Retrieval Ordering of Documents Using the Entropy Distribution by IPF Algorithm

Jung Jin Lee*
Soongsil University, Seoul, Korea – jjlee@ssu.ac.kr

Hyun Jo You
Soongsil University, Seoul, Korea – youhyunjo@daum.net

Lee and Kantor [JASIS, 1991, 1998] developed a formulation of the Maximum Entropy Principle (MEP) for Information Retrieval to estimate the effectiveness of term combinations to a specific query. The strong MEP asserts that the actual distribution of relevant and non-relevant documents across the collection of documents corresponds to the distribution generated by the MEP. The Rank MEP orders the possible term combinations in the most effective order for optimal retrieval. However, the MEP formulation is a nonlinear optimization problem which is not easy to solve and the work has had almost no impact. This paper proposes an iterative proportional fitting(IPF) algorithm to solve the MEP formulation constrained on user's judgments. Application of this model to OHSUMED database shows some promising results if expert judgments are close to actual relevance.

Keywords: Information Retrieval, Maximum Entropy Distribution, Iterative Proportional Fitting Algorithm

1. Introduction

The maximum entropy principle (MEP) based on Shannon's measure has been used with great success in many areas. Cooper and Huizinga (1982) applied the MEP to the design of probabilistic information retrieval systems. Specifically, we consider a collection of documents which are categorized into Boolean components by attributes. The MEP estimates the probability of "relevance" of each Boolean component by integrating expert judgments about the "relevance" of attributes with the observed distribution of the Boolean components. The MEP retrieval system, in response to a user's request, provides an ordering of the Boolean components using this estimated probability of "relevance."

Several refinements of the MEP retrieval system design have been developed by Kantor (1982, 1984), Kantor and Lee (1986) using what is called the dual problem to the original constrained optimization. Also, Lee and Kantor (1991) resolved problems of the inconsistent expert judgment in an MEP retrieval system. However, the MEP formulation is a nonlinear optimization problem which is not easy to solve and the work has had almost no impact. This paper proposes an iterative proportional fitting (IPF) algorithm to solve the MEP formulation constrained on user's judgments. Application of this model to OHSUMED database shows some promising results if expert judgments are close to actual relevance.

2. Maximum Entropy Retrieval Model

We shall use the term "documents" to refer in general to any retrievable report or item that might contain further information relevant to the problem at hand. In the simplest example situation the entire set of "documents" may have neither, either, or both of two index terms *A* and *B*. We represent the four possible Boolean components of the entire set of "documents," *R*, using the notations of set operations as follows.

$$\mathrm{R} = \bar{A}\,\bar{B} \cup \bar{A}\mathrm{B} \cup \mathrm{A}\bar{B} \cup \mathrm{AB}$$

Let f_i, $i = 1,2,3,4$, be the fraction of all documents lying in each Boolean component. Suppose that the value of a document is either 0 or 1 which represents "not relevant" or "relevant" respectively. Let p_{iv} denote the probability of document value *v* in the Boolean component *i* where *i*

= *1,2,3,4* and *v = 0,1*. The situation may be described by [Table 1].

[Table 1] Notation used for four Boolean components

Component No	Boolean Component	Probability of non-relevant	Probability of relevant	Fraction of component
1	$\bar{A}\ \bar{B}$	p_{10}	p_{11}	$f_1 = p_{10} + p_{11}$
2	$\bar{A}$B	p_{20}	p_{21}	$f_2 = p_{20} + p_{21}$
3	A$\bar{B}$	p_{30}	p_{31}	$f_3 = p_{30} + p_{31}$
4	AB	p_{40}	p_{41}	$f_1 = p_{40} + p_{41}$

Since p_{iv}'s are the joint probabilities, We have the following probability constraints

$$\sum_{i=1}^{4} \sum_{v=0}^{1} p_{iv} = 1,$$
$$p_{iv} \geq 0 \ i = 1,2,3,4 \text{ and } v = 0,1 \ (1)$$

We want to provide an ordering of the Boolean components, in response to a user's request, by estimating the conditional probability of "relevance" of the documents in each component. Note that the conditional probability that, for example, an item in Boolean component 3 be relevant is $p(3,1)/f_3$. Since the fraction f_i lying in each Boolean component can usually be determined using a computer, the question is how to estimate the joint probability p_{iv}, the "relevance decomposition" of the documents in each component. If we know all the p_{iv}'s, then the Boolean components can be ranked by the order of the conditional probability $p(i,1)/f_i$. The component which has the highest conditional probability will be the first candidate for "relevant" information retrieval.

