Automatic Query Expansion Based on Divergence

D. Cai
University of Glasgow
Glasgow G12 8QQ, UK
caid@dcs.gla.ac.uk

C. J. van Rijsbergen
University of Glasgow
Glasgow G12 8QQ, UK
keith@dcs.gla.ac.uk

J. M. Jose
University of Glasgow
Glasgow G12 8QQ, UK
jj@dcs.gla.ac.uk

ABSTRACT

In this paper we are mainly concerned with discussion of a formal model, based on the basic concept of divergence from information theory, for automatic query expansion. The basic principles and ideas on which our study is based are described. A theoretical framework is established, which allows the comparison and evaluation of different term scoring functions for identifying good terms for query expansion. The approaches proposed in this paper have been implemented and evaluated on collections from TREC. Preliminary results show that our approaches are viable and worthy of continued investigation.

Keywords

Automatic Query Expansion, Divergence, Relative Entropy, Term Selection, Scoring Functions, Query Quality

1. INTRODUCTION

One of the major difficulties in textual information retrieval is the description and representation of information needs in terms of a query. Usually, the original query statement consists of just a few terms related to the subject of interest. Given such a scenario a retrieval system cannot be expected to accurately distinguish between relevant and irrelevant documents, which in general are long and often have complex structures.

Automatic query expansion and relevance feedback techniques have been proposed to address this issue. These techniques supplement the original query with additional good terms, and they can be expected to produce an improvement in retrieval performance. In relevance feedback, the expansion terms come from the user-identified relevant documents. In pseudo-relevance feedback systems, the expansion terms come from the top retrieved documents which are assumed to be relevant. An extensive bibliography of papers on the approaches to accomplish these techniques was reviewed by Efthimiadis [1]. One of the approaches is to use analysis of term distributions. The underlying assumption of such an analysis is that the diversity between the different sets of documents might be associated with certain semantic relationships between terms. A theoretical argument that supported the assumption by using of the difference of term distributions to select and re-weight the expansion terms was presented by Robertson [2].

Based on the term distribution analysis, Carpineto et al. [3] [4] proposed a novel pseudo-relevance feedback approach using the concept of divergence studied in information theory. A basic condition that must be satisfied in applications of divergence is that the two components of divergence must be absolutely continuous with respect to one another [5], the divergence does not converge or, is meaningless, otherwise. Usually, the condition may not be satisfied when we attempt to derive the probability distributions from the different sets of documents for the purpose of query expansion. It is therefore a key issue that, for the rationality of applying the concept of divergence to feedback technique, needs to be carefully analysed and rigorously proven. Carpineto, et al. thoroughly discussed this issue, and suggested a scheme that attempted to find out a discounting factor μ $(0 < \mu < 1)$ for discounting the probability distribution of terms in order to solve the problem. In their work, however, it seemed that the factor μ was not really derived, and the main experiments described still relied on $\mu = 1$. In fact, the theoretical problem of applying divergence to query expansion technique still remains an open problem, and is one of the focal points of this study.

This leads to the research questions we address in this paper. We aim to develop a formal model that resolves some of the difficulties in existing approaches that apply information theory ideas to information retrieval. The developed model does not require any a priori knowledge about relevance information in a sample set of documents, is computationally simple, can be easily implemented, and is effective in improving retrieval performance.

In the remainder of this paper, we proceed as follows. First, in section 2, we discuss the representing scheme for documents and queries. We then describe a formal approach based on the concept of divergence for constructing a scoring function for query expansion, and illustrate how it can be applied in a realistic retrieval system. Followed to that, in section 4, we evaluate the retrieval effectiveness of our proposed approach and expound on how the effectiveness varies according to the quality of original query. In section 5, we describe an alternative formal approach based on the concept of relative entropy for query expansion. Finally, we comment on the necessity of this study, and conclude our discussions by emphasising some of the difficulties encountered in the design of a query expansion retrieval system.

CIKM'01, November 5-10, 2001, Atlanta, Georgia, USA.
Copyright 2001 ACM l-581 13-436-3/01/001 1...$5.00.

