

Probability Models for Information Retrieval based on Divergence from Randomness

Giambattista Amati

Thesis submitted for the degree of Doctor of Philosophy,

Department of Computing Science

Faculty of Information and Mathematical Sciences

University of Glasgow

June 2003



© Giambattista Amati
Glasgow, 9th June 2003

Abstract

This thesis devises a novel methodology based on probability theory, suitable for the construction of term-weighting models of Information Retrieval. Our term-weighting functions are created within a general framework made up of three components. Each of the three components is built independently from the others. We obtain the term-weighting functions from the general model in a purely theoretic way instantiating each component with different probability distribution forms.

The underpinning idea on which we are able to systematically construct the term-weighting models is based on the notion of divergence from randomness. The leading theme of the divergence-from-randomness approach is that the informative content of a term can be measured by examining how much the term-frequency distribution departs from a “benchmark” distribution, that is the distribution described by a random process.

Following this idea, the first two components of the framework provide an explanation to the duality existing in Information Retrieval between the distributions of topic-terms in a small set of documents (the elite set of a topic) and in the rest of the collection. The third component deals with the term-frequency normalization and is able to compare term frequencies within documents of different lengths. As a consequence, different probability distributions can be used in the framework of the divergence-from-randomness approach. Our experiments utilise some of them to show that the framework is sound and robust and generates different but highly effective Information Retrieval models.

The thesis begins with investigating the nature of the statistical inference involved in Information Retrieval. We explore the estimation problem underlying the process of sampling. De Finetti’s theorem is used to show how to convert the frequentist approach into Bayesian inference and we display and employ the derived estimation techniques in the context of Information Retrieval.

We initially pay a great attention to the construction of the basic sample spaces of Information Retrieval. The notion of single or multiple sampling from different populations in the context of Information Retrieval is extensively discussed and used throughout the thesis. The language modelling approach and the standard probabilistic model are studied under the same foundational view and are experimentally compared to the divergence-from-randomness approach.

In revisiting the main information retrieval models in the literature, we show that even language modelling approach can be exploited to assign term-frequency normalization to the models of divergence from randomness.

We finally introduce a novel framework for the query expansion. This framework is based on the models of divergence-from-randomness and it can be applied to arbitrary models of IR, divergence-based, language modelling and probabilistic models included.

We have done a very large number of experiments and results show that the framework generates highly effective Information Retrieval models.

Acknowledgments

I would like to express my gratitude to my supervisor, Prof. Cornelis Joost Van Rijsbergen, for his scientific input, knowledge, advice, understanding, feedback, generosity and for allowing me to freely explore new ideas.

I am grateful to Dr. Roderick Murray-Smith for his kindness and optimism.

I am especially indebted to Juliet Van Rijsbergen who greatly improved the readability of the manuscript by adding scientific comments, providing suggestions for improving its structure, and even going through the mathematics and pointing out where my thoughts and explanations were unclear.

I thank the Committee members, Prof. Norbert Führ, Dr. Ronald R. Poet and Dr. Lewis M. Mackenzie for having appreciated my work.

And finally my thanks to my wife, Carmen, who encouraged me to believe only the important things in my life.

Legenda

Symbols and notations

D	a text collection
t	a term
q	a query
d	a document
$w(q d)$	the weight of the query q given the document d
tf_q	the term–frequency of t in the query q
tf	the term–frequency of t in the document d
E	a sample of the collection
E_q	<i>the elite set of the query</i> , the set of topmost documents satisfying the query q according to the weight $w(q _)$
E_t	<i>the elite set of the term</i> , the set of documents containing the term t
N	the number of documents in the collection D
$avg.l$	the average length of a document in the collection
l, l_d	the length of the document d
F, F_t, F_E	the total number of tokens of t in the collection, in E_t , and in an arbitrary subset E
$TotFr_D, TotFr_E$	the total number of tokens in the collection D and in a subset E of D
$p_D, p_D(t)$	the relative frequency $\frac{F}{TotFr_D}$ of t in the collection
$p_d, p_d(t)$	the relative frequency $\frac{tf}{l_d}$ of t in the document
n, n_t	the document–frequency, the cardinality of E_t , $n = n_t = E_t $

Symbols and notations

$B(F, k, p)$	the binomial distribution of F trials with probability p of success and k successes
n_e	the number of documents containing a term according to the binomial distribution, $N \cdot (1 - B(F_t, 0, p))$
r, r_t	the number of relevant documents containing the term t
R	the number of relevant documents of a query
μ	the parameter of the Dirichlet priors
α	the parameter of the query expansion
c	the parameter of the term frequency normalization
	$H2$
D	the Divergence of two distributions
χ	the χ divergence of two distributions
KL	the Kullback-Leibler divergence of two distributions
$Inf(t E),$	the informative content of the term in E
$Inf_E(t)$	

Basic Divergence-based Models

D	the Divergence approximation of the binomial
P	the Poisson approximation of the binomial
B_E	the Bose-Einstein distribution
G	the geometric approximation of the Bose-Einstein
$I(n)$	the Inverse Document Frequency model
$I(F)$	the Inverse Term Frequency model
$I(n_e)$	the Inverse Expected Document Frequency model

