Are There New BM25 "Expectations"?

Emanuele Di Buccio and Giorgio Maria Di Nunzio

Dept. of Information Engineering – University of Padua
[dibuccio,dinunzio]@dei.unipd.it

Abstract. In this paper, we present some ideas about possible directions of a new interpretation of the Okapi BM25 ranking formula. In particular, we have focused on a full bayesian approach for deriving a smoothed formula that takes into account a-priori knowledge on the probability of terms. In fact, most of the efforts in improving the BM25 were done in capturing the language model (frequencies, length, etc.) but missed the fact that the constant equal to 0.5 used as a correction factor can be one of the parameters that can be modelled in a better way. This approach has been tested on a visual data mining tool and the initial results are encouraging.

1 Introduction

The relevance weighting model, also known as RSJ by the name of its creators (Roberston and Sparck-Jones), has been one of the most influential model in the history of Information Retrieval [1]. It is a probabilistic model of retrieval that tries to answer the following question:

> What is the probability that this document is relevant to this query?

'Query' is a particular instance of an information need, and 'document' a particular content description. The purpose of this question is to rank the documents in order of their probability of relevance according the Probability Ranking Principle [2]:

> If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

The probability of relevance is achieved by assigning weights to terms, the RSJ weight hereafter named as w_i, according to the following formula:

$$w_i = log \frac{p_i}{(1-p_1)} \frac{(1-q_i)}{q_i} \ , \tag{1}$$

where p_i is the probability that the document contains the term t_i given that the document is relevant, and q_i is the probability that the document contains the term t_i given that the document is not relevant. If the estimates of these

probabilities are computed by means of a maximum likelihood estimation, we obtain the following results:

$$p_i = \frac{r_i}{R} \tag{2}$$

$$q_i = \frac{n_i - r_i}{N - R} \tag{3}$$

where r_i is the number of relevant documents that contain term 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. However, this estimation leads to arithmetical anomalies; for example, if a term is not present in the set of relevant documents, its probability p_i is equal to zero and the logarithm of zero will return a minus infinity. In order to avoid this situation, a kind of smoothing is applied to the probabilities. By substituting Equation 2 and 3 in Equation 1 and adding a constant to smooth probabilities, we obtain:

$$w_i = log \frac{r_i + 0.5}{(R - r_i + 0.5)} \frac{(N - R - n_i + r_i + 0.5)}{n_i - r_i + 0.5} \ , \tag{4}$$

which is the actual RSJ score for a term. The choice of the constant 0.5 may resemble some Bayesian justification related to the binary independence model.[1] This idea is wrong, as Robertson and Sparck Jones explained in [3], and the real justification can be traced back to the work of Cox [4].

The Okapi BM25 weighting schema takes a step further and introduces the property of eliteness [5]:

> Assume that each term represent a concept, and that a given document is about that concept or not. A term is 'elite' in the document or not.

BM25 estimates the full eliteness weight for a term from the RSJ score, then approximates the term frequency behaviour with a single global parameter controlling the rate of approach. Finally, it makes a correction for document length. For a full explanation of how to interpret eliteness and integrate it into the BM25 formula read [6–9]. The resulting formula is summarised in the following way:

$$w'_i = f(tf_i) \cdot w_i \tag{5}$$

where w_i is the RSJ weight, and $f(tf_i)$ is a function of the frequency of the term t_i parametrized by global parameters.

In this paper, we concentrate on the RSJ weight and in particular to a full Bayesian approach for smoothing the probabilities and on a visual data analysis to assess the effectiveness of these new smoothed probabilities. In Section 2, we present the Bayesian framework, then in Section 3 we describe the visualisation approach; in Section 4, we describe the initial experiments on this approach. Some final remarks are given in Section 5.

[1] In this model; documents are represented as binary vectors: a term may be either present or not in a document and have a 'natural' a priori probability of 0.5.