2. KNOWLEDGE REPRESENTATION

In a textual information retrieval system, thc *objects* one deals with are documents and *queries.* In order to develop an appropriate quantitative retrieval model, one has to design a reasonable scheme to represent documents and query.

In information retrieval, each object x is represented by means of a set of concepts, in which, the semantics involved is a key issue. So far, the simplest way of characterising each concept involved in a object is to use index terms that appear in the object.

Usually, an individual index term may contain a piece of information, and there exists complex semantic relationships between the index terms. Each object may therefore be composed of an arbitrary amount of information. No assumptions are made regarding the structure of the objects, although in practice structured subdivisions of the documents can be accommodated.

To arrive at a precise representation of entire information content of an object by means of a set of index terms is arduous task because it is very difficult to obtain sufficient statistical data for the estimation of the amount of information contained in index terms, and for the indication of the semantic relationships between index terms.

Let $D = \{d_1, d_2, \cdots, d_i, \cdots, d_N\}$ be a document collection, and let a finite ordered tuple $V = \{t_1, t_2, \ldots, t_j, \ldots, t,\}$ be the vocabulary of terms indexed from the whole collection D. Let q be a query. Assume that the distribution of the amount of information of terms $t_j \in V$ (called information distribution) in each object x = $d_i \in D$ or x = q can be approximately quantitatively represented, with a *weighted vector* $v_x = (w_x(t_1), w_x(t_2), \cdots, w_x(t_j), \cdots, w_x(t_n))$. With such a knowledge representation, the relationships between the objects will become transparent when dealt with a specific quantitative retrieval model.

Under an idealised retrieval environment, the knowledge representation will be perfect and the interpretation of *weights* of terms should be entirely independent of any individual model. However, a feasible scheme for computing accurately the amount of information contained terms and also capturing effectively the semantic relationships between terms is not available. Thus interpretation of the weights has to depend on a specific model itself, and it is frequently consistent with the statistical nature of indexing procedure. Generally, the weights $w_x(t_j)$ are considered to 'indicate' the relative importance of the term $t_j \in V$ for the object x. The terms with higher weights are regarded to 'contain' more information than those with lower weights.

3. A MODEL FOR TERM SELECTION

This section focuses on the development of a formal model for automatic query expansion. The approach proposed is based on the concept of divergence studied in information theory. We will prove that the methodology discussed in this study is theoretically justified.

3.1 Background

In probabilistic retrieval, the random *event* 'which term will occur' can be viewed as a random variable on V, denoted by ξ, which may be a mapping $\xi : V \to \{1, \cdots, n\}$ if we suppose that $\xi(t_j) = j$. The random variable $\xi(t)$ have some kind of uncertainty, which can be characterised by means of a ***probability distribution*** $P_\pi(\{\xi(t_j) = j\})$ $(j = 1, \cdots, n)$, denoted by $P_\pi(t)$ for convenience, over ***probability space*** $(V, 2")$. The interpretation of $P_\pi(t)$ will depend on the population π from which it is derived and on the statistical nature associated with π in a specific indexing procedure.

Let populations $\pi_k = D_k$ C D $(k = 1, 2)$ be two sets of documents, and $P_{D_k}(t)$ probability distributions over $(V, 2^V)$, derived from D_k $(k = 1, 2)$. Here $P_{D_k}(t)$ are interpreted as information distributions of D_k $(k = 1, 2)$.

The divergence between the distributions $P_{D_1}(t)$ and $P_{D_2}(t)$ due to Kullback & Leibler [5] is defined by

$$J(P_{D_1}(t), P_{D_2}(t)) = \sum_{t \in V} (P_{D_1}(t) - P_{D_2}(t)) \log \frac{P_{D_1}(t)}{P_{D_2}(t)},$$

which can be used to measure the average *difference* of the information contained in $P_{D_1}(t)$ and that contained in $P_{D_2}(t)$ about $P_{D_1}(t)$, and vice versa.

