# Geometric Perspectives of the BM25
## (Extended Abstract)*

Giorgio Maria Di Nunzio

Department of Information Engineering – University of Padua
dinunzio@dei.unipd.it

**Abstract.** In this paper, we present the initial findings about a possible geometric interpretation of the BM25 model and a comparison of the BM25 with the Binary Independence Model (BIM) on a two-dimensional space. A Web application was developed in R to show an example of this geometric view on a standard TREC collection. The application is accessible at the following link:
http://gmdn.shinyapps.io/shinyRF04

## 1 Introduction

The Binary Independence Model (BIM) [4] is a probabilistic retrieval model that considers documents as binary vectors and ranks the documents according to their probability of relevance given a query. The BIM assigns a weight $w_i$ to each term $t_i$ that appears in both the query and the document:

$$w_i = log\left(\frac{p_i}{q_i}\frac{(1-q_i)}{(1-p_i)}\right) \ , \tag{1}$$

where $p_i$ (or $q_i$) is the probability that a relevant (or non-relevant) document contains the term $t_i$. The estimates of these probabilities are:

$$p_i = \frac{r_i + \alpha}{R + \alpha + \beta} \ , \ q_i = \frac{n_i - r_i + \alpha}{N - R + \alpha + \beta} \tag{2}$$

where $r_i$ is the number of relevant documents that contain $t_i$, $n_i$ the number of documents that contain term $t_i$, $R$ and $N$ the number of relevant documents and the total number of documents, respectively. The parameters $\alpha$ and $\beta$ are used to smooth $p_i$ and $q_i$ in order to avoid arithmetical anomalies (in [4], $\alpha = \beta = 0.5$).

The BM25 model goes one step further by introducing the frequency of the term and the length of the document in the weight of the term $t_i$ [5]:

$$w_i' = \frac{tf_i}{tf_i + K} \cdot w_i \tag{3}$$

where $tf_i$ the frequency of the term $t_i$ in the document, and $K$ is a function of some parameters about the global statistics of the collection of documents:

$$K = k_1 * ((1-b) + b * dl/\Delta) \tag{4}$$

* This work is an extended abstract of [3]

where $k_1$ and $b$ are two parameters (usually set to 1.2 and 0.75, respectively), $dl$ is the length of document $d$, and $\Delta$ is the average document length.

In this paper, we want to study the problem of the optimisation of the parameters of two models, the BIM and the BM25, and to show a direct comparison of the two models by means of a visual interpretation of probabilities based on the idea of Likelihood Spaces [6, 1]. For this purpose, we have developed a Web application which allows users to be directly involved in the optimisation of the retrieval function and to study the effect of the variation of the parameters by means of visual inspection. [1] As a showcase, the test collection used for this application is based on the TREC2004 Robust collection. [2]

## 2 Mathematical Background

The BIM ranks documents according to the probability of relevance ($R = 1$) given a document $d$ and a query $q$, $P(R = 1|d, q)$. This probability can be approximated by the sum of the weights $w_i$ defined in Eq. 1 (see [5]):

$$P(R = 1|d, q) \approx \sum_{t_i \in d \cap q} w_i \tag{5}$$

The BM25 ranking formula can be expressed with the same sum over the terms that appear in the document and the query, replacing $w_i$ with $w_i'$ of Eq. 3.

In the two-dimensional representation of probabilities, we keep $P(R = 1|d, q)$ distinct from the probability of a document being not relevant $P(R = 0|d, q)$. With some algebraic manipulation (see [2]), we obtain the following decision (or ranking) function:

$$\underbrace{\sum_{t_i} \log\left(\frac{p_i}{1-p_i}\right)}_{x} - \underbrace{\sum_{t_i} \log\left(\frac{q_i}{1-q_i}\right)}_{y} > 0 \tag{6}$$

which is an alternative interpretation of the relevance weight of a document of the original work [4]. The two sums, $x$ and $y$, can be interpreted as two coordinates of a two-dimensional space, and documents are ranked according to the value of the difference of the two sums. With a more general approach (described in [2]) which involves Bayesian Decision Theory, we can add two more parameters $M$ and $Q$:

$$\underbrace{M \sum_{t_i} \frac{p_i}{1-p_i}}_{x} + Q - \underbrace{\sum_{t_i} \frac{q_i}{1-q_i}}_{y} > 0 \tag{7}$$

which can be interpreted as a ranking line (or a decision line) $y < Mx + Q$. This formulation allows us to study the problem on a two-dimensional space where documents are represented by two coordinates shown and ranking can be optimised according to the parameters of the decision line.

[1] `http://gmdn.shinyapps.io/shinyRF04`

[2] `http://trec.nist.gov/data/t13_robust.html`

## 3 Description of the Interface

The Web application that we developed takes into account many important parameters which characterise the two retrieval models:

- we can change the smoothing parameters $\alpha$ and $\beta$ and see how the probabilities $p_i$ and $q_i$ change (for both BIM and BM25);
- we can decide whether the ranking is computed over the terms of the document or the terms of the query, and study the differences when pseudo-relevance feedback is available (both for BIM and BM25).
- we can change the BM25 parameters $k_1$ and $b$;
- we can change the proportion of training documents to estimate $p_i$ and $q_i$ and change the number of terms of the vocabulary (both BIM and BM25);
- we can adjust the decision line by changing the angular coefficient $M$ (and the intercept $Q$ which does not affect ranking).