2 Bayesian Framework

In Bayesian inference, a problem is described by a mathematical model M with parameters θ and, when we have observed some data D, we use Bayes' rule to determine our beliefs across different parameter values θ [10]:

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)} \ , \tag{6}$$

the posterior distribution of our belief on θ is equal to a likelihood function $P(D|\theta,M)$, the mathematical model of our problem, multiplied by a prior distribution $P(\theta|M)$, our belief in the values of the parameters of the model, and normalized by the probability of the data $P(D|M)$. We control the prior by choosing its distributional form along with its parameters, usually called *hyper-parameters*. Since the product between $P(D|\theta,M)$ and $P(\theta|M)$ can be hard to calculate, one solution is to find a "conjugate" prior of the likelihood function [10].

In the case of a likelihood function which belongs to the exponential family, there always exists a conjugate prior. Naïve Bayes (NB) models have a likelihood of this type and, since the RSJ weight is related to the Binary Independence Model which is a multi-variate Bernoulli NB model, we can easily derive a formula to estimate the parameter θ. The multi-variate Bernoulli NB model represents a document d as a vector of V (number of words in the vocabulary) Bernoulli random variables $d = (t_1, ..., t_i, ..., t_V)$ such that:

$$t_i \sim Bern(\theta_{t_i}) \ . \tag{7}$$

We can write the probability of a document by using the NB assumption as:

$$P(d|\theta) = \prod_{k=1}^{V} t_i = \prod_{k=1}^{V} \theta_i^{x_k}(1-\theta_i)^{1-x_k} \ , \tag{8}$$

where x_i is a binary value that is equal either to 1 when the term t_i is present in the document or to 0 otherwise. With a Maximum Likelihood estimation, we would end up with the result shown in Equation 2 and 3; instead, we want to integrate the conjugate prior which in this case of a Bernoulli random variable is the *beta* function:

$$beta_i = \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \ , \tag{9}$$

where i refers to the ith random variable t_i. Therefore, the new estimate of the probability of a term t_i that takes into account the prior knowledge is given by the posterior mean of Eq. 6 (see [10] for the details of this result). For the relevant documents we obtain:

$$\hat{\theta}_{t_i|rel} = \frac{r_i + \alpha}{R + \alpha + \beta} = \hat{p_i} \ , \tag{10}$$

where $\hat{p_i}$ is the new estimate of the probability p_i. Accordingly, the probability of a term in the non-relevant documents is:

$$\hat{\theta}_{t_i|\overline{rel}} = \frac{n_i - r_i + \alpha}{N - R + \alpha + \beta} = \hat{q_i} \ . \tag{11}$$

With this formula, we can recall different smoothing approaches; for example, with $\alpha = 0$ and $\beta = 0$ we obtain the Maximum Likelihood Estimation, with $\alpha = 1$, $\beta = 1$ the Laplace smoothing. We can even recall the RSJ score by assigning $\alpha = 0.5$ and $\beta = 0.5$.

3 Probabilistic Visual Data Mining

Now that we have new estimates for the probabilities p_i and q_i, we need a way to assess how the parameters α and β influence the effectiveness of the retrieval system. In [11, 12], we presented a visual data mining tool for analyzing the behavior of various smoothing methods, to suggest possible directions for finding the most suitable smoothing parameters and to shed the light into new methods of automatic hyper-parameters estimation. Here, we use the same approach for analyzing a simplified version of the BM25 (that is Equation 5 ignoring the term frequency function).

In order to explain the visual approach, we present the problem of retrieval in terms of a classification problem: classify the documents as relevant or non relevant. Given a document d and a query q, we consider d relevant if:

$$P(rel|d,q) > P(\overline{rel}|d,q) \ , \tag{12}$$

that is when the probability of being relevant is higher compared to the probability of not being relevant. By using Bayes rule, we can invert the problem and decide that d is relevant when:

$$P(d|rel,q)P(rel|q) > P(d|\overline{rel},q)P(\overline{rel}|q) \ . \tag{13}$$

Note that we are exactly in the same situation of Equation (2.2) of [9] where:

$$P(rel|d,q) \propto \frac{P(d|rel,q)P(rel|q)}{P(d|\overline{rel},q)P(\overline{rel}|q)} \ . \tag{14}$$

In fact, if we divide both members of Equation 13 by $P(d|\overline{rel},q)P(\overline{rel}|q)$ (we assume that this quantity is strictly greater than zero), we obtain:

$$\frac{P(d|rel,q)P(rel|q)}{P(d|\overline{rel},q)P(\overline{rel}|q)} > 1 \ , \tag{15}$$

where the ranking of the documents is given by the value of the ratio on the left (as in the BM25); moreover, we can classify a document as 'relevant' if this ratio is greater than one.

