BM25 for Non-Textual Modalities in Social Book Search

Melanie Imhof[1,2]

[1] Université de Neuchâtel, Neuchâtel, Switzerland
[2] Zurich University of Applied Sciences, Winterthur, Switzerland
imhf@zhaw.ch

Abstract. The Social Book Search (SBS) lab at CLEF 2016 provides a complex test collection that gives the opportunity to experiment with retrieval methods that combine various modalities in order to achieve the best possible ranked list. We show how the idea of being "characteristic", which is used as the core concept in most of the weighting schemes used for textual modalities, can be applied to non-textual modalities. Our approach re-defines BM25 for the three non-textual modalities found in the SBS collection: ratings, price and number of pages. A fuzzy query is constructed from the user preferences inferred from the user's catalog. The results are used to re-rank a textual baseline, which significantly improves the retrieval effectiveness.

Keywords: BM25, non-textual modalities, fuzzy query.

1 Introduction

The suggestion track of the INEX Social Book Search (SBS) at CLEF 2016 allows researchers to evaluate their methods on a multimodal collection with queries constructed from real LibraryThing user requests. For the books in the collection not only the book meta-information from Amazon (description, binding, number of pages, price etc.) is available but also user generated information such as book ratings. Also, the personal catalogs of the users are given and can be used to infer user preferences.

In our SBS 2015 participation [2], we found that the user preferences can be used to improve the retrieval effectiveness, by incorporating the books read by the users in a random forest based learning to rank approach. In this participation, we focus on taking the user preferences into account using a different approach. BM25 is a well known weighting scheme that has widely been used in text retrieval. It was originally developed for the English language but it has proven to be useful for other languages as well as for image retrieval [3]. We show how BM25 can be applied to the modalities ratings, price and number of pages. The BM25 scores of these non-textual modalities are then used to re-rank the textual baseline to significantly improve it.

2 Retrieval Models

2.1 Textual Models

Similar to our participation in 2015, we employ a textual baseline [2] as a basis for our methods. For the textual score, we merge all textual fields of the document into a single textual index field and construct queries from the two topic fields *title* and *request* that are analogously merged into a single textual representation. Further, we use the example books mentioned by the topic creators to expand the queries with the 35 most characteristic terms. Hereby, the most characteristic terms of the example books are computed by BM25.

Additionally, we filter the books already read by the topic creator from the final ranked list, since this is a hard criterion in the relevance assessments [1]. Hereby, we determine the read books from the catalog of the topic creator.

2.2 BM25 Model for Non-Textual Modalities

BM25 can be described in terms of how it combines three components; the feature frequency (*ff*), the document frequency (*df*) and the document length normalization component [5]. Although, it was originally developed for retrieval on English language text, it has generalized well to many related tasks, such as multilingual retrieval, multimedia retrieval and others. The *ff* and the *df* make sure that "characteristic" terms are weighed heavily. Hereby, a characteristic term is one that appears frequently in the document in consideration (*ff*) and rarely in the remainder of the collection (*df*). This concept of "being characteristic" is quite general and therefore applicable to other (non)-textual modalities [4]; i.e. bag of visual words in image retrieval, locations in geographical IR or timestamps in time-aware IR.

The retrieval status value (RSV) of document d_j w.r.t. query q when using BM25 is defined as

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

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) \cdot \log\left(\frac{0.5 + N - \mathrm{df}(\varphi_k)}{0.5 + \mathrm{df}(\varphi_k)}\right) \tag{2}$$

$$\mathrm{RSV}_{\mathrm{BM25}}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q), \tag{3}$$

where k_1 is the *ff* saturation parameter and b is the document length normalization parameter. The k_1 parameter controls the amount an incremented *ff* will to contribute to the score. The notation used for the BM25 and its non-textual adaptions is described in Table 1.

For the suggestion track of the SBS lab at CLEF 2016, we adapt BM25 for three non-textual modalities the ratings, the price and the number of pages and use it to re-rank the textual baseline.

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	$\mathrm{ff}(\varphi_k, d_j)$	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	Δ	average document length in number of tokens

Ratings For the ratings, we do not have a per-user query information, but we assume, that in general users will prefer books with higher ratings. Therefore, we define the query in the following way

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

$$\mathrm{ff}(\varphi_k, q) := \varphi_k. \tag{5}$$

With this definition, each possible rating (1-5) is part of the query, however, a rating 5 is weighted 5 times heavier than a rating 1. The definition of the feature frequencies $\mathrm{ff}(\varphi_k, d_j)$, document frequencies $\mathrm{df}(\varphi_k)$ and document lengths l_j is analogous to the definition used for text. Hence, the *ff* is the number of times a given rating appears in a 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.

