Document-Query Duality Meets Maximum Likelihood: The Answer is 3/15/24?

David Bodoff
Hong Kong University of Science and Technology
dbodoff@ust.hk

A maximum likelihood approach to relevance feedback is introduced. This approach has the additional benefit of resolving document-query duality. The main idea is that in order to know whether and how much to modify a document and/or query in response to relevance feedback data, we need to have error models for documents, queries, and relevance feedback data. A maximum likelihood approach can then use these models to "decide" which document and/or query representations to modify. If this method is to be used for probabilistic models, then the probabilistic models must allow us to modify document and query representations in response to feedback data. Previously, the probabilistic models would never modify these representations, only give them low weights. Finally, we report preliminary results of an empirical test of this method using Cranfield data.

1 Introduction

This work grew out of an attempt to resolve query-oriented with document-oriented approaches in IR. This "duality" [1] was first highlighted in a series of publications by Robertson, but has plagued almost every approach to information retrieval. In this paper we present a theoretical approach we have been pursuing to resolve this duality. Our method introduces probabilistic models and maximum likelihoods in a manner they have not, to our knowledge, been previously used in IR research. Also, our method resolves the document-query duality in relation to the concept of relevance. Our paper thus revolves around three issues that are of central importance to the theory of IR -- document/query duality, relevance, and probabilistic models. For these reasons, we believe this work helps to integrate theoretical results in IR. In addition, this paper reports the first empirical application of our approach.

2 Document-Query Duality

The document-oriented probabilistic approach goes back to [2], while the document and query oriented approaches -- and a unified model -- were both presented in [3]. Robertson has cogently framed the duality issue in the context of these probabilistic models [1, 4]. In [5], we review how this duality issue also pertains to other approaches to IR. In the context of the Vector Space Model (VSM), for example, the duality arises as two competing approaches to relevance feedback data -- query modification $Q_{new} = Q_{old} + D$ versus document modification $D_{new} = D_{old} + Q$. In the context of logical models, for another example, the duality arises as the difference between $P(d \rightarrow q)$ and $P(q \rightarrow d)$. In all these cases, the issue is not fully resolved, in the sense that it appears we should combine the document-oriented view with the query-oriented view, but it is not obvious how.

It is our observation that the duality issue is more pointed for the probabilistic models, than for the vector space model. It is perhaps for this reason that the duality question has been associated mostly with the probabilistic models. In this section we describe in what sense the problem is more pointed for the probabilistic models than for VSM. This is important because our solution to the duality question is influenced by the VSM view. If our solution is to apply to the probabilistic models as well, then the probabilistic models need to adopt some VSM thinking, as we will see below.

As it was first explored in (Robertson et al 1982), the question is how to estimate a joint distribution that best fits two marginal distributions. Conceptually, the problem arises because the so-called document- and query-oriented models don't actually estimate anything about documents or queries, but *about relevance*, *of* documents *to* queries, or vice versa.

In contrast, let us consider how the problem arises in VSM. In response to feedback data, the query-oriented view would adjust queries according to a formula such as Qnew=Qold+D, where D is a relevant document. The document-oriented approach would do the opposite, adjusting the document according to Dnew=Dold+Q, where Q is a query to which it is relevant.

Continuing with VSM, the question arises, since both approaches appear valid, how should we combine them? If we want to apply both in sequence, then the question arises, in which sequence? For example, when we apply Qnew=Qold+D, what value do we use for "D", Dold or Dnew (vice versa when adjusting documents)? The simplest method of combining the two approaches is to simply apply both in isolation:
Dnew=Dold+Qold
Qnew=Qold+Dold
The solution we have been working on, allows us to apply both simultaneously:
D*new*=Dold+Q*new*
Q*new*=Qold+D*new*