The main idea of the two-dimensional visualization of probabilistic model is to maintain the two probabilities separated and use the two numbers as two coordinates, X and Y, on the cartesian plane:

$$\underbrace{P(d|rel,q)P(rel|q)}_{X} > \underbrace{P(d|\overline{rel},q)P(\overline{rel}|q)}_{Y} \ . \tag{16}$$

If we take the logs, a monotonic transformation that maintains the order, and if we model the document as a multivariate binomial (as in the Binary Independence Model [1]), we obtain for the coordinate X:

$$\underbrace{\sum_{i \in V} x_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + \sum_{i \in V} \log(1-\hat{p}_i)}_{P(d|rel,q)} + \underbrace{\log(P(rel|q))}_{P(rel|q)} \ . \tag{17}$$

Since we are using the Bayesian estimate $\hat{p}_i$, we can modulate it by adjusting the hyper parameters α and β of Equation 10. If we want to consider the terms that appear in the query, the first sum is computed over the terms $i \in q$, which corresponds to Equation (2.6) of [9].

We intentionally maintained explicit the two addends that are independent of the document, respectively $\sum_{i \in V} \log(1-\hat{p}_i)$ and $\log(P(rel|q))$. These two addends do not influence the ordering among documents (it is a constant factor independent of the document) but they can (and they actually do) affect the classification performance. If we rewrite the complete inequality and substitute these addends with constants we obtain: [2]

$$\sum_{i \in q} x_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + c_1 > \sum_{i \in q} x_i \log\left(\frac{\hat{q}_i}{1-\hat{q}_i}\right) + c_2 \tag{18}$$

$$\sum_{i \in q} x_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) - \sum_{i \in q} x_i \log\left(\frac{\hat{q}_i}{1-\hat{q}_i}\right) > c_2 - c_1 \tag{19}$$

$$\sum_{i \in q} x_i \underbrace{\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\frac{1-\hat{q}_i}{\hat{q}_i}\right)}_{RSJ} > c_2 - c_1 \tag{20}$$

that is exactly the same formulation of the RSJ weight with new estimates for p_i and q_i, plus some indication about whether we classify a document as relevant or not.

3.1 A simple example

Let us consider a collection of 1,000 documents, suppose that we have a query with two terms, $q = \{t_1, t_2\}$, and the following estimates:

$$\hat{p}_1 = \frac{3+\alpha}{10+\alpha+\beta} \ , \ \hat{q}_1 = \frac{17+\alpha}{990+\alpha+\beta} \ ,$$
$$\hat{p}_2 = \frac{2+\alpha}{10+\alpha+\beta} \ , \ \hat{q}_2 = \frac{15+\alpha}{990+\alpha+\beta} \ ,$$

which means that we have

[2] Note that we need to investigate how this reformulation is related to Cooper's linked dependence assumption [13].

– 10 relevant document ($R = 10$) for this query;
– 20 documents that contain term t_1 ($n_1 = 20$) and three of them are known to be relevant ($r_1 = 3$);
– 17 documents that contain term t_2 ($n_2 = 17$) and two of them are known to be relevant ($r_2 = 2$).