In order to avoid meaningless expressions in the discussion that follows, we use the following notational conventions:

$$0 \cdot \log(\tfrac{0}{a}) = 0, \qquad 0 . \log(\tfrac{0}{0}) = 0$$

$$(0 - a) . \log(\tfrac{0}{a}) = \lim_{\varepsilon \to +0}(\varepsilon - a) . \log(\$) = +\infty,$$

$$(a - 0) . \log(\tfrac{a}{0}) = \lim_{\varepsilon \to +0}(a - \varepsilon) . \log(\tfrac{a}{\varepsilon}) = +\infty.$$

where $0 < a < +\infty$. That is, for some $t' \in V$, if $P_{D_1}(t') = 0$ (but $P_{D_2}(t') \neq 0$), or if $P_{D_2}(t') = 0$ (but $P_{D_1}(t') \neq 0$), the conventions that $(0 \quad P_{D_2}(t')) \log \frac{0}{P_{D_2}(t')} = +\infty$ and $(P_{D_1}(t') \quad 0) \log \frac{P_{D_1}(t')}{0} = +\infty$ are accepted.

It can be easily verified that the divergence has the following properties: $J(P_{D_1}, P_{D_2}) \geq 0$ with equality if and only if $P_{D_1}(t) = P_{D_2}(t)$ for all $t \in V$, and $J(P_{D_1}, P_{D_2}) = J(P_{D_2}, P_{D_1})$. This is, the divergence is non-negative and symmetric.

In practice, it might be sometimes desired to have a consistency in the measure of the average difference between distributions. The divergence $J(\cdot, \cdot)$ is explored so as to product a symmetric *dissimilarity* measure between two distributions when we have no particular reason to emphasise either of them. Thus, it should be more natural and reasonable for us to think of the divergence $J(\cdot, \cdot)$ as a 'distance' measure (even it is not) between distributions in a realistic retrieval environment.

3.2 Some Explanations and Notation

The divergence is a basic concept in information theory. It has different applications in a variety of research areas, in particular, it has been becoming a useful tool in estimation problems. We believe that this is also certainly true in the statistical and probabilistic information retrieval.

Perhaps the usefulness of divergence can be best illustrated by the following situation. In an information retrieval context, it is desirable or necessary to consider the divergence between the information distributions derived from some specific populations (sets of documents) can expose some semantic relations between terms. A feasible scheme of capturing true semantic relations of complicated semantics is not yet available. But if the divergence in the form of population distributions can be obtained, and if the distributions can approximately reflect information distributions in the populations, then one will know for sure that the divergence meets one's needs.

Underlying all of our discussions in this study is the assumption that the divergence between information distributions derived from the different sets of documents can reveal some semantic relations between terms.

Let the population π = ***R*** be the set of relevant documents with respect to original query ***q***, and V^R be the sub-vocabulary consisted of the terms that appear in the relevant documents, which constitute a source of candidate terms. The set of candidate terms contains terms that may be added to the query, whereas the set of *expansion* terms contains terms that the system actually adds to the query. Let $V^{e(q)}$ be the set of expansion terms of the query ***q*** (then $V^{e(q)}$ C V^R). Let V^x be the set of terms that appear in the object x.

We denote $|\pi|$ as the size of population π (e.g., $|D|$ = N when $\pi = D$), $|V^\pi|$ and $|V^x|$ as the sizes of sub-vocabularies V^π and ***V”,*** respectively. And we denote $||x||$ as the ***length*** of object x, $||v_x|| = (\sum_{t \in V} w_x^2(t))^{\frac{1}{2}}$ as the norm (or length) of vector v_x representing object x. Also we denote $f_x(t)$ as the occurrence *frequency* of the term t in the object x, $F_D(t)$ ***as*** the ***document*** frequency of term t relative to the collection ***D,*** and ***awe(D)*** = $\frac{1}{|D|} \sum_{d \in D} ||d||$ is the average length of documents in ***D.***

Note that, in this study we denote $t \in V^x$ instead oft $\in x$ (e.g., $t \in V^d$ rather than $t \in$ ***d)*** because from the ***set theory*** viewpoint the duplicate terms in the set of terms should be represented with one element. Thus we have, for instance, $||d|| \geq |V^d|$.