This second solution is based on Multi-Dimensional Scaling (MDS). Relevancy data is used as an indication that the target "distance" between a document and a query should be small or smaller (as in [6]). A method based on MDS then simultaneously finds positions for all documents and queries in the space, in a manner that minimizes total distances between objects whose target distances are small. Two additional terms in the MDS objective allow us to account for the initial representations, with penalties for moving a document or a query from their initial positions. Let a binary function *B* denote known relevance of a document to a query, B: D×Q→0/1; let *r* be a distance function that predicts the relevance of a document to a query, and let s_D and s_Q be distance functions between two document or two query representations, respectively. Then the MDS style objective function we proposed is to find values for all Di_new and Qj_new, to minimize:

$$\Sigma\Sigma\ r(Di\_new, Qj\_new)*B(Di, Qj) + \Sigma\ s_D(Di\_new, Di\_old) + \Sigma\ s_Q(Qj\_new, Qj\_old) \qquad (1)$$

Note that this function requires an estimate of every dimension of every single document and every single query representation. Each of these terms can also be weighted. The idea is simply that for every Di,Qj for which B(Di,Qj)=1, we want then to be as close as possible (first term), subject to penalties for moving an object too much (latter terms). This simultaneous estimation of documents and queries is a second solution to the duality problem.

The serial and simultaneous solutions to the duality problem in VSM, are not applicable to the probabilistic models as they are usually understood. The first solution, i.e. to separately apply each approach, simply makes no sense for the probabilistic models. *In VSM*, the document-oriented approach actually estimates a document, while the query-oriented approach estimates a query. Is it possible to estimate both separately? Certainly. But the probabilistic models *each estimates a probability of relevance*, and one is left to wonder which probability of relevance to use. One cannot separately "do both", as one can with the VSM one-sided approaches.

The second method we outlined above has some analogy with Robertson et al.'s proposed model 3. But there is one difference. Our method *compromises* the two views, so that each one is qualified, whereas Robertson et al.'s model 3 *combines* the two marginal views by estimating a joint distribution that best fits them. But why are the marginal views known and fixed in probabilistic approaches but not in VSM? Underlying this difference is the following important observation: Whereas VSM methods use relevance feedback to estimate document and query representations, the probabilistic models take as given all document and query representations, and use relevance feedback to estimate conditional probabilities of relevance. The probabilistic models will give a low weight to a document or query term that does not seem to predict relevance, but they will not modify the document or query representation.

It is our belief that the probabilistic approaches have unnecessarily straight-jacketed themselves by accepting as given the document and query estimates. By accepting these as immutable, the effect of relevance data is two known marginal probabilities of relevance, for which we can only hope for a good joint fit. There are simply not as many degrees of freedom, as compared with an approach that allows us to adjust our document and query estimates (i.e. representations) in addition to our estimate of relevance.

How is it possible to estimate documents and queries in the probabilistic model? If VSM can use relevance feedback data to estimate better document and query representations, why can a probabilistic model not use probability theory to estimate better document and query representations? To adopt the VSM philosophy within a binary probabilistic framework, we would use relevance feedback data to estimate the probability that a binary document or query term is represented correctly, and if that seemed improbable, then that bit would be flipped from zero to one or vice versa. This is the corollary to VSM document and query modifications in response to relevance feedback data. Fuhr [7] points exactly in this direction, by explicitly modelling $P(\mathbf{x} \mid d_m)$, the probability that representation **x** is correct for document d_m, separately from the probability of relevance $P(R| \mathbf{x}, f_k)$. That model contributed to the development of our approach. We went one step further: instead of (merely) estimating whether a document or query representation appears correct, we go ahead and correct it if it doesn't, as in VSM.

The approach outlined above for VSM can work identically for the probabilistic models: Let a binary function *B* show actual relevance of a document to a query, $B: D \times Q \rightarrow 0/1$; let $r: D \times Q \rightarrow [0,1]$ denote a function to predict probability of relevance (e.g. Model 0), and let s_D: Di_new × Di_ old → [0,1] denote a probability that Di_new is the correct representation for the document we originally observed as Di_old (respectively s_Q for queries). Then the MDS style objective function we proposed is to ***find binary values for every bit in all Di_new and Qj_new***, to minimize:

$\Sigma\Sigma$ r(Di_new, Qj_new)*B(Di, Qj) + $\Sigma\, s_D$(Di_new, Di_old) + $\Sigma\, s_Q$(Qj_new, Qj_old)

The idea, once again, is that any Di,Qj pair for which B(Di,Qj)=1 ought to be moved closer together -- this time, in terms of a probabilistic distance (first term) -- subject only to probabilistic penalties for adjusting objects from their initial representations.

