2. Logistic Models

2.1 General description

The object of a probabilistic retrieval model is to provide some method of estimating the probability that a particular document is relevant. The information which is to be used to obtain this estimate is essentially the specific combination of index terms assigned to (or present in) the particular document.

Thus from the point of view of the probabilistic model, any two documents with exactly the same set of index term assignments are indistinguishable. The index terms in effect serve to allocate documents to exclusive classes or cells, each all being defined by a particular combination of terms. The probability of relevance is the probability p_j that a randomly selected document in cell j will be relevant.

A "pure" probabilistic approach would require past evidence about the relevance of documents from any particular cell, in order to make statements about new documents in that cell. But since most cells are empty and most of the rest contain only one or two documents, this is not in general possible in information retrieval. So the usual procedure has been to devise a model on the basis of some assumptions of independence between index terms, which imply relationships between the various probabilities p_j . Assuming such a model means that the estimate of p_j for cell j may be based on evidence about the relevance of documents from other cells. For example, in the Robertson/Sparck Jones model, strong independence assumptions are made; as a consequence, the estimate of p_j uses evidence from all previously-judged documents containing <u>any one</u> of the terms that define cell j.

The model or class of models proposed here actually reverses part of this process. Instead of making independence assumptions and deriving relationships between the p_j's, it makes direct assumpumptions about these relationships.

The relationships are expressed in terms of the logistic transform of the probabilities, so the class of models might generally be described as "logistic". One particular model in the class, the "independent logistic" model, is roughly equivalent to the Robertson/ Spark Jones model. Dependencies between terms can also be introduced. Further, the framework includes a natural range of estimation methods, as will be seen.

2.2 Mathematical Description

2.2.1 Notation and likelihood function

We assume that there are d distinguishable classes (or cells) of documents. We also assume that in unit time, the number of documents in class j, n_j, has a Poisson distribution with mean c_j, and that the n_j are independent. It will be seen below that the particular assumption about the n_j is not important. We also assume that the probability of a document being relevant given that it is in class j is p_j . The number of relevant documents in cell j in unit time is called k_j. It is not hard to see that, conditional on n_j, k_j will have a binomial distribution with parameters p_j and n_j.

Using vector notation $\underline{n}$ is the d-dimensional column vector with entries n_j, similarly $\underline{p}$, $\underline{c}$, $\underline{k}$ etc. We can write down the likelihood function for $\underline{p}$ and $\underline{c}$ given $\underline{k}$ and $\underline{n}$

$$L(\underline{p}, \underline{c}; \underline{k}, \underline{n}) = \prod_{j=1}^{d} \frac{(p_j c_j)^{k_j}}{k_j\,!} \; \frac{((1-p_j)c_j)^{(n_j-k_j)}}{(n_j-k_j)!} \; e^{-c_j}$$

and the log-likelihood

$$\ell(\underline{p},\ \underline{c};\ \underline{k},\ \underline{n}) = \sum_{j=1}^{d} \Big\{ k_j \log P_j + (n_j - k_j)\log(1-p_j) + n_j \log c_j - c_j - \log k_j! - \log (n_j - k_j)! \Big\}$$

The logistic models are models of the form $(\log \frac{P_j}{1-p_j}) \in \mathfrak{M}$ where $\mathfrak{M}$ is a vector subspace of $\mathbb{R}^d$. For brevity we will just talk about the model $\mathfrak{M}$. If we reparametrize by writing

$$\pi_j = \log \frac{P_j}{1-P_j} \quad \text{for} \quad 1 \leq j \leq d$$

then the log-likelihood becomes

$$\ell(\underline{\pi},\ \underline{c},\ \underline{k},\ \underline{n}) = \sum_{j=1}^{d} \Big\{ k_j \pi_j - n_j \log(1+e^{\pi_j}) + n_j \log c_j - c_j - \log k_j! - \log (n_j - k_j)! \Big\}$$

2.2.2 Estimation

If we consier the maximum likelihood estimates (*m.l.e.*) of $\underline{\pi}$ and $\underline{c}$ (assuming $\underline{\pi} \in \mathfrak{M}$) we see that the *m.l.e.* of $\underline{c}$ is $\underline{n}$ and does not depend on $\mathfrak{M}$. The *m.l.e.* of $\underline{\pi}$ if it exists is denoted $\underline{\hat{\pi}}_m$ and is that $\underline{\rho} \in \mathfrak{M}$ which maximises

$$f(\underline{\rho}) = f(\underline{k}, \underline{n}, \underline{\rho}) = \sum_{j=1}^{d} \Big\{ k_j \rho_j - n_j \log(1+e^{\rho_j}) \Big\} \qquad (1)$$

(The existence of the *m.l.e.* is discussed below.)

We may easily introduce Bayesian ideas into this model. In general, we multiply the likelihood function by some factor which describes the prior distribution, thus obtaining the posterior distribution. Then instead of a maximum likelihood estimate, we may consider a maximum posterior estimate (or posterior mode). This will introduce an additive factor into the function $f(\underline{p})$ which must be maximised - an example is given below.

