2012 International Conference on Information and Knowledge Management (ICIKM 2012)
IPCSIT vol.45 (2012) © (2012) IACSIT Press, Singapore

Optimizing the Value c in the DFR Framework for Small Collections

Fadi Yamout

Lebanese International University, Beirut, Lebanon
fadi.yamout@liu.edu.lb

Abstract. Divergence From Randomness is a methodology for constructing nonparametric models of Information Retrieval where the term weighting, DFR weight, is derived by measuring the divergence of the actual term distribution from that obtained under a random process. A parameter c was introduced in the literature as part of the second term frequency normalization, and was shown to have definite values for Web test collections. In this paper, research thru testing was done to determine the values of c that give the best precision for Text REtrieval Conference (TREC)[1] small test collections MED, CISI, Cranfield, LISA, and NPL following the submission of the query and query reformulation using relevance feedback. The values are determined for different relevance feedback models mainly pseudo, positive, and negative-relevance.

Keywords: Deviation From Randomness, Bernoulli Model, Relevance Feedback

1. Introduction

Probabilistic models were first suggested by Maron and Kuhns [1], and afterwards developed by Robertson and Sparck-Jones [2], and Van Rijsbergen [3]. Term's weights, in a probabilistic model are given by a probability [4], and the documents and queries are viewed as vectors. More details on the Probabilistic model are found in [2] and [5]. A probabilistic matching function is used to measure the similarity between the documents and the query. This matching function estimates the probability that a document will be relevant to the query. Therefore, the documents are ranked based on an estimate of the probability of relevance of a document to a query [2], as opposed to vector-space model which ranks the documents in decreasing similarity of query and document [6]. Probabilistic models have been extended in different models mainly Deviation from Randomness (DFR) [4].

2. Deviation From Randomness

DFR is a methodology for constructing nonparametric models of Information Retrieval [4] where the term weighting is derived by measuring the divergence of the actual term distribution from that obtained under a random process. One main advantage of using nonparametric approach is the generation of different models using different choices of probability distribution. DFR has five basic IR models: Bernoulli Model of Randomness, Binomial, Bose–Einstein, the Inverse document Frequency model, tf-idf (In), the Inverse term frequency model, tf-itf (IF), and the Inverse expected document frequency model , tf-expected idf (In_exp).

One of the probabilistic weighting models for IR in the DFR framework [4] is the Bernoulli model. The Bernoulli Model is approximated by two other models. The first one is the approximation of the binomial – Poisson model (P) and the second is the Divergence approximation of the binomial (D). It is shown in [4] that these two approximation models perform equally under all normalizations. In this paper, we employ the DB2 and in further work we plan to experiment with other weighting models [4] different than the Bernoulli model. The next section describes the weighting model DB2 with the corresponding equations.

[1] Information can be found at http://trec.nist.gov/

2.1. DB2 Document Weighing Model

The DB2 document weighting model is the Bernoulli Model of Randomness approximated by the Divergence approximation of the binomial (D) with first normalization using the Ratio B of two binomial distributions (B) and with the second term frequency normalization (2). The parameter c is an adjustable parameter (hyperparameter) that could be set automatically [7] [8]. The value of c differs with different sizes of topics. In [9] for example, c is set to 1 for title-only topics and 7 for long topics. In Terrier [10], a list of estimated parameter values on TREC collections is given, and accordingly, c is set to 13.13 for WT10G and 1.28 for GOV (title-only queries). However, more research was done in this paper to determine the best values for the small test collections Medline, CISI, Cranfield, LISA, and NPL.

3. Experimental Design

3.1. Test Collections

Many test collections with different numbers of documents and terms are available for testing an IR system [11]. The experiments in this paper are conducted on some of these test collections. The commonly used small test collections for testing an IR system are the Medline, Cranfield, and CISI. These are of small to medium size with just over a thousand documents with an average size of 1 Megabytes. Some of the information related to these test collections are summarized below (Table 1):

Table 1. Details of Small Test Collections

Test Collection	Size in Megabytes	No of Documents	No of queries	No of Terms	Topics
Medline	1.05	1,033	30	8,915	Medicine
Cranfield	1.40	1,400	225	4,217	Aeronautical engineering
CISI	1.98	1,460	76	5,591	Information Science
NPL	3.02	11,429	93	7,934	Electronic Engineering
LISA	3.04	6,003	35	11,291	Library & Information Science