When a new query is submitted to this system, it will be matched using function *r* against documents with their new representations (the new query will be in its original form, since when it is first submitted we do not have any feedback data for it). Model 0 is the most obvious choice for *r*. Are we using the feedback data available for the specific document and specific query, as is the aim of Model 3? Yes, because the query-specific and document-specific information in the relevancy assessments has been used to update the representations of the individual documents and queries. The function in Model 0 is general to all documents and queries, but the data to which it is applied -- what terms are present/absent in the specific document and query in question -- have been adjusted for each individual document and query, based on the totality of evidence.

As we wrote in [8], this is a radical departure from the usual approach to parameter estimation in probabilistic models, in that feedback data is used to separately estimate documents, queries, and relevance, rather than take as given all but relevance. The advantage is that we are not stuck with the original marginal probabilities, and we are not left wondering how to combine the two fixed marginal probabilities of relevance. Instead, we have separately estimated individual document and query representations, and predict relevance based on these adjusted representations.

We have contrasted the VSM versus probabilistic approaches to the use of feedback data, and explained how the VSM approach can also be used with probabilistic models. The purpose of this discussion has been to clarify further our idea for resolving the document-query duality. Our MDS-based proposal is to allow both document and query representations as simultaneously free variables, being pulled away from their initial positions by the relevance feedback "distance" data. We have pointed out that to do this within the probabilistic models, would require us to estimate the probability of each document and query representation, along with a single probability of relevance.

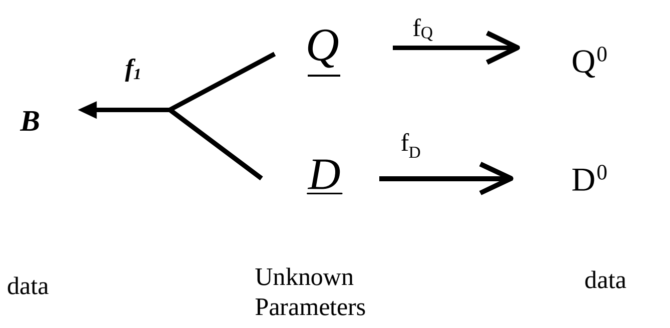
The only shortcoming of all this is that we have not presented any particular theoretical basis for the MDS objective function we introduced. Even a one-sided approach that estimates only documents or queries based on relevance feedback and the initial representations, ought to have a theoretical basis for the particular algorithm employed. If we aim to simultaneously estimate both documents and queries, this issue is just more glaring -- what is the theoretical basis for the MDS objective function above? And even if we accept its basic structure, what is the theoretical basis for the particular forms of functions *r* and *s*, and for the weights we may give to each?

Explicit Models

To fully resolve these shortcomings, we derive the MDS function introduced above in a maximum likelihood framework. More generally, we show how to derive different functions for simultaneously estimating documents and queries, depending on one's probabilistic models.

In [5] we introduce the following model (figure 1):

Figure 1



Hypothesized Probabilistic Meta-Structure

In this model, the final representations $\underline{D}$ and $\underline{Q}$ are the parameters we seek, using three sources of data: the relevancy feedback data "B", and the initial document and query representations D^0 and Q^0. This diagram suggests the explicit modelling of f_1, f_D, and f_Q as stochastic functions that relate the three sources of data we have (i.e. B, D^0, Q^0) with the parameters we seek -- i.e. the final $\underline{D}$ and $\underline{Q}$. Following the notation of a probability density as f(x; θ) with observations x and parameters θ, we denote these stochastic functions as $f_D(D^0; \underline{D})$, $f_Q(Q^0; \underline{Q})$, and $f_1(B; \underline{D},\underline{Q})$, respectively. What remains is only to choose the form of these models, either through theoretical or empirical means. Once these stochastic relationships are modelled, we can construct a maximum likelihood function to simultaneously estimate both documents and queries based on the relevancy feedback data.