Many empirical results show that the difference of distributions of terms in the relevant and non-relevant document sets might reflect some semantic relations between terms. One would expect, the terms related to the query ***q*** will occur frequently in relevant documents to ***q,*** rather than in non-relevant ones. A reasonable function to score candidate terms for query expansion should be the one that can reflect semantic relations between the candidate terms $t \in V^R$ and the terms $t \in V^q$.

Suppose that the selection of the expansion terms $t \in V^{e(q)}$ are based on the scores, obtained by means of the divergence $J(\cdot, \cdot)$, of candidate terms $t \in V^R$. The basic idea of the selection is that the candidate terms assigned high scores are considered as being likely to contain information relevant to the query, and that the terms relevant to the query should be much more concentrated on the sub-vocabulary V^R than the whole vocabulary ***V.***

In a practical retrieval environment, the relevant document set ***R*** is frequently replaced with the pseudo-relevant document set $\mathfrak{R}$ for the purpose of the query expansion. That is, we will generate $V^{e(q)}$ from $V^{\mathfrak{R}}$, the sub-vocabulary consisted of the terms that appear in the pseudo-relevant documents in $\mathfrak{R}$, instead of from V^R. In what follows, we attempt to theoretically propose a approach of constructing a scoring function by using divergence $J(P_{\mathfrak{R}}, P_D)$, which is assumed to be able to reveal some semantic relations inherent in the candidate terms $t \in V^{\mathfrak{R}}$ and the terms $t \in V^q$.

3.3 A Scoring Function

We are now in a position to investigate the query expansion by applying the divergence $J(P_{\mathfrak{R}}, P_D)$.

Let $P_{\mathfrak{R}}(t)$ be a probability distribution on ***V*** derived from the pseudo-relevant document set $\mathfrak{R}$, and $P_{\mathfrak{R}}$ ***(t) > 0*** for every $t \in V^{\mathfrak{R}}$ and $P_{\mathfrak{R}}(t) = 0$ for every $t \in V - V^{\mathfrak{R}}$. Let $P_D(t)$ be a probability distribution on ***V*** derived from the whole collection ***D,*** and $P_D(t)$ ***> 0*** for every $t \in$ ***V”.***

Obviously, if $|\mathfrak{R}| = |D|$ (in this case, $P_{\mathfrak{R}}(t) > 0$ and $P_D(t)$ ***> 0*** for every $t \in V^{\mathfrak{R}}$ = ***V)*** the divergence $J(P_{\mathfrak{R}} : P_D)$ is entirely meaningful for the distributions $P_{\mathfrak{R}}(t)$ and $P_D(t)$ over the same probability space $(V, 2^V) = (V^{\mathfrak{R}}, 2^{V^{\mathfrak{R}}})$. We can therefore directly apply the $J(P_{\mathfrak{R}}, P_D)$ as a dissimilarity measure to construct a score function

$$score_q(t) = (P_{\mathfrak{R}}(t) - P_D(t)) \log \frac{P_{\mathfrak{R}}(t)}{P_D(t)} \qquad (t \in V^{\mathfrak{R}})$$

for query expansion (for the interpretation of this see the last paragraph in section 3.4).

Otherwise, without losing generality, assume that $|\mathfrak{R}| < |D|$. Let us denote the ***item*** of divergence $J(P_{\mathfrak{R}}, P_D)$ ***as a*** function of terms $t \in$ ***V,*** that is,

$$f(t) = (P_{\mathfrak{R}}(t) - P_D(t)) \log \frac{P_{\mathfrak{R}}(t)}{P_D(t)} \qquad (t \in V).$$

Note that $P_{\mathfrak{R}}(t) = 0$ for every $t \in$ ***V*** $- V^{\mathfrak{R}}$, therefore $f(t) = (0 - P_D(t)) \log \frac{0}{P_D(t)} = +\infty$ (when $P_D(t) \neq 0$), and this results in the $J(P_{\mathfrak{R}}, P_D)$ being meaningless. That is, we can not directly apply $J(P_{\mathfrak{R}}, P_D)$ as a dissimilarity measure for $P_{\mathfrak{R}}(t)$ and $P_D(t)$ in a query expansion application.