It is impossible for an expert (or an expert system) to estimate all the p_{iv}'s. However, an expert might be able to provide an opinion in the form "the chance that documents indexed by the term *A* (or *B*) are relevant is V_A (or V_B). The expert estimates of V_A and V_B provide partial information on the data structure and can be used to estimate the p_{iv}'s using the MEP. The resulting constrained optimization problem is to maximize the entropy function of the probabilities p_{iv} subject to V_A and V_B. Note that V_A and V_B can be represented using p_{iv} and f_i as follows

$$V_A = \frac{p_{31}+p_{41}}{f_3+f_4}$$

$$V_B = \frac{p_{21}+p_{41}}{f_2+f_4}$$

Hence the MEP optimization problem to estimate the p_{iv}'s can be formulated as follows.

Find p(i,v), i = 1,2,3,4 and v = 0,1 which

Maximize $-\sum_{i=1}^{4} \sum_{v=0}^{1} p_{iv} \ln p_{iv}$
Subject to

$$p_{31} + p_{41} = V_A (f_3 + f_4)$$
$$p_{21} + p_{41} = V_B (f_2 + f_4)$$
$$\sum_{i=1}^{4} \sum_{v=0}^{1} p_{iv} = 1,$$
$$p_{iv} \geq 0, \ i = 1,2,3,4 \text{ and } v = 0,1$$

This is a nonlinear programming problem which has linear constraints. If we have the solution of the optimization problem, then the Boolean component which has the highest value of the p_{i1} / f_i, i = 2,3,4, will be the best candidate for "relevant" information retrieval.

If there are *k* index terms, then there are 2^k Boolean components. The MEP optimization problem

is generalized with *k* index terms as follows:

Find p(i,v), i = 1,2, ... ,2^k and v = 0,1 which

Maximize $-\sum_{i=1}^{2^k} \sum_{v=0}^{1} p_{iv} \ln p_{iv}$

Subject to

$$\sum_{x \in S_m} p_{x1} = V_m \sum_{x \in S_m} f_x \text{ , } m=1,2,\ldots, M$$

$$\sum_{i=1}^{2^k} \sum_{v=0}^{1} p(i,v) = 1,$$
$$p_{iv} \geq 0, \; i = 1,2, \ldots, 2^k \text{ and } v = 0,1$$

where V_m, $m = 1,2,\ldots,M$, represents the probability of relevance for index term *m*, *M* is the total number of possible index terms whose relevance can be given by experts, and S_m is the set of the Boolean components which are constrained by the index term *m*.

3. Iterative Proportional Fitting Algorithm

In order to estimate the parameters of the maximum entropy distribution, the nonlinear optimization problem should be solved. However, if the number of index terms is increasing, the total number of parameters of the maximum entropy distribution to be estimated is increasing exponentially and the nonlinear programming problem is not easy to solve. In order to overcome this problem, the iterative proportional fitting (IPF) method with I-projection is used to estimate the parameters of the multinomial distribution. The IPF method was originally proposed by Ireland . Kullback(1968) to estimate the distribution of multidimensional table using the marginal distributions. The IPF method assumes the initial estimation of $\{p_{iv}\}$ is an uniform distribution. The IPF method searches iteratively a distribution which minimizes the cross entropy function given the marginal distributions. Existence and convergence of the IPF method has been studied by Ruschendorf(1995) and Cramer(2000). The modified IPF algorithm for the MEP formulation is as follows.

Assume $p_{iv}^{(0)} = {}^{1}/_{2^{k+1}}$

Repeat the following until converges:

For m=1,2, ... , M

Minimize $-\sum_{i=1}^{2^k} \sum_{v=0}^{1} p_{iv}^{(m)} \ln \frac{p_{iv}^{(m)}}{p_{iv}^{(m-1)}}$

Subject to

$$\sum_{x \in S_m} p_{x1}^{(m)} = V_m \sum_{x \in S_m} f_x$$
$$p_{i0}^{(m)} + p_{i1}^{(m)} = f_i, \; i = 1,2, \ldots, 2^k$$
$$p_{iv} \geq 0, \; i = 1,2, \ldots, 2^k \text{ and } v = 0,1$$

It can be shown that the above optimization problem has an explicit solution by using the Lagrangian multiplier method as follows.

$$p_{x1}^{m} = V_m \frac{p_{x1}^{(m-1)}}{\sum_{x \in S_m} p_{x1}^{(m-1)}}, \text{ if x } \in S_m$$

$$p_{x1}^{m} = (1 - V_m) \frac{p_{x1}^{(m-1)}}{\sum_{x \text{ not in } S_m} p_{x1}^{(m-1)}}, \text{ if x not in } S_m$$

Therefore the nonlinear optimization problem can be solved easily by iterative proportional fitting.