For the log odds, we have:

$$\phi_1 = \log\left(\frac{\hat{p_1}}{1-\hat{p_1}}\right) = \log\left(\frac{3+\alpha}{7+\beta}\right) \ , \ \psi_1 = \log\left(\frac{\hat{q_1}}{1-\hat{q_1}}\right) = \log\left(\frac{17+\alpha}{973+\beta}\right) \ ,$$
$$\phi_2 = \log\left(\frac{\hat{p_2}}{1-\hat{p_2}}\right) = \log\left(\frac{2+\alpha}{8+\beta}\right) \ , \ \psi_2 = \log\left(\frac{\hat{q_2}}{1-\hat{q_2}}\right) = \log\left(\frac{15+\alpha}{975+\beta}\right) \ .$$

Suppose that we want to rank two document d_1 and d_2, where d_1 contains both terms t_1 and t_2, while d_2 contains only term t_1. Let us draw the points in the two-dimensional space, we assume the two constants c_1 and c_2 equal to zero:

$$\begin{aligned} X_{d_1} &= x_{1,d_1} * \phi_1 + x_{2,d_1} * \phi_2 = 1 * \phi_1 + 1 * \phi_2 \simeq -2.86, \\ Y_{d_1} &= x_{1,d_1} * \psi_1 + x_{2,d_1} * \psi_2 = 1 * \psi_1 + 1 * \psi_2 \simeq -11.77, \\ X_{d_2} &= x_{1,d_2} * \phi_1 + x_{2,d_2} * \phi_2 = 1 * \phi_1 + 0 * \phi_2 \simeq -1.10, \\ Y_{d_2} &= x_{1,d_2} * \psi_1 + x_{2,d_2} * \psi_2 = 1 * \psi_1 + 0 * \psi_2 \simeq -5.80 \end{aligned}$$

where $x_{i,d_j} = 1$ if term t_i occurs in document d_j, $x_{i,d_j} = 0$ otherwise.

In Figure 1, the two points (X_{d_1}, Y_{d_1}) and (X_{d_2}, Y_{d_2}) are shown. The line is a graphical help to indicate which point is ranked first: the closer the point, the higher the document in the rank. The justification of this statement is not presented in this paper for space reasons, refer to [14] for further details. What is important here is the possibility to assess the influence of the parameter α and β on the RSJ score. The objective is to study whether these two parameters can drastically change the ranking of the documents or not. In graphical terms, if we can "rotate" the points such that the closest to the line becomes the furthest.

Moreover, there are some considerations we want to address:

– when the number of terms in the query is small, it is very difficult to note any change in the ranking list. Remember that with 'n' query terms, we can only have 2^n points (or RSJ scores). In the event of a query constituted of a single term, all the documents that contain that query term collapse in one point.
– the Okapi BM25 weight 'scatters' the documents that are collapsed in one point in the space by multiplying the RSJ score with a scaling factor $f(tf_i)$ proportional to the frequency of the term in the document. Therefore, we expect this Bayesian approach to be more effective on the BM25 rather than on the simple RSJ score.

3.2 Visualization Tool

The visualisation tool was designed and developed in R [15]. It consists of three panels:

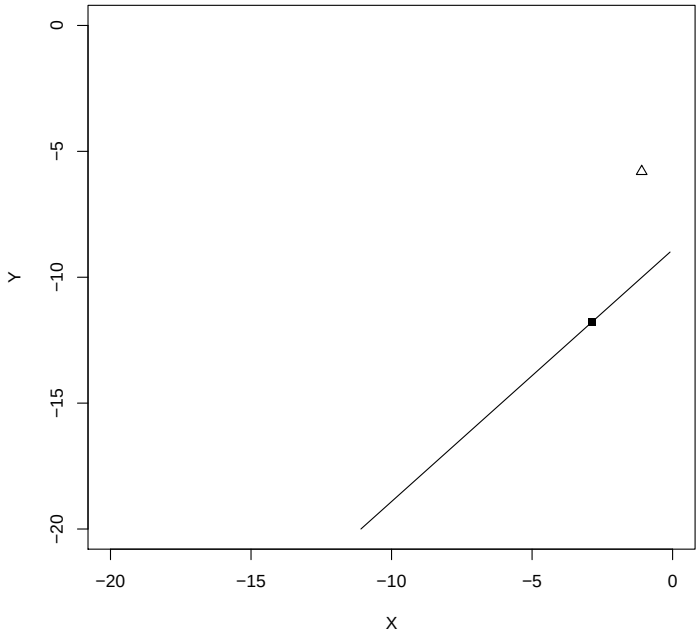

Fig. 1: Example for the documents d_1 and d_2 represented respectively by the points (X_{d_1}, Y_{d_1}) and (X_{d_2}, Y_{d_2}).