In order to resolve this problem theoretically, we design ***decomposed probability distributions*** over probability space $(V_{t^*}^{\mathfrak{R}}, 2^{V_{t^*}^{\mathfrak{R}}})$ for both $P_{\mathfrak{R}}(t)$ and $P_D(t)$, where $V_{t^*}^{\mathfrak{R}} = V^{\mathfrak{R}} \cup \{t^*\}$ (let $t^* \notin$ ***V*** be a ***fictitious ‘term’*** without containing any information content). The scheme adopted in this study is simply based on discounting some value of density of $P_{\mathfrak{R}}(t)$, say $P_{\mathfrak{R}}(t_0)$ of the term $t_0 \in V^{\mathfrak{R}}$ (satisfying $P_{\mathfrak{R}}(t_0) \neq 1$), with a discounting factor $\mu = P_{\mathfrak{R}}(t_0)$ (then $0 < \mu < 1$). Whereas the discounted value of density $P_{\mathfrak{R}}(t_0) - \mu P_{\mathfrak{R}}(t_0) = P_{\mathfrak{R}}(t_0) - P_{\mathfrak{R}}^2(t_0)$ is restored by redistributing it onto the fictitious term t^*. We may formulate this scheme by the decomposed probability distribution $P_{\mathfrak{R}}(t,$ ***to,*** $t^*)$ as follows.

$$P_{\mathfrak{R}}(t, \boldsymbol{to}, t^*) = \begin{cases} P_{\mathfrak{R}}(t) & \text{when } t \in V_{t^*}^{\mathfrak{R}} - \boldsymbol{\{to\}} - \{t^*\} \\ P_{\mathfrak{R}}^2(t_0) & \text{when } t = t_0 \in V_{t^*}^{\mathfrak{R}} \\ P_{\mathfrak{R}}(t_0) - P_{\mathfrak{R}}^2(t_0) & \text{when } t = t^* \in V_{t^*}^{\mathfrak{R}}. \end{cases}$$

Similarly, for the distribution $P_D(t)$, the decomposed probability distribution $P_D(t, t_0, t^*)$ can be expressed by

$$P_D(t, t_0, t^*) = \begin{cases} P_D(t) & \text{when } t \in V_{t^*}^{\mathfrak{R}} - \boldsymbol{\{to\}} - \{t^*\} \\ P_{\mathfrak{R}}(t_0) P_D(t_0) & \text{when } t = t_0 \in V_{t^*}^{\mathfrak{R}} \\ P_D(t_0) - P_{\mathfrak{R}}(t_0) P_D(t_0) + \sum_{t \in V - V^{\mathfrak{R}}} P_D(t) & \\ \qquad \textbf{when } t = t^* \in V_{t^*}^{\mathfrak{R}}. & \end{cases}$$

It is easily seen that both $P_{\mathfrak{R}}(t,$ ***to, t*)*** and $P_D(t, t_0, t^*)$ satisfy two axioms of probability distribution, they are hence probability distributions on $V_{t^*}^{\mathfrak{R}}$. In fact, it is readily viewed that $P_{\mathfrak{R}}(t,$ ***to,*** $t^*) > 0$ and $P_D(t, t_0, t^*) > 0$ hold for every $t \in V_{t^*}^{\mathfrak{R}}$.