4. A Simulation Experiment

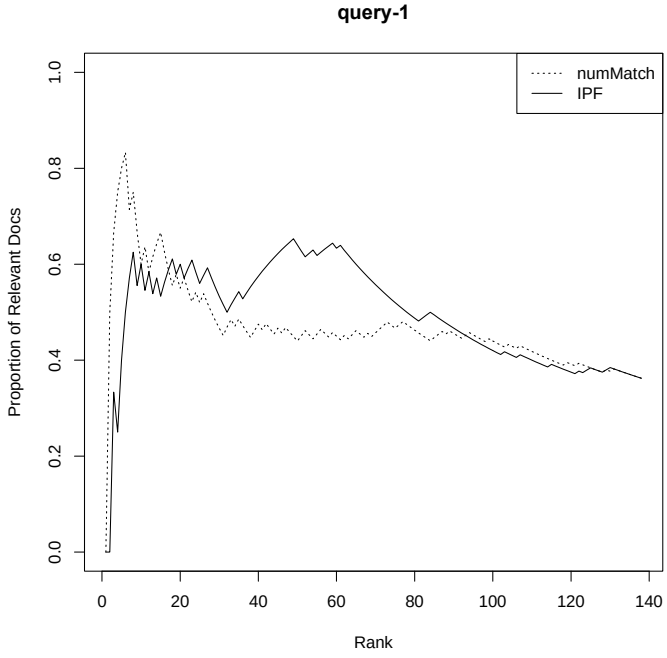
MEDLINE subset in OHSUMED database is used for a simulation experiment. The database includes 348,566 references, 106 queries and 16,140 query-documents pairs have been judged for relevance. For example, query 1 is 'Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?'. After stemming, the most frequent terms are w1=advers, w2=effect, w3=lipid, w4=progesteron, w5=estrogen, w6=replac, w7=therapy and the top 10 documents ranked by the number of matched terms are as follows:

rel	doc-id	w1	w2	w3	w4	w5	w6	w7	numMatch
1	48192	0	1	1	1	1	1	1	6
1	256570	0	1	1	1	1	1	1	6
1	165240	0	1	1	1	1	1	1	6
1	335079	1	1	1	0	1	1	1	6
0	276804	1	1	1	0	1	1	1	6
0	244338	0	1	0	1	1	1	1	5
1	143821	0	1	1	0	1	1	1	5
0	285257	0	1	0	1	1	1	1	5
0	201684	1	1	0	0	1	1	1	5
1	111457	0	1	1	0	1	1	1	5

P(rel doc) given each term estimated by the relevance feedback is as follows.

	w1	w2	w3	w4	w5	w6	w7
P(rel|w)	1/3	6/10	6/7	6/10	6/10	6/10	6/10

<Figure 1> shows the precision of relevant document retrieval using the IPF algorithm with P(rel doc). It shows a high precision after certain number of document retrieval which is a promising result. Similar experiments are currently under study for all other queries.



<Figure 1> Precision using the relevance feedback of query 1

5. Conclusions

If expert judgments and/or pilot judgments are close to actual relevance, MEP retrieval ordering shows some promising results. However, it requires more extensive study using real data. Effect of various types of expert judgment will be studied.

References

Cooper, W.S. and Huizinga, P. (1982). The maximum entropy principle and its application to the design of probabilistic retrieval systems. Information Technology: Research and Development, 1, 99-112.

Cramer, E. (2000). Probability Measures with Given Marginals and Conditionals: I Projections and Conditional Iterative Proportional Fitting, Statistics and Decisions, Vol 18, 311-329.

Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals, Biometrika, Vol 55, 1, 179-188.

Kantor, P.B. and Lee, J.J. (1998). Testing the Maximum Entropy Principle for Information Retrieval. Journal of American Society for Information Science, Vol 49, 6, 557-566.

Lee, J.J. and Kantor, P.B. (1991). A Study of Probabilistic Information Retrieval Systems in the Case of Inconsistent Expert Judgments. Journal of American Society for Information Science, Vol 42, 166-172.

Ruschendorf, L, (1995) Convergence of the Iterative Proportional Fitting Procedure. The Annals of Statistics, 23, No 4, 1160-1174