There is a standard set of queries, and for each query, experts have chosen which documents are relevant, and this information is useful for evaluating machine performance. These documents are put in a relevance judgment list. Medline, for example, comes with 30 queries and a relevance judgment list of 696 entries (on the average 23 relevant documents per query). The table shows also the number of queries and relevance judgment lists for these collections. After the indexing process, the test collections contain different numbers of terms. Medline for example consists of 1033 documents with 8915 unique terms.

3.2. Querying and Relevance Feedback

When submitting a query in DFR, documents are ranked using the DB2 document weighting model from the DFR framework as explained in section 2 “Divergence From Randomness”. When reformulating a query in a probabilistic model, probabilistic weights of the documents’ terms are re-weighted based on the documents chosen relevant by the user. Consequently, the similarity coefficient for a given document is obtained by summing these new weights and the documents with the highest weights are highly ranked in the list. The relevant process, used in this paper for the probabilistic model uses Information-theoretic query expansion [12]. It calls the topmost documents to be assessed the “Elite set T of documents” and the most informative terms are selected by using the information theoretic approach to automatic query expansion [12] based on the Kullback-Leibler divergence function KL [9]. As a result, the documents with the highest weights are highly ranked in the list.

4. Experimental Analysis

Retrieval performance is measured using precision at recall level 0.1, precision at recall level 0.3, and the Interpolated Mean Average Precision (MAP) which is the average precision at recall levels=0.0, 0.1… 1.0 [13]. The assessment is done as follows: A query is run, resulting in a ranking of documents. In simulated-positive feedback, the relevant ones (these are found in the list of user judgments that comes with the test collection) from the top N retrieved are selected, whereas in PRF the top N documents are selected as relevant. Consequently, these relevant ones are used to modify the query and a new retrieval is done. In a simulated-negative feedback, the terms found in the non-relevant ones in the top N documents are removed from the modified query. The recall and precision figures are shown in the experiments at different levels of recall such as 0.1 (10% recall) and 0.3 (30% recall).

5. Experimental Results

For each test collection, different values of c were found for querying, pseudo, simulated positive-feedback, and simulated negative-feedback. In this paper, a single value for c was chosen for each test collection based on the following criterion: the value should give the best precision at recall values 0.1 (10% recall), 0.3 (30% recall) in addition to a high MAP value, however, at 10% recall under pseudo-relevance feedback was given a precedence. For the Medline collection, and for the querying process, a value of c between 1.2 and 2.1 gives the best precision at recall value 0.1 (precision at 10% recall equal 0.88). Whereas for simulated pseudo-feedback, the best precision at recall value 0.1 (Figure 1) is given when c is 2.8.

Medline - Query

c	0.3		1.2	↔	1.6		1.8	↔	2.1	Max
MAP	0.55		0.56	same	0.56		0.56	same	0.56	0.56
P@10	0.86		0.88	same	0.88		0.88	same	0.88	0.88
P@30	0.75		0.71	same	0.71		0.71	same	0.71	0.75

Medline - Pseudo Feedback

c	2.8	Max
MAP	0.29	0.29
P@10	0.67	0.67
P@30	0.43	0.43

Medline - Positive Feedback

1	1.1		2.5	Max
0.3	0.3		0.29	0.3
0.68	0.68		0.66	0.68
0.44	0.43		0.45	0.45

Medline - Negative Feedback

0.5	Max
0.19	0.19
0.41	0.41
0.27	0.27

Figure 1. Value of Parameter c assigned to Medline Test Collections

As for the Cranfield test collection, the best precision for pseudo-feedback at recall value 0.1 (Figure 2) is when c is between 4.8 and 5.8.