Thus the divergence $J(P_{\mathfrak{R}}(t), P_D(t))$ is modified to $J(P_{\mathfrak{R}}(t,$ ***to,*** $t^*), P_D(t,$ ***to, t”)),*** which is rewritten as follows.

$$\begin{aligned} & J(P_{\mathfrak{R}}(t, t_0, t^*), P_D(t, t_0, t^*)) \\ & = \textstyle\sum_{t \in V_{t^*}^{\mathfrak{R}}} (P_{\mathfrak{R}}(t, t_0, t^*) \quad P_D(t, t_0, t^*)) \log \frac{P_{\mathfrak{R}}(t, t_0, t^*)}{P_D(t, t_0, t^*)} \\ & = \textstyle\sum_{t \in V^{\mathfrak{R}} - \{t_0\} - \{t^*\}} f(t) + \varepsilon_1(t_0) + \varepsilon_2(t^*), \end{aligned}$$

in which,

$$\varepsilon_1(t_0) = (P^2_{\Re}(t_0) - P_{\Re}(t_0)P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)},$$

$$\varepsilon_2(t^*) = [(P_{\Re}(t_0) - P^2_{\Re}(t_0)) -$$

$$-(P_D(t_0) - P_{\Re}(t_0)P_D(t_0) + \textstyle\sum_{t\in V-V^{\Re}} P_D(t))]\times$$

$$\times \log \frac{P_{\Re}(t_0)-P^2_{\Re}(t_0)}{P_D(t_0)-P_{\Re}(t_0)P_D(t_0)+\sum_{t\in V-V^{\Re}} P_D(t)}.$$

Thus, the $J(P_{\Re}(t, t_0, t^*), P_D(t, t_0, t^*))$ is entirely meaningful for the distributions $P_{\Re}(t, t_0, t^*)$ and $P_D(t, t_0, t^*)$ ***over*** the same probability space $(V^{\Re}_{t^*}, 2^{V^{\Re}_{t^*}})$.

As discussed above, the modified divergence $J(P_{\Re}(t, t_0, t^*), P_D(t, t_0, t^*))$ can be used to measure the difference of the information contained in $P_{\Re}(t, t_0, t^*)$ and that contained in $P_D(t, t_0, t^*)$ about $P_{\Re}(t, to, t^*)$, and vice versa. While the difference is assumed probably to reveal some semantic relations between terms $t \in V^{\Re}_{t^*}$ and terms $t \in V^q$. Thus, the expansion terms selected (from $t \in V^{\Re}_{t^*}$) should be those which ***mostly contribute*** to the difference. The greater difference the terms make, the more likely they are to be related semantically to the query q. These statements may be further formulated by means of a scoring function

$$score_q(t, t_0, t^*) = \begin{cases} f(t) & \text{when } t \in V^{\Re}_{t^*} \quad \{t_0\} - \{t^*\} \\ \text{El (to)} & \text{when } t = t_0 \in V^{\Re}_{t^*} \\ \varepsilon_2(t^*) & \text{when } t = t^* \in V^{\Re}_{t^*}, \end{cases}$$

in conjunction with a decision *rule*: If $score_q(t_j, to, t^*) > score_q(t_k, t_0, t^*)$ then t_j is more relevant to q than t_k.

Thus our task is reduced to calculating the scores using $score_q(t, t_0, t^*)$ for every $t \in V^{\Re}_{t^*}$, and then to decide which of them should be selected as expansion terms for query expansion by means of the decision rule.

Note that, in the discussion above, $t_0 \in V^{\Re}$ is completely arbitrary, and the values of scoring function $score_q(t, to, t^*)$ are entirely independent to *to* when $t \in V^{\Re}_{t^*} - \{to\} - \{t^*\} = V^{\Re} - \{t_0\}$.

3.4 Simplification of the Scoring Function

We have formally derived a scoring function based on the modified divergence as discussed above. We now wish to make a further simplification so that it is reasonable for us to use the scoring function directly for the terms $t \in V^{\Re}$ without considering what the terms t_0 and t^* should be. In other words, we wish to be able to ignore the score values $\varepsilon_1(t_0)$ and $\varepsilon_2(t^*)$ of terms t_0 and t^* at all when we use $score_q(t_j, t_0, t^*)$ as a query expansion approach.