For example, we may adopt the following three probability models:

$f_1(B; \underline{D},\underline{Q}) = \exp(\cos(\underline{D},\underline{Q})^B)$ as a simplified approximation to: $\exp(\cos(\underline{D},\underline{Q})^B + (1-\cos(\underline{D},\underline{Q}))^{(1-B)}$

$f_D(D^0; \underline{D}) = k*\exp(\cos(\underline{D},D^0))$

$f_Q(Q^0; \underline{Q}) = k*\exp(\cos(\underline{Q},Q^0))$

Then after taking logs, the objective is to find representations for all $\underline{D}$ and $\underline{Q}$ to maximize

$$L = \Sigma\Sigma\cos(\underline{D},\underline{Q}) + \Sigma\cos(\underline{D},D^0) + \Sigma\cos(\underline{Q},Q^0) \qquad (3)$$

In this case, we have modelled *r* and *s* from expression (1) above as cosine functions.

Note that every dimension of every single document and query is a free parameter to be estimated. Additional weighting factors are also possible. In terms of intuitive meaning, the model of f1, for example, means that relevance is a stochastic function of the cosine between the correct representations of the document and query. The choice of the same cosine function for modelling f_D and f_Q is similar to the idea of self-relevance -- a final document or query representation should have a high predicted relevance to its original location.

For binary probabilistic models, we might use Model 0 rather than cosine as the basis for all these three distributions:
$f_1(B; \underline{D},\underline{Q})$ = Model 0 of $(\underline{D},\underline{Q})$
$f_D(D^0; \underline{D})$ = Model 0 of $(\underline{D},D^0)$ (i.e. treat either D^0 or $\underline{D}$ as the "query")
$f_Q(Q^0; \underline{Q})$ = Model 0 of $(\underline{Q},Q^0)$ (i.e. treat either Q^0 or $\underline{Q}$ as the "document")

The idea of this objective function is as follows: Documents and queries should be represented in a way that relevant document-query pairs have a high score using Model 0. At the same time, the final document and query representations should have a high predicted relevance to their original representations. Intuitively, this means we "flip a bit" in the document or query representation if that would explain relevancy data more than it would cause surprise (in the likelihood sense) regarding the author's adding/omitting this extra term.

There is one additional benefit of this approach -- when we observe which stochastic models result in improved performance, then "for free" we will have discovered actual models that describe "what they say, and what they really mean". We may, for example, find that the model f_Q has higher (lower) variance than the model for f_D, indicating that query terms should be given relatively less (or more) weight in the simultaneous estimation of documents and queries. We believe such results may have far-reaching implications for IR theory and practice. We have been eagerly waiting for such results since this line of research was initiated. The next and final section discusses our very preliminary results.

Empirical Work

We have recently begun the first empirical test of our simultaneous approach, using Cranfield data. Our first efforts use VSM models. We have adopted the stochastic models:
$f_1(B; \underline{D},\underline{Q}) = k*\exp(\alpha\cos(\underline{D},\underline{Q})^B)$
$f_D(D^0; \underline{D}) = k*\exp(\beta\cos(\underline{D},D^0))$
$f_Q(Q^0; \underline{Q}) = k*\exp(\phi\cos(\underline{Q},Q^0))$
and the resulting likelihood function:
$\max_{\underline{D},\underline{Q}}: L = \Sigma\Sigma\cos(\underline{D},\underline{Q})^B + \Sigma(\beta/\alpha)\cos(\underline{D},D^0) + \Sigma(\phi/\alpha)\cos(\underline{Q},Q^0)$

After maximizing this function for training queries, we test the performance of the new document representations on set-aside queries. Any resulting performance benefit is good news in itself, but we view as even more exciting the fact that performance benefits are evidence to the validity of the model.