- *View Panel*: this displays the two-dimensional plot of the dataset according to the choices of the user.
- *Interaction Panel*: this allows for the interaction between the user and the parameters of the probabilistic models.
- *Performance Panel*: this displays the performance measures of the model.

Figure 2 shows the main window with the three panels. In the centre-right, there is the main view panel, the actual two-dimensional view of the documents as points, blue and red for relevant and non-relevant, respectively. The green line represents the ranking line, the closer the point the higher the rank in the retrieval list. At the top and on the left, there is the interaction panel where the user can choose different options: the type of the model (Bernoulli in our case), the type of smoothing (conjugate prior), the value of the parameters α and β. The bottom of the window is dedicated to the performance in terms of classification (not used in this experiment).

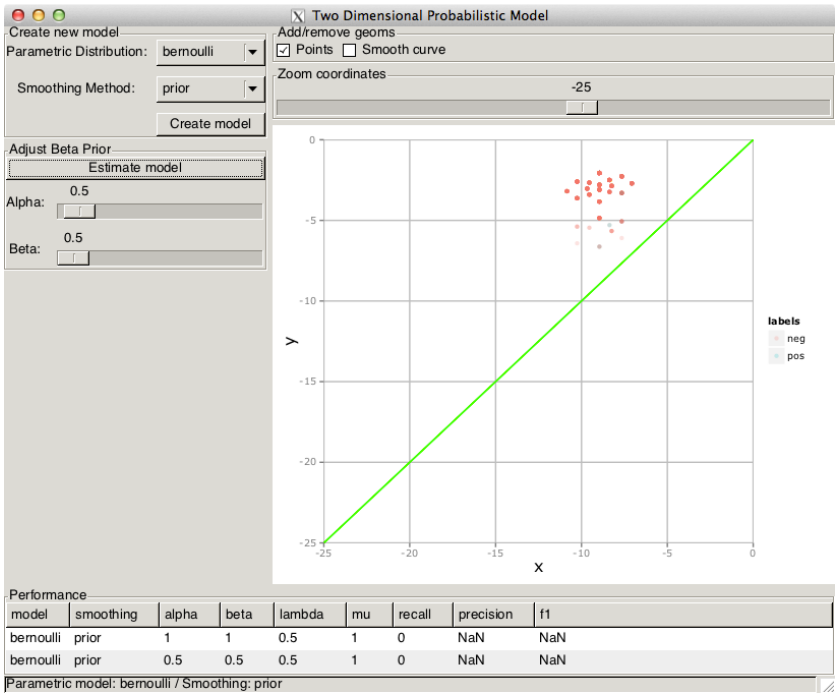

Fig. 2: Two-dimensional visualization tool: Main window.

4 Preliminary Experiments

Preliminary experiments were carried out on some topics of the TREC2001 Adhoc Web Track test collection.[3] The content of each document was processed during indexing except for the text contained inside the `<script></script>` and the `<style></style>` tags. When parsing, the title of the document was extracted and considered as the beginning of the document content. Stop words were removed during indexing.[4] For each topic we considered the set of documents in the pool, therefore those for which explicit assessment are available.

We considered two different experimental settings: (i) query-term based representation and (ii) collection vocabulary-based representation of the documents. In the former case, each document was represented by means of the descriptor extracted from the title of the TREC topics, used as queries: therefore V consisted of query terms; in the latter case V consisted of the entire collection vocabulary — both settings did not consider stopwords as part of V.