Cranfield - Query

c	0.1	0.2		1.2	↔	3.9	Max
MAP	0.09	0.09		0.08	same	0.08	0.09
P@10	0.17	0.17		0.18	same	0.18	0.18
P@30	0.11	0.11		0.09	same	0.08	0.11

Cranfield - Pseudo Feedback

1.4		4.8	↔	5.8	Max
0.08		0.07	same	0.07	0.08
0.13		0.21	same	0.21	0.21
0.1		0.08	same	0.08	0.1

Cranfield - Positive Feedback

c	0.6	0.7		1.1		3.4	↔	4.1	Max
MAP	0.06	0.06		0.07		0.06	same	0.06	0.07
P@10	0.15	0.15		0.14		0.15	same	0.15	0.15
P@30	0.05	0.06		0.11		0.07	same	0.06	0.11

Cranfield - Negative Feedback

1.7	↔	1.9	Max
0.08	same	0.08	0.08
0.13	same	0.13	0.13
0.09	same	0.09	0.09

Figure 2. Value of Parameter c assigned to Cranfield Test Collections

For the CICS test collection. The best precision for pseudo-relevance feedback at recall level 0.2 (Figure 3) is when c equal to 4.2. Almost any value of c will give a good precision for Positive-feedback. For LISA and NPL test collections the value of c is almost the same at recall level 0.1 (Figures 4 and 5) whenever Pseudo-feedback is applied; for LISA for instance, c is between 7.3 and 7.9 whereas for NPL c is between 7.4 and 7.6

CISI - Query

c	1.9		3.1	3.2	Max
MAP	0.19		0.19	0.19	0.19
P@10	0.35		0.36	0.36	0.36
P@30	0.22		0.21	0.21	0.22

CISI - Pseudo Feedback

4.2	Max
0.15	0.15
0.25	0.25
0.17	0.17

CSCI - Positive Feedback

c	1.1	↔	2.3		3.1	↔	4.5	Max
MAP	0.14	same	0.14		0.14	same	0.14	0.14
P@10	0.26	same	0.26		0.26	same	0.26	0.26
P@30	0.17	same	0.17		0.17	same	0.17	0.17

CISI - Negative Feedback

1.7	↔	1.9	Max
0.08	same	0.08	0.08
0.13	same	0.13	0.13
0.09	same	0.09	0.09

Figure 3. Value of Parameter c assigned to CISI Test Collections

LISA - Query

c	0.3		2.2	↔	2.5	Max
MAP	0.15		0.17	same	0.17	0.17
10%	0.38		0.34	same	0.34	0.38
30%	0.2		0.25	same	0.25	0.25

LISA - Pseudo Feedback

3	↔	4.6		5.5		7.3	↔	7.9	Max
0.05	same	0.05		0.05		0.05	same	0.05	0.05
0.1	same	0.1		0.11		0.11	same	0.11	0.11
0.06	same	0.06		0.05		0.05	same	0.05	0.06

LISA - Positive Feedback

c	2.9	↔	4.6		7.3	↔	7.5	Max
MAP	0.05	same	0.05		0.05	same	0.05	0.05
10%	0.12	same	0.1		0.13	same	0.13	0.13
30%	0.06	same	0.06		0.05	same	0.05	0.06

LISA - Negative Feedback

0.1	↔	2.5		5.6	↔	8	Max
0.01	same	0.01		0.01	same	0.01	0.01
0.01	same	0.01		0.01	same	0.01	0.01
0.01	same	0.01		0.01	same	0.01	0.01

Figure 4. Value of Parameter c assigned to LISA Test Collections

NPL - Query

c	3.6	↔	4	Max
MAP	0.27	same	0.27	0.27
10%	0.54	same	0.54	0.54
30%	0.34	same	0.34	0.34

NPL - Pseudo Feedback

7.4	↔	7.6	Max
0.14	same	0.14	0.14
0.33	same	0.33	0.33
0.18	same	0.18	0.19

NPL - Positive Feedback

c	3	↔	6.6		7.2	↔	8	Max
MAP	0.14	same	0.14		0.15	same	0.15	0.15
10%	0.31	same	0.31		0.34	same	0.34	0.34
30%	0.19	same	0.19		0.18	same	0.18	0.19

NPL - Negative Feedback

2.4	↔	3	Max
0.04	same	0.04	0.04
0.07	same	0.07	0.07
0.05	same	0.05	0.05

Figure 5. Value of Parameter c assigned to NPL Test Collections

Table 2 shows the values of the parameter c used in the experiments. The precision values are at recall value 0.1 (10% recall) using Pseudo Feedback.