In our first few experimental runs, we have found performance degradation with $\beta/\alpha = \phi/\alpha = 1$, and modest performance benefits with $\beta/\alpha = 5$ and $\phi/\alpha = 8$. Currently, the search for the best weightings is heuristic. More generally, the choice of the probability functions -- we chose functions proportional to cosine -- is also currently heuristic. For the relevancy model $f_1(B; \underline{D},\underline{Q}) = k*\exp(\alpha\cos(\underline{D},\underline{Q})^B)$, a value of $\alpha=3$ results in a relevancy patterns similar to the Cranfield relevancy data. Taken together, our results provide evidence that "document authors choose words proportional to exp(15*cosine) of their true intention, and information searchers submit queries whose words are proportional to exp(24*cosine) of their true intention. We believe this kind of result may have very far-reaching implications and benefits to research in IR. After working with the VSM framework, we plan to use the same approach for binary probabilistic models, using Model 0 as outlined above.

Practicalities

We had previously encountered two sources of doubt over our approach: It was believed that it is impossible to estimate such a large number of parameters with such a small amount of data, and it was believed that it would be impossible to estimate such a large number of parameters in reasonable time. Our simulations had always shown otherwise. Our objective function on the Cranfield data takes about 5-10 iterations of a MATLAB optimisation routine, and this takes about 3-4 hours on a dual processor Pentium III 733 MhZ PC/Linux. Our total optimisation approach is of course intended to improve the document index offline, and is not intended for ad-hoc query adjustment, for which one-sided heuristics are ideal. Performance remains as an issue, but the approach is not obviously infeasible. In terms of the ability to estimate parameters with such a small amount of data, this is due to the elegant structure shown in figure 1: the three stochastic elements triangulate on the free parameters, and convergence is swift. The likelihood function is constructed such that if a document or query has no relevancy feedback data, it is not moved at all from its original position, and it is moved proportionally more in accordance with the available evidence. For this reason, our formulation does not suffer from the problems often encountered when trying to apply a formal parameter estimation procedure with small amounts of data.

Conclusion

We have proposed a method for simultaneous estimation of document and query representations, in response to relevance feedback data. The approach is grounded in a maximum likelihood framework, and will allow us to discover models that accurately describe how people chose words relative to what they really intend. In order to apply to the probabilistic models, one must estimate the probability of the correctness of a document or query representation, and modify those representations if that will help explain the observed document, query, and relevancy data.

References

1. Robertson, S.E., *Query-Document Symmetry and Dual Models.* Journal of Documentation, 1994. **50**(3): p. 233-238.
2. Maron, M.E. and J.L. Kuhns, *On Relevance, Probabilistic Indexing and Information Retrieval.* Journal of the ACM, 1960. **7**: p. 216-244.
3. Robertson, S.E., M.E. Maron, and W.S. Cooper, *Probability of Relevance: A Unification of Two Competing Models For Document Retrieval.* Information Technology -- Research and Development, 1982. **1**: p. 1-21.
4. Robertson, S.E., M.E. Maron, and W.S. Cooper. *The Unified Probabilistic Model for IR*. 1982. Berlin: Springer-Verlag.
5. Bodoff, D., *et al.*, *A Unified Maximum Likelihood Approach to Document Retrieval.* Journal of the American Society for Information Science and Technology, 2001(forthcoming).
6. Bartell, B.T., G.W. Cottrell, and R.K. Belew, *Representing Documents Using an Explicit Model of their Similarities.* Journal of the American Society for Information Science, 1995. **46**(4): p. 254-271.
7. Fuhr, N., *Models for Retrieval with Probabilistic Indexing.* Information Processing and Management, 1989. **25**(1): p. 55-72.
8. Bodoff, D., *A Re-Unification of Two Competing Models for Document Retrieval.* Journal of the American Society for Information Science, 1999. **50**(1): p. 49-64.