For this purpose, let us look at again the modified divergence and the scoring function as discussed above. We wish that, for every $t \in V^{\Re} - \{to\}$, the item $f(t)$ of divergence, i.e., the score $score_q(t, to, t^*)$ satisfies the inequality

$$f(t) = (P_{\Re}(t) - P_D(t))\log\frac{P_{\Re}(t)}{P_D(t)}$$

$$\geq (P^2_{\Re}(t_0) - P_{\Re}(t_0)P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)} = \varepsilon_1(t_0).$$

This inequality explicitly indicates that the difference between $P_{\Re}(t, t_0, t^*)$ and $P_D(t, t_0, t^*)$ mostly comes from the terms $t \in V^{\Re} - \{to\}$, rather than from the term *to.*

In fact, it is easily seen that the following inequality

$$f(t_0) = (P_{\Re}(t_0) - P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)}$$

$$\geq (P^2_{\Re}(t_0) - P_{\Re}(t_0)P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)} = \varepsilon_1(t_0)$$

holds iff (if and only if)

$$(P_{\Re}(t_0) - P_D(t_0)) - (P^2_{\Re}(t_0) - P_{\Re}(t_0)P_D(t_0))$$

$$= P_{\Re}(t_0)(1 - P_{\Re}(t_0)) - P_D(t_0)(1 - P_{\Re}(t_0))$$

$$= (P_{\Re}(t_0) - P_D(t_0))(1 - P_{\Re}(t_0)) \geq 0$$

holds iff $P_{\Re}(t_0) \geq P_D(t_0)$ holds.

Therefore, taking $t_0 \in V^{\Re}$ such that

$$f(t_0) = (P_{\Re}(t_0) - P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)}$$

$$= \min\{ f(i?) \quad t \in V^{\Re}\},$$

we immediately obtain

$$f(t) = (P_{\Re}(t) - P_D(t))\log\frac{P_{\Re}(t)}{P_D(t)}$$

$$\geq (P_{\Re}(t_0) - P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)} = f(t_0)$$

$$\geq (P^2_{\Re}(t_0) - P_{\Re}(t_0)P_D(t_0))\log\frac{P_{\Re}(t_0)}{P_D(t_0)} = \varepsilon_1(t_0)$$

holds, for every $t \in V^{\Re}$, iff $P_{\Re}(t_0) \geq P_D(t_0)$ holds.

In a real retrieval environment, the set of pseudo-relevant documents is usually very small compared with the whole collection. Thus the size of V (i.e., $|V|$) is much larger than the size of $V^{\Re}$ (i.e., $|V^{\Re}|$). Accordingly, the densities of $P_{\Re}(t)$ are relatively much greater than the densities of $P_D(t)$ for all $t \in V^{\Re}$. Therefore, from the viewpoint of applications, it should not be a problem for satisfying the constraint $P_{\Re}(t_0) \geq P_D(t_0)$ for $t_0 \in V^{\Re}$ at all.

On the other hand, t^* is a fictitious term without containing any information content, it is of course impossible to be relative to any given query. So there is no need to consider the score $\varepsilon_2(t^*)$ for the term t^* at all during query expansion procedure.

Note that the scoring function $score_q(t, t_0, t^*)$ is independent to t_0 and t^* when $t \in V^{\Re} - \{to\}$. From the discussion above, $score_q(t, t_0, t^*)$ can then be actually simplified to

$$score_q(t) = f(t) = (P_{\Re}(t) - P_D(t))\log\frac{P_{\Re}(t)}{P_D(t)} \qquad (t \in V^{\Re}),$$

along with a decision rule: If $score_q(t_j) > score_q(t_k)$ then t_j is more relevant to q than t_k. This turns out to be simple computationally and constructively.