There are several possible ways of finding a maximum, i.e. of deriving the maximum likelihood or maximum posterior distribution estimates. It is necessary, however, to use an iterative procedure.

2.2.3 Existence of estimates

Considering the maximum likelihood estimates, it is very convenient to define $\underline{N}$ to be the diagonal matrix where diagonal elements are the n_j. Now suppose there is a vector $\underline{\sigma} \in \mathcal{M}$ such that $\underline{N}\ \underline{\sigma} = 0$. Then clearly $f(\underline{\rho} + \underline{\sigma}) = f(\underline{\rho})$ for all $\underline{\rho}$. Hence a necessary condition for the existence of an $m.l.e.$ for $\underline{\pi}$ is that there is no non-zero $\underline{\sigma} \in \mathcal{M}$ $s.t.$ $\underline{N}\ \underline{\sigma} = \underline{0}$. Imposing this condition on $\underline{N}$ and $\mathcal{M}$ will ensure that $f(\underline{\rho})$ is strictly concave and hence has at most one finite maximum and that any stationary value of $\underline{\rho} \in \mathcal{M}$ is a maximum but it is not sufficient to ensure f has a maximum in $\mathcal{M}$.

It is unlikely that there is any easily verified condition on $\underline{N}$, $\underline{k}$ and $\mathcal{M}$ which is equivalent to the existence of an $m.l.e.$ (although some results can be proved). Further, it is likely that in many practical situations, for some models which one might want to use, the $m.l.e.$ does not exist. The Bayesian prior, however, may be used as a mechanism to ensure the existence of an estimate (now a maximum posterior estimate).

2.2.4 Discussion

The above descibes a general class of models for the probabilities p_j. The particular model is expressed as a linear subspace $\mathcal{M}$ of the space generated by the logistic transforms π_j. An appropriate way to obtain estimates of the parameters of any model from data would be the maximum likelihood method. However, this method will not necessarily possess solutions at all, particularly where there is little data. We may further introduce prior information into the model by standard Bayesian methods, and maximise the posterior distribution rather than the likelihood. This procedure may be used to force the existence of a solution. Also, the more restricted the model (i.e. the smaller the dimension of $\mathcal{M}$), the more likely it is that the maximum likelihood method will give a solution.

In section 2.3 this formalism is used to set up an "independence" model roughly equivalent to the Robertson-Sparck Jones model. We then discuss the possible theoretical advantages and disadvantages of the class of models defined here.

Given the form of the class of models, the question naturally arises : under what conditions should one try to relax or restrict the model (i.e. increase or reduce its dimension)? In section 4 we return to this question in theoretical terms.

2.3 Example : the Independent Logistic Model

We now consider a specific model which corresponds, in some sense, to the Robertson/Sparck Jones model.

Our basic assumption is that probability of relevance will be modelled by a simple sum-of-weights, each weight being associated with one of the matching terms. That is, we will assume that

$$\pi_j = \Sigma \; w_t + w_o$$

where w_o and $\{w_t\}$ are the quantities to be estimated, and the sum is over all query terms t whose presence defines the particular cell j . This model can be simply expressed as a vector subspace of $\mathbb{R}^d$, as required.

To formulate this model more precisely, we define Q as the set of query terms, and $T \subseteq Q$ as a set of terms defining a particular cell, i.e. those documents containing all the terms T and none of the terms $Q \setminus T$. (In fact each "cell" may be subdivided by the presence or absence of terms other than query terms; however, give this model, such subdivision makes no difference to the maximum likelihood estimates derived from (1).) In order to deal with the w_o weight in the formula for π_j, we may define a dummy term t_o which occurs in all documents, and which is automatically included in T. Then $\mathcal{M}$ is defined as follows :

$$\pi_T = \sum_{t \, \varepsilon \, T} w_t \qquad (2)$$

So, from (1),

$$f(\underline{w}) = \sum_{T \subseteq Q} \left[k_T \sum_{t \, \varepsilon \, T} w_t - n_T \log \left(1 + \exp\left(\sum_{t \, \varepsilon \, T} w_t\right)\right) \right]$$

$$= \sum_{t \, \varepsilon \, Q} w_t \sum_{T : t \, \varepsilon \, T} k_T - \sum_{T \subseteq Q} n_T \log \left(1 + \exp\left(\sum_{t \, \varepsilon \, T} w_t\right)\right) \qquad (3)$$

$\sum_{T : t \varepsilon T} k_T$ is the total number of relevant documents assigned to term t. The w's which maximize $f(\underline{w})$ give the maximum likelihood estimates of the π's.

We may add a Bayesian prior : suppose for example that the w_t's are assumed to be independent, normally distributed with means μ_t and common variance σ^2 (the means may, for example, be derived from the traditional collection frequency weights). Then the function to be maximised is :

$$g(\underline{w}) = \sum_{t \varepsilon Q} w_t \sum_{T : t \varepsilon T} k_T - \sum_{T \subseteq Q} n_T \log\left(1 + e^{\sum_{t \varepsilon T} w_t}\right) - \qquad (4)$$

$$1 \sum (\quad)^2$$

2.4 Comparison of present models with previous approaches

We first describe how the Independent Logistic model might be applied in a practical situation, and then discuss its apparent advantages and disadvantages, in comparison with the Robertson/ Sparck Jonesmodel. We also discuss its possible extension to deal with term dependence.