[3] http://trec.nist.gov/data/t10.web.html

[4] The stop words list is that available at the url http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

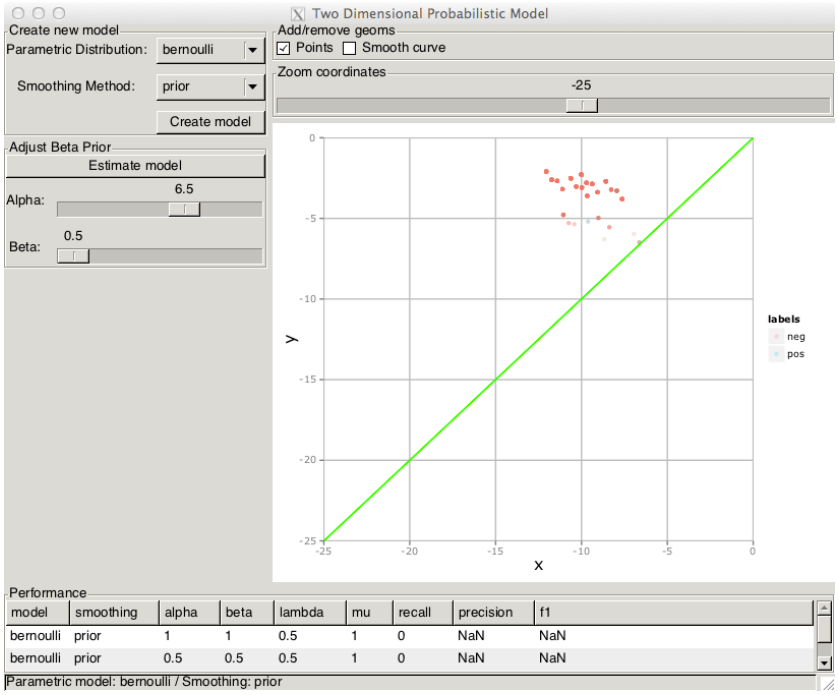

Fig. 3: Query 528: changed parameter alpha. Documents are stretched along the x-axis and rotate anti-clockwise.

In this paper, we report the experiments on topic 528. We selected this query because it contains five terms and it is easier to show the effect of the hyper-parameters. In Figure 2, the cloud of points generated by the two-dimensional approach is shown. Parameters α and β are set to the standard RSJ score constant 0.5. The line corresponds to the decision line of a classifier, and it also correspond to the 'ranking' line: imagine this line spanning the plane from right to left, each time the line touches a document, the document is added to the list of retrieved documents.

In Figure 3, the hyper-parameter α was increased and β was left equal to 0.5. When we increase α, the probability $\hat{p}_i$ tends to one, and the effect, in terms of the two dimensional plot, is that points rotate anti-clockwise. In Figure 4, the opposite effect is obtained by increasing β and leaving α equal to 0.5. In both situations, the list of ranked documents was significantly different from the original list produced by using the classical RSJ score.

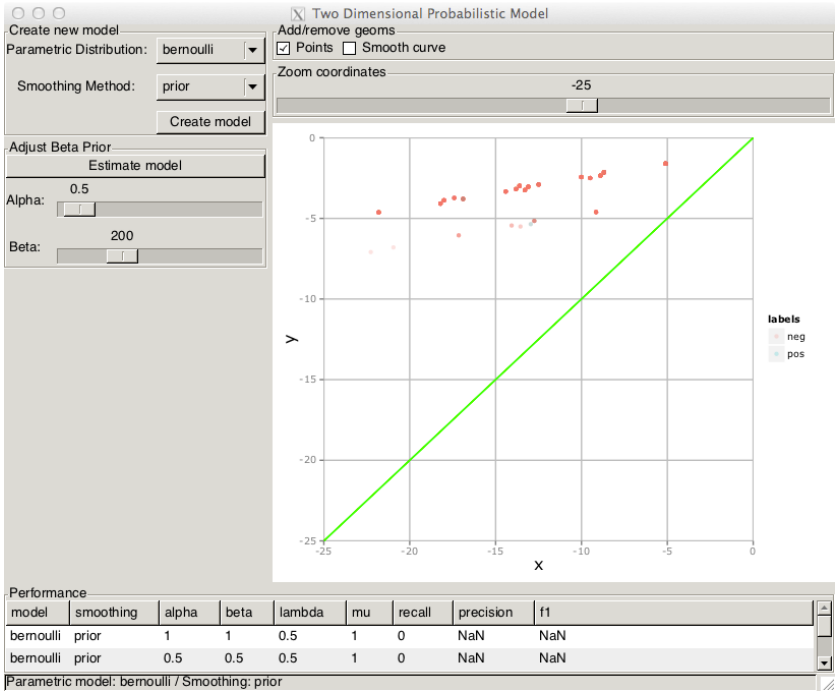

Fig. 4: Query 528: changed parameter beta. Documents are stretched along the x-axis and rotate clockwise.