First Normalization Models

L	the Laplace normalization
B	the Bernoulli ratio normalization

Second Normalization: Term Frequency Normalization

$H1$	the uniform distribution of term frequencies
$H2$	the logarithmic normalization
$H3$	the Dirichlet normalization
Z	the Zipfian normalization

Normalized Models

$DL1$	the divergence basic model D , normalized by Laplace normalization L and by term frequency normalization H1
\vdots	\vdots
$I(n_e)BZ$	the Inverse Expected Document Frequency basic model, normalized by the Bernoulli ratio normalization B and by the Zipfian term frequency normalization Z

Contents

1	Theoretical Information Retrieval	16
1.1	The intention of this Thesis	19
1.2	The origins of the proposal	20
1.3	The generating term–weighting formula	23
1.4	The first component: informative content	23
1.4.1	An exemplification of informative content: Bernoulli model of divergence from randomness	25
1.5	The second component: apparent aftereffect of sampling	26
1.5.1	An example of aftereffect model: Laplace’s law of succession	28
1.6	The third component: term-frequency normalization	29
1.7	The probabilistic framework	32
1.8	The naming of models	33
1.9	The component of query expansion	33
1.10	Experimental work	35
1.11	Outline of the Thesis	35
2	Probability distributions for divergence based models of IR	37
2.1	The probability space in Information Retrieval	39
2.1.1	The sample space V of the terms	40
2.1.2	Sampling with a document	41
2.1.3	Multiple sampling: placement of terms in a document collection	43
2.2	Binomial distribution: limiting forms	45
2.2.1	The Poisson distribution	45

2.2.2	The divergence D	46
2.2.3	Kullback-Leibler divergence	47
2.2.4	The \mathcal{X} divergence	47
2.3	The hypergeometric distribution	49
2.4	Bose-Einstein statistics	50
2.4.1	The geometric distribution approximation	51
2.4.2	Second approximation of the Bose-Einstein statistics	52
2.5	Fat-tailed distributions	53
2.5.1	Feller-Pareto distributions	56
2.6	Mixing and compounding distributions	59
2.6.1	Compounding the binomial with the Beta distribution	59
2.7	Summary and Conclusions	61
3	The estimation problem in IR	62
3.1	Sampling from different populations	63
3.2	Type I sampling	63
3.3	Type II sampling: De Finetti's Theorem	65
3.3.1	Estimation of the probability with the posterior probability	66
3.3.2	Bayes-Laplace estimation	66
3.3.3	Maximum likelihood estimation	67
3.3.4	Estimation with the loss function	68
3.3.5	Small binary samples	68
3.3.6	Multinomial selection and Dirichlet's priors	69
4	Models of IR based on divergence from randomness	71
4.1	Basic Models	72
4.2	The informative content Inf_1 in the basic probabilistic models	76
4.3	The basic binomial model	76
4.3.1	The model P	78
4.3.2	The model D	78
4.4	The basic Bose-Einstein model	78
4.4.1	The model G	78

4.4.2	The model B_E	79
4.5	The tf-idf model	79
4.5.1	The model $I(n)$	80
4.5.2	The model $I(n_e)$	80
4.5.3	The model $I(F)$	81
4.6	First normalization of the informative content	81
4.6.1	The first normalization L	83
4.6.2	The first normalization B	84
4.7	Relating the aftereffect probability $Prob_2$ to Inf_1	86
4.8	First Normalized Models of Divergence from Randomness	88
4.8.1	Model PL	88
4.8.2	Model PB	89
4.8.3	Model DL	89
4.8.4	Model DB	89
4.8.5	Model GL	89
4.8.6	Model GB	89
4.8.7	Model B_EL	89
4.8.8	Model B_EB	90
4.8.9	Model $I(n)L$	90
4.8.10	Model $I(n)B$	90
4.8.11	The model $I(n_e)L$	90
4.8.12	The model $I(n_e)B$	90
4.8.13	Model $I(F)L$	90
4.8.14	Model $I(F)B$	90
5	Related IR models	91
5.1	The vector space model of IR	91
5.2	The standard probabilistic model of IR	92
5.2.1	The 2-Poisson model	95
5.2.2	The $BM25$ matching function	96
5.3	Inference Network Retrieval	99
5.4	The language model	101

<i>CONTENTS</i>	10
5.4.1 Ponte and Croft's model	102
5.5 Language model: Dirichlet's prior for IR	104
5.6 Language model: mixtures of probability distributions	106
6 Term-frequency normalization	108
6.1 Related works on term-frequency normalization	115
6.2 Term-frequency normalizations H1 and H2	116
6.2.1 A discussion on the Second Normalization H2	119
6.3 Term-frequency normalization based on the classical Pareto distribution	121
6.3.1 The relationship between the vocabulary and the text length	124
6.3.2 Example: the Paretian law applied	126
6.3.3 The Paretian term-frequency normalization formula	126
6.4 Term-frequency normalization Dirichlet priors	128
7 Normalized models of IR based on divergence from randomness	131
7.1 A derivation of BM25 and INQUERY formula	133
7.2 Experimental data	134
7.3 Experiments with long queries	135
7.3.1 Results from experiments with long queries	137
7.4 Experiments with short queries	146
7.4.1 Results from experiments with short queries	149
7.5 Conclusions	151
8 Query expansion	153
8.1 Introduction	153
8.2 Term-weighting in the expanded query	155
8.3 Query expansion	156
8.4 Rocchio's method	160
8.5 The Binomial Law for query expansion	161
8.6 The hypergeometric model of query expansion	162
8.6.1 Approximations of the Binomial	163
8.7 Query expansion with the Bose-Einstein distribution	165
8.8 Normalized term-frequency in the expanded query	165