It should be pointed out that the comparison between the present and previous approaches is not intended to provide a simple decision between the two. Rather we hope to shed some light on the whole problem, by looking at it from different points of view.

2.4.1 Possible application of the independent logistic model

We assume a relevance feedback situation : that is, we assume an initial search which has revealed some relevant documents. We want to use this information to estimate the w's (by maximising expression (3) or (4)), so as to derive an improved ranking of the documents on the next run. The next run may be on the same collection (as in retrospective searching) or a new one (as in SDI).

The result of the initial search is a small number (typically) of documents known to be relevant, also a small number (perhaps zero) of documents known to be non-relevant, and the large bulk of the collection of unknown relevance status. Perhaps the obvious way to use this data in a probablistic model would be to use only the documents of known relevance status in each cell in the derivation of estimates. However, it has been common in earlier systems to use the known relevant as such and all the remainder (known-non-relevant and unknown) as "non-relevant". This is known as the "complement" procedure (Harper and Van Rijsbergen, 1978). The justification for this procedure is that almost all the unknown are certainly non-relevant; the resulting increase in precision of the estimates is likely to far outweigh the slight bias resulting. (Some experiments by Harper and Van Rijsbergen, using another formula, support this view).

We assume, then, that the complement procedure is adopted for the independent logistic model; this has consequences which are discussed below.

In that case, the quantities we need in order to maximise expression (3) or (4) are :

(a) $\sum_{T:t\varepsilon T} k_T$ i.e. the total number of known relevant documents assigned to each term t.

(b) n_T, i.e. the total number of documents, known or unknown, in each individual cell T.

As indicated above, the estimates of the $w's$ would be obtained by means of an iterative maximisation algorithm. The particular procedure used in our experiments is described in section 3.3.

2.4.2 Main points of comparison

We may indicate the main differences betweeen the independent logistic model and the Robertson/Sparck Jones model :

(a) The new model requires a litte more data, namely the individual n_T values (as opposed to the total for each term).

(b) The new model uses an iterative algorithm for obtaining the estimates, rather than a simple formula. The effect of this will depend very much on the number of query terms $\| Q \|$, since the procedure requires manipulation of a $\| Q \| \times \| Q \|$ matrix. This it may affect computation time insignificantly for a 4-term query, but require a totally unrealistic amount of time for a 40-term SDI profile. The methods used and the resources required in the present experiments are discussed in section 3.

(c) The independence assumptions of the new model are considerably weaker than those of the Robertson/Sparck Jones model. The definition of $\mathcal{M}$ given by expression (2) is equivalent to assuming, in some sense, that the degree of dependence of any set of terms is the same in the non-relevant set as it is in the relevant set (Robertson/Sparck Jones assume no dependence in either set).

(d) One major advantage claimed for the logistic models in the medical context is that the estimation process does not involve assuming that the sample on which the estimates are based is random : it is only necessary to assume that the items in the sample from a given cell are a random sample of items in that cell. This property should confer a considerable advantage on the logistic models, since the estimation sample is drawn on the basis of cell membership. However, it appears that this advantage is nullified by the use of the Complement method for non-relevant documents (for which there is no equivalent in the medical application).

(e) The major reason for the choice of the maximum likelihood method of estimating parameters is its mathematical tractability. Maximum likelihood estimates do have a number of desirable properties, but these properties tend to be asymptotic : that is they only hold absolutely for large samples. The estimation method used in the Robertson/Sparck Jones model (i.e. the "*0.5*" formula, expression (1) in section 1.2) was chosen on the basis of a specific property of unbiasedness; this is also an asymptotic property, but the small-sample error is not too great, since it varies with n^{-2} rather than n^{-1} for sample size n.

(f) The Bayesian prior is a very much more flexible device than the equivalent in the Robertson/Sparck Jones method, namely the "+ 0.5" in the formula for calculating weights. Different choices of the means μ_t reflect different assumptions about possible prior indications of term value; different choices of the variance σ^2 reflect different assumptions about the reliance to be placed on these prior indications. Some experiments with different values are discussed in section 3.

(g) Finally, the model itself is more flexible. Thus we can consider the addition of a new term, or of a term-pair (to allow for term-dependence not conforming to the independence assumptions) by simply adding an appropriate dimension to $\mathcal{M}$. (Adding a term-pair as a new component of $\mathcal{M}$ does not conflict with the independence assumptions discussed in (c) above, since $A \;\&\; B \Longrightarrow A$ applies to both the relevant and the non-relevant set).

Thus we see that the logistic model, or class of models, differs from previous approaches in a number of interacting theoretical and practical ways. Section 3 is devoted to a series of experiments with various versions of the logistic model on two test collections.