Table 2. Value of Parameter c assigned to Small Test Collections

Test Collection	Precision @ 10	Value of parameter c
Medline	0.67	2.8
Cranfield	0.21	4.8 ↔ 5.8
CISI	0.25	4.2
LISA	0.11	5.5 and 7.3 ↔ 7.9
NPL	0.33	7.4 ↔ 7.6

Figures 1, 2, 3, 4 and 5 show the best values of c for the small test collections with precision given at recall value 0.1 (10% recall), and recall value 0.3 (30% recall) along with the MAP values. In these tables the maximum precisions are shaded with grey.

6. Conclusions

In this paper, the DB2 weighting model in the DFR framework is used as the probabilistic model. In the weighting scheme of this model, the parameter c was introduced as part of the second term frequency normalization, exclusively in the variable tfn (see equation 5). It was shown in [7] [8] that this parameter has definite values for WT10G and WT18G. Research thru testing was done in this paper to determine the values of c that give the best precision after submitting the query and reformulating the query using relevance feedback. The test is done on small test collections for different relevance feedback features mainly pseudo, positive, and negative-relevance feedback.

The best values of c are determined for these small test collections with precision given at recall value 0.1 (10% recall), and recall value 0.3 (30% recall) along with the MAP values. These values are shaded with grey in figures 1, 2, 3, 4 and 5. Table 2 shows the final best values of the parameter c at recall value 0.1 (10% recall) chosen in the experiments

7. References

[1] Maron M. E. and Kuhns J. L. (1997) “On Relevance Probabilistic Indexing and Information Retrieval” Journal of the Association for Computing Machinery. 15. pp 8-36. 1960. Reprinted in Readings in Information Retrieval. K. Sparck-Jones and P Willet (eds). Morgan Kaufman. pp 39-46. 1997.

[2] Robertson S E & Sparck-Jones K (1976). Relevance weighting of search terms, Journal of the American Society of Information Science, pp.129-146

[3] van Rijsbergen C J (1979). Information retrieval, Buttersworth, London

[4] Amati G & van Rijsbergen C J (2002). Probabilistic models of information retrieval based on measuring divergence from randomness. ACM Transactions on Information Systems, 20(4):357-389

[5] Sparck-Jones K., Walker S. and Robertson S.E. (2000) “A Probabilistic Model of Information Retrieval: Development and Comparative Experiments”. Information Processing & Management 36(6), pp. 779-840, 2000.

[6] Salton G. and Lesk, M. (1971) "Information Analysis and Dictionary Construction" In Salton G. editor, The SMART Retrieval System, Chapter 6. Prentice-Hall, New Jersey

[7] He B. and Ounis I. (2003) “A Study of Parameter Tuning for Term Frequency Normalization”, in Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, 2003.

[8] He B. and Ounis I. (2005) “Term Frequency Normalisation Tuning for BM25 and DFR Model”, in Proceedings of the 27th European Conference on Information Retrieval (ECIR'05), 2005.

[9] Amati, G (2003). “Probabilistic Models for Information Retrieval based on Divergence from Randomness”, PhD thesis, Department of Computing Science, University of Glasgow.

[10] Ounis I., Amati G., Plachouras V., He B., MacDonald C., and Johnson D. (2005). “Terrier Information Retrieval Platform”, In Proceedings of the 27th European Conference on IR Research (ECIR 2005) Volume 3408 of Lecture Notes in Computer Science, pages 517-519. Springer, 2005.

[11] Baeza-Yates R. and Ribeiro-Neto B. (1999) “Modern Information Retrieval”. New York: Addison-Wesley. P 118 1999

[12] Carpineto C., De Mori R., Romano G., and Bigi B., (2001) “An Information Theoretic Approach to Automatic Query Expansion”, ACM Transactions on Information Systems, 19(1):1-27, 2001.

[13] Kent A., Berry M. Leuhrs F.U. and Perry J.W. (1955) “Machine Literature Searching VIII: Operational Criteria for Designing Information Retrieval Systems". American documentation, 6(2):93-101 1955