text retrieval conference (TREC-5). *TREC*, *97*, 1–28.

Wilkins, P., Ferguson, P., & Smeaton, A. F. (2006) Using score distributions for query-time fusion in multimediaretrieval. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval* (pp. 51–60). ACM.

Woolf, M. (2014). A statistical analysis of 1.2 million amazon reviews. http://minimaxir.com/2014/06/reviewing-reviews/

Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007) Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on multimedia information retrieval* (pp. 197–206). ACM