Price For the price, we use the average price of the books that the user has already read $\Delta_p(q)$ as the query information. Since an exact match of the price is not meaningful, we use a fuzzy search with $\Delta_p(q)$ as the search parameter. We assume, that a user would also like books that are at most 20% cheaper and at most 30% more expensive than the average price of the books in his library. Although, we assume that generally a cheaper book is always acceptable, we still set a lower bound, because we assume that people tend to like similar kinds of books, that are usually in the same price range. The query's set of features and feature frequencies are defined as

$$\Phi(q) :=]0.8 \cdot \Delta_p(q), 1.3 \cdot \Delta_p(q)[\tag{6}$$

$$\mathrm{ff}(\varphi_k, q) := \begin{cases} \frac{1.3 \cdot \Delta_p(q) - \varphi_k}{0.3 \cdot \Delta_p(q)} & \text{if } \varphi_k \geq \Delta_p(q) \\ \frac{\varphi_k - 1.2 \cdot \Delta_p(q)}{0.2 \cdot \Delta_p(q)} & \text{if } \varphi_k < \Delta_p(q). \end{cases} \tag{7}$$

For the definition of the *df*, we bin the prices into bins with a quadratically increasing width as shown in Figure 1. The bin index for the price p is defined as

$$\mathrm{bin}(p) = \left\lfloor \frac{\sqrt{p}}{2} \right\rfloor. \tag{8}$$

This is based on the assumption, that with increasing prices, the tolerance for two book prices to be comparable is larger. The *df* is then defined as the number of documents with a price in a given bin. Since a book only has a single value for the price, the *ff* and the document length are always 1.

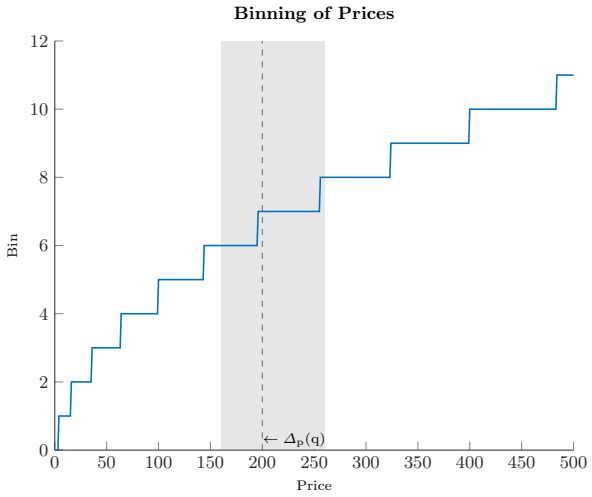

Fig. 1. Binning of prices to compute the document frequencies as well as the fuzzy query range with the average price of the books of the topic creator $\Delta_p(q)$.

Number of Pages For the number of pages of the books, we defined the *ff*, *df* and the document length as well as the query, analogous to the price.

3 Experimental Evaluation

Our goal in the experiments is to use the information present in the non-textual modalities to refine the result lists so that they reflect the users preferences.

3.1 Experimental Setup

For the textual baseline we used Lucene[3] for indexing and searching. For all modalities we used BM25 with a document length normalization parameter $b = 0.75$ and a *ff* saturation parameter $k_1 = 1.2$. For the textual modalities, we used the built-in *EnglishAnalyzer*, which removes a small set of stopwords and stems terms using the Porter stemming algorithm. For the re-ranking, we used a linear

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

combination of the scores from the different modalities. Hereby, the weights of the linear combination sum up to one.

$$\mathrm{RSV}_{\mathrm{BM25}} = \alpha \cdot \mathrm{RSV}^{\mathrm{text}}_{\mathrm{BM25}} + \beta \cdot \mathrm{RSV}^{\mathrm{rating}}_{\mathrm{BM25}} + \gamma \cdot \mathrm{RSV}^{\mathrm{price}}_{\mathrm{BM25}} + \delta \cdot \mathrm{RSV}^{\mathrm{pages}}_{\mathrm{BM25}} \quad (9)$$

In order to validate the effectiveness of our approaches and to find the optimal re-ranking parameters, we used the topics and relevance assessments from SBS 2015.

For our participation to INEX SBS 2016 track, we built six runs by applying different configurations (the re-ranking parameters equal to zero are omitted):

- **Run1**: Textual baseline using BM25 with example based relevance feedback using 35 terms and read book filtering with re-ranking parameters: $\alpha = 1$.
- **Run2**: Textual baseline re-ranked with a query-independent BM25 model for ratings with re-ranking parameters: $\alpha = 0.7818, \beta = 0.2182$.
- **Run3**: Textual baseline re-ranked with a user catalog based BM25 model for the number of pages with re-ranking parameters: $\alpha = 0.3118, \delta = 0.6882$.
- **Run4**: Textual baseline re-ranked with a user catalog based BM25 model for the price with re-ranking parameters: $\alpha = 0.2332, \gamma = 0.7668$.
- **Run5**: Textual baseline re-ranked with a user catalog based BM25 model for the price and the number of pages with re-ranking parameters: $\alpha = 0.2225, \gamma = 0.3033, \delta = 0.4742$.
- **Run6**: Textual baseline re-ranked with a query-independent BM25 model for ratings and a user catalog based BM25 model for the price and the number of pages with re-ranking parameters: $\alpha = 0.265, \beta = 0.045, \gamma = 0.225, \delta = 0.465$.

In the next section we discuss the evaluation results of our official submission.

3.2 Results and Discussion

Table 2 summarizes our official results of SBS 2016 evaluated using nDCG@10 (Normalized Discounted Cumulative Gain), MRR (Mean Reciprocal Rank), MAP (Mean Average Precision) and R@1000 (Recall), with nDCG@10 being the official evaluation measure.

The submitted runs using all non-textual modalities to re-rank the textual baseline (Run6), the run using the price and the number of pages (Run5) as well as the run using the number of pages (Run3) significantly improve the nDCG@10 over the textual baseline (Run1). The significance is computed using a paired randomization test [7] with significance level $\alpha = 5\%$. Using just the number of pages (Run3) leads to the highest nDCG@10 amongst our submitted runs. Using just the ratings for the re-ranking (Run2) increases the nDCG@10 over the textual baseline, although not significantly. Our re-ranking with the scores calculated based on the price (Run4) does not help to find a better ranked list.

To further analyze the results, we also evaluated the performance of the non-textual modalities on their own. Therefore, we used the documents retrieved

Table 2. Official results at SBS 2016. The runs are ranked according to nDCG@10.[4]

Rank	Run	Features	nDCG@10	MRR	MAP	R@1000
25	Run3	text, pages	0.0674	0.1512	0.0472	0.2556
26	Run6	text, price, pages, ratings	0.0667	0.1499	0.0462	0.2556
27	Run5	text, price, pages	0.0665	0.1442	0.0461	0.2556
30	Run2	text, ratings	0.0584	0.1332	0.0419	0.2556
31	Run1	text	0.0561	0.1251	0.0396	0.2556
32	Run4	text, price	0.0542	0.1114	0.0386	0.2556

with the textual baseline and ranked them only based on the score of each non-textual modality. This will not lead to a fully textual baseline independent ranked list (e.g. the recall will not change), however it gives an indication how well they would perform on their own. Using this approach the nDCG@10 for the ratings is 0.0206, for the price it is 0.0258 and for the number of pages 0.0135. Surprisingly, we see that the price on its own results in the highest nDCG@10, although this is not reflected in the runs that combine the non-textual modalities with the textual baseline. We also trained the weights for modalities with the relevance assessments for the 2016 task, and found, that with the optimal weights, the textual baseline can also be improved by taking the price into account. Hence, the weights chosen based on the 2015 task, are not optimal. Nevertheless, the nDCG@10 for the runs using the number of pages (0.0706) and the ratings (0.0647) using optimal weights is still higher than for the run with the price (0.0596). This shows, that either there is a higher information overlap between the price and the textual modality than between the other modalities and the text, or the linear combination merging is not as effective for the price as for the others.

4 Conclusions

In this paper, we described our participation to the suggestion track of the INEX SBS 2016 lab. We investigated how the weighting scheme BM25 can be applied to non-textual, continuous modalities. Therefore, we proposed a method to discretize the continuous modalities in order to define a document frequency and a fuzzy query that takes into account that the query does not require an exact match. By using our approach on the ratings, prices and number of pages, we showed that the effectiveness of the system can be significantly increased over the textual baseline using a simple linear score combination. However, the performance of our random forest based learning to rank approach from 2015, can not be reached.

Our experiments, have shown that the merging the scores of the prices with the textual scores leads to a smaller improvements as could be expected based

[4] We have underlined any statistically significant differences in performance according to nDCG@10 to the textual baseline (Run1) resulting from a paired randomization test [7] (significance level $\alpha = 5\%$).

on the performance of the non-textual modalities individually. So far, we did not yet investigate the merging in more depth. It is possible that a different merging method could improve the merging with the price. For example, we could use a non-linear combination of the scores, or a per-query normalization strategy, like the z-score [6], to avoid that the per-query optimal weights are far apart.

Further, we would like to investigate if the function used for the fuzzy search is the best possible. We could for example use different parameters or a non-linear falloff of the weighting.

So far, we approximated the user preferences by the average price and number of pages of the books read by the user. However, it could also be possible to construct the query such that each price and number of pages is part of the query and therefore the loss of information due to the averaging is avoided.

References

1. Bogers, T., Koolen, M., Jaap, K., Kazai, G., Preminger, M.: Overview of the inex 2014 social book search track. In: Conference and Labs of the Evaluation Forum. pp. 462–479 (2014)
2. Imhof, M., Badache, I., Boughanem, M.: Multimodal social book search. In: Sixth International Conference of the CLEF Association, CLEF (2015)
3. Moulin, C., Barat, C., Ducottet, C.: Fusion of tf. idf weighted bag of visual features for image classification. In: Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on. pp. 1–6. IEEE (2010)
4. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)
5. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)
6. Savoy, J., Berger, P.Y.: Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, chap. Monolingual, Bilingual, and GIRT Information Retrieval at CLEF-2005, pp. 131–140. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
7. 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 Conference on information and knowledge management. pp. 623–632. ACM, New York, NY, USA (2007)