<i>CONTENTS</i>	11
8.9 Experiments with query expansion	166
8.10 Results from query expansion	166
8.11 Conclusions	168
9 Conclusions	172
9.1 Summary of the results from the experiments	172
9.2 Research Contributions and Future Research	173
A Evaluation	177
A.1 Evaluation measures	177
B Functions and probability distributions	182
B.1 Functions and distributions	182

List of Figures

1.1	Informative content of the term “osteoporosis” with the Poisson approximation (model P) of the Bernoulli model over TREC-8 collection.	27
1.2	Informative content gain (model PL) of the term “osteoporosis” over TREC-8 collection.	29
1.3	Score distribution using the term-frequency normalization component (model $PL2$) over the TREC-8 collection.	31
2.1	Relation between the logarithms of term rank and term-frequency in TREC-10 collection.	54
2.2	Relation between the logarithms of term rank and term-frequency in TREC-8 collection.	55
6.1	The average correlation coefficient between the document length and the term-frequencies normalized by H2 of Formula 6.10. The sample of terms are from the queries of TREC-7, TREC-8 and TREC-10 respectively.	119
6.2	Comparison of the the correlation coefficient as in Figure 6.1 to the performance. The model is $I(n_e)B2$. The best matching value of MAP for TREC 7 data is 0.1904 at $c = 13$. Best Pr@10 is 0.4400 at $c = 8$	120
6.3	Comparison of the the average correlation coefficient as in Figure 6.1 to the performance. The model is $I(n_e)B2$. The best matching value of MAP for TREC 10 data is 0.2107 at $c = 12$. Best Pr@10 is 0.3720 at $c = 7$.121	

List of Tables

1.1	Comparison among the basic model (P), the gain (PL), and the term-frequency normalization model ($PL2$) with the TREC-8 data. The evaluation measures are defined in Appendix B.1.	32
1.2	Models are made up of three components. For example $B_E B2$ uses the limiting form B_E of Bose-Einstein Formula 2.33, normalized by the incremental rate B of the Bernoulli process of Formula 4.22. The within-document term-frequency is normalized under hypothesis $H2$ of Formula 6.10	34
5.1	The contingency table in the probabilistic model.	94
6.1	The Correlation coefficient between length and term-frequency with terms of the first 12 queries of TREC -7	110
6.2	Terms of TREC -7 in decreasing ordering of term-frequency-document length correlation.	111
6.3	Performance of $B_E L$ with different term-frequency normalizations on TREC-10 data.	129
6.4	The performance of the Pareto term-frequency normalization for TREC-8 data. The run $Z = 0.2942$ is that relative to the value of Z corresponding to the slope $\alpha = 1.399$ for the 2 GB collection of TREC-8.	129
6.5	The performance of the Pareto term-frequency normalization for TREC 9 data. The run $Z = 0.2972$ is that relative to the value of Z corresponding to the slope $\alpha = 1.365$ for the wt10g collection.	129

7.1	The probability $\Phi(\beta)$ is the probability computed by the standard normal distribution that a random document has length $\left \frac{l}{avg_l} - 1 \right < 1$ in a collection with mean avg_l and variance σ^2	134
7.2	Results from TREC-1 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	138
7.3	Results from TREC-2 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	139
7.4	Results from TREC-3 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	140
7.5	Results from TREC-6 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	141
7.6	Results from TREC-6 with the long queries and removing long documents. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	142
7.7	Best performing models for each test collection and for different precision measures. The basic probability models $I(F)$, D and B_E are not considered here, as they do not differ significantly from their alternative approximations $I(n_e)$, P and G respectively. See Section 1.8 and Table 1.2 for an explanation of the model names.	142
7.8	Results from TREC-7 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	143
7.9	Results from TREC-8 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.	144
7.10	Baselines for short queries of TREC-8	148
7.11	Baselines for short queries of TREC-9	148

7.12	Baselines for short queries of TREC-10	148
7.13	Comparison of models with TREC-10 data without using Porter's stemming algorithm.	149
8.1	The highest informative terms for the query 502 (Prime factor?) of TREC-10 data. The last column shows the weights of the terms in the new expanded query.	159
8.2	Precision obtained by different expansion methods averaged over all models and TREC collections.	169
8.3	Increment of precision obtained by different expansion methods averaged over all models and TREC collections.	169
8.4	Best expansion methods for each model and TREC collection. The best values for each TREC data are in bold.	170