5 Conclusions

This paper presents a new direction for the study of the Okapi BM25 model. In particular, we have focused on a full Bayesian approach for deriving a smoothed formula that takes into account our a-priori knowledge on the probability of terms. In fact, we think that many of the efforts in improving the BM25 were done mostly in capturing the language model (frequencies, length, etc.) but missed the fact that the 0.5 correction factor could be one of the parameters that can be modelled in a better way.

By starting from a slightly different approach, the classification of documents into relevant and non relevant classes, we derived the exact same formula of the RSJ weight but with more degrees of interaction. The two-dimensional visualization approach helped in understanding why some of the constants factors can be taken into account for the case of the classification and, more important, how the hyper-parameters can be tuned to obtain a better ranking.

After this preliminary experiment, we can draw some considerations: for the first time, it was possible to visualize the cluster of points that are generated by the RSJ scores; it was clear that very short queries tend to create a very small

number of points making it hard to perform a good retrieval; hyper-parameters do make a difference in both classification and retrieval.

There are still many open research questions we want to investigate in the future:

- so far, we have assumed that all the beta priors associated to each term use exactly the same values for hyper-parameters α and β. A more selective approach may be more effective;
- the coordinate of the points in the two-dimensional plot take into account the two constants of Equation 17. In particular, the addend $\sum_{i \in V} \log(1-\hat{p}_i)$ may be the cause of the 'rotation' of the points, hence the radical change of the ranking list;
- The current approach assumes that the value of R and r_i are known for each term in the query: indeed these values are adopted to estimate the coordinates of each document. A further research question is the effect of estimation based on feedback data on the capability of the probabilistic visual data mining approach adopted in this paper.

Acknowledgments. This work has been partially supported by the QONTEXT project under grant agreement N. 247590 (FP7/2007-2013).

References

1. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. In Willett, P., ed.: Document retrieval systems. Taylor Graham Publishing, London, UK, UK (1988) 143–160
2. Robertson, S.E.: The Probability Ranking Principle in IR. Journal of Documentation **33** (1977) 294–304
3. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage. **36** (2000) 779–808
4. Cox, D., Snell, D.: The Analysis of Binary Data. Monographs on Statistics and Applied Probability Series. Chapman & Hall (1989)
5. Robertson, S.: Understanding inverse document frequency: On theoretical arguments for idf. In: Journal of Documentation. Volume 60. (2004)
6. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Croft, W.B., van Rijsbergen, C.J., eds.: SIGIR, ACM/Springer (1994) 232–241
7. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference (TREC), Gaithesburg, USA (1994)
8. Robertson, S.E., Walker, S.: On relevance weights with little relevance information. SIGIR Forum **31** (1997) 16–24
9. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval **3** (2009) 333–389
10. Kruschke, J.K.: Doing Bayesian Data Analysis: A Tutorial with R and BUGS. 1 edn. Academic Press/Elsevier (2011)

11. Di Nunzio, G., Sordoni, A.: How well do we know bernoulli? In: IIR. Volume 835 of CEUR Workshop Proceedings., CEUR-WS.org (2012) 38–44
12. Di Nunzio, G., Sordoni, A.: A visual tool for bayesian data analyisis: The impact of smoothing on naïve bayes text classifiers. In: Proceeding of the 35th International ACM SIGIR 2012. Volume 1002., Portland, Oregon, USA (2012)
13. Cooper, W.S.: Some inconsistencies and misnomers in probabilistic information retrieval. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '91, New York, NY, USA, ACM (1991) 57–61
14. Di Nunzio, G.: Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. Int. J. Approx. Reasoning **50** (2009) 945–956
15. Di Nunzio, G., Sordoni, A.: A Visual Data Mining Approach to Parameters Optimization. In Zhao, Y., Cen, Y., eds.: Data Mining Applications in R. Elsevier (2013, In Press)