Thus each candidate $(t \in V^{\Re})$ can be assigned a score by the 'intelligent judgementor' ***score,(t).*** The candidate terms with the top scores should be given a top priority as the expansion terms $t \in V^{e(q)}$, and these expansion terms actually make the most contribution to the $J(P_{\Re}, P_D)$, the difference between $P_{\Re}(t)$ and $P_D(t)$, among the terms $t \in$ **V”.** Consequently, as the assumption given in section 3.2, they might be more relevant to the query q than the others or, in other words, might be regarded ***as*** good ***discriminators*** to distinguish relevant documents in the whole collection from non-relevant ones with respect to the query.

3.5 Application of the Scoring Function

An appropriate question is 'How should I set components of $J(P_{\Re}, P_D)$ for applying the scoring function ***score,(t)*** to effectively improve retrieval performance?'. This is very tricky problem indeed. In any probabilistic retrieval model, the information distribution of a certain population will have to be approximately estimated by means of the estimation of term *distribution* based on the statistics derived from the population. The population may be a small set of sample documents, or be any set of documents, whatever the population may be, depends on a specific application. Consequently, the estimation of term distribution plays an essential part in determining retrieval effectiveness for any probabilistic retrieval model. In fact, the accuracy and validity to approximately estimate information distribution by the estimation of term distribution are for a long time a crucial and open problem. This is because, as mentioned in section 2, it is very difficult to obtain sufficient statistics for the estimation of the amount of information contained in terms, and for the indication of the semantic relations between terms. So far, almost all existing probabilistic models might still suffer from the same problem. This study does not give estimation rules for the distributions $P_{\Re}(t)$ and $P_D(t)$, which will be regarded as a significant subject for further study.

However, we have made some preliminary attempts and studies of estimating components and testing results. One of them is illustrated here, and its experimental result will be given in the following section.

In practice, we have only observations, i.e., the statistics of the occurrence frequency of terms, from the different sets of documents $\Re$ and *D*. In this case, for each term $t \in V^d$ and $d \in \Re$, let

$$agr_d(t, q) = p_d(t) \times p_q(d) = \frac{f_d(t)}{||d||} \times \frac{sim(d,q)}{\sum_{d\in\Re} sim(d,q)},$$

indicates the agreement of term $t \in V^d$ with the query *q*. The probability *p,(d)* is relative similarity defined over $\Re$ and derived based on the similarity measure $sim(d, q)$ obtained in initial retrieval procedure. The choice of $sim(d, q)$ depends on a specific model itself. Therefore, when $t \in V^{\Re}$, the component $P_{\Re}(t)$ can be approximately estimated by

$$P_{\Re}(t) = \frac{\sum_{d\in\Re} agr_d(t,q)}{\sum_{t\in V^{\Re}}(\sum_{d\in\Re} agr_d(t,q))}$$
$$= \frac{\sum_{d\in\Re} p_d(t)p_q(d)}{\sum_{t\in V^{\Re}}(\sum_{d\in\Re} p_d(t)p_q(d))} = \sum_{d\in\Re} p_d(t)p_q(d),$$

since $\sum_{t\in V^d} p_d(t) = 1$, $\sum_{d\in\Re} p_q(d) = 1$, and $p_d(t) = 0$ when $t \in V^{\Re} - V^d$, hence, for denominator in the expression of $P_{\Re}(t)$, *we* have

$$\sum_{t\in V^{\Re}}(\sum_{d\in\Re} p_d(t)p_q(d)) = \sum_{d\in\Re}(\sum_{t\in V^{\Re}} p_d(t)p_q(d))$$
$$= \sum_{d\in\Re}(p_q(d) \sum_{t\in V^d} p_d(t)) = 1.$$

It is easily verified that $P_{\Re}(t) > 0$ for every $t \in V$”.

Similarly, for each term $t \in V^d$ and $d \in D$, let

$$imp_D(t,d) = p_d(t) \times idf_D(t) = \frac{f_d(t)}{||d||} \times log\frac{|D|}{F_D(t)},$$

indicates the importance of term $t \in V^d$ to document *d* (relative to *D)*. Therefore, when $t \in V$, the component $P_D(t)$ can be approximately estimated by

$$P_D(t) = \frac{\sum_{d\in D} imp_D(t,d)}{\sum_{t\in V}