The main window is split into two parts: the sidebar with the interaction widgets on the left and the main panel with the output on the right (in Fig. 1 we show only half of the interface for space limits). [3]

**Interaction** The user can interact with the following widgets

1. Select the topic of interest from the drop-down menu.
2. Select the retrieval model (if BM25 is not selected, the BIM is on).
3. Change the value of the parameters $\alpha$ and $\beta$.
4. Choose the number of folds that we need to compute the probabilities of the terms of relevant and non relevant documents.
5. Change the parameters $M$ and $Q$ of the ranking line.

**Visualization** The main panel is divided into two columns: the first column shows the results on the validation set, the second column (not shown in the figure) the results on the test set. Both columns contain the following pieces of information: the text box shows the total number of objects used for validation and the number of positive examples (red points, the pseudo-relevant documents of the chosen topic). The table shows performance measures in terms of precision-at-j (j = 5, 10, 20, 100, 500, 1000). The two-dimensional plot shows in red the relevant documents of the chosen topic (pseudo-relevant for validation, true relevant for test) and in black all the other documents of the collection.

## 4 Description of the Interface

In this paper, we have presented a Web application developed in R which allows users to interact with two retrieval models, the BIM and the BM25 models, on a standard TREC collection. The two models are projected on a two-dimensional space based on the idea of Likelihood spaces. The interactive application shows, in a real machine learning setting, how the human pattern recognition capabilities can immediately detect whether the model is close to the optimal solution or not. We believe that this interactive approach may be a crucial step in setting the initial parameters of an automatic procedure that optimises these parameters.

[3] `http://gmdn.shinyapps.io/shinyRF04`

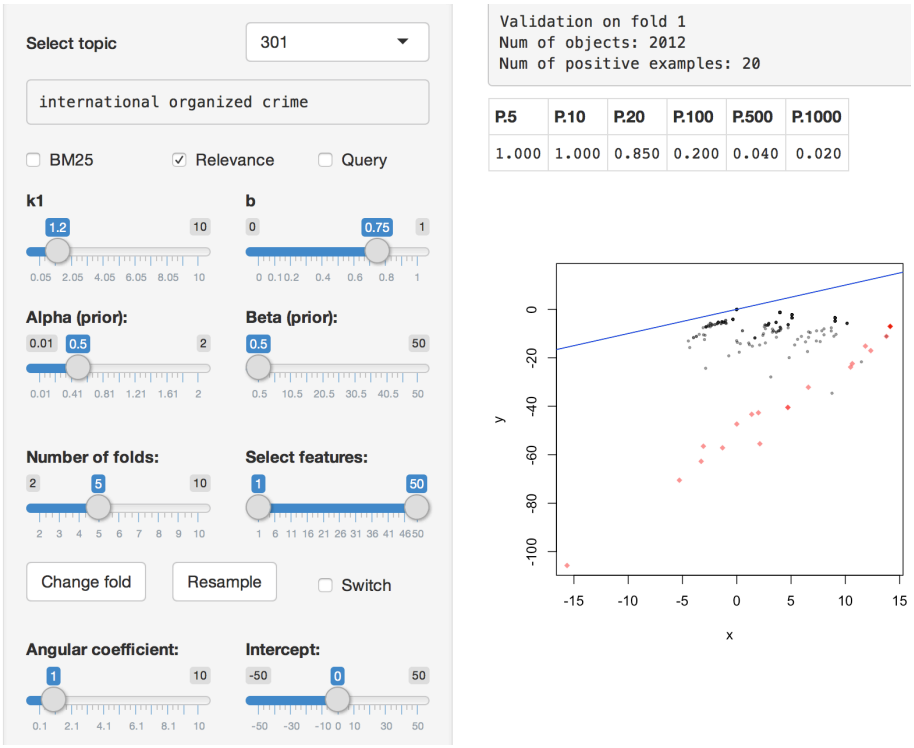


**Fig. 1.** Main window of the Web application developed in Shiny.

## References


1. Giorgio Maria Di Nunzio. Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. *Int. J. Approx. Reasoning*, 50(7):945–956, 2009.
2. Giorgio Maria Di Nunzio. A new decision to take for cost-sensitive naïve bayes classifiers. *Information Processing & Management*, 50(5):653 – 674, 2014.
3. Giorgio Maria Di Nunzio. Shiny on your crazy diagonal. In *Proceedings of the SIGIR 2015*, pages in press, http://dx.doi.org/10.1145/2766462.2767867, 2015.
4. Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. In Peter Willett, editor, *Document retrieval systems*, chapter Relevance weighting of search terms, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.
5. Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
6. Rita Singh and Bhiksha Raj. Classification in likelihood spaces. *Technometrics*, 46(3):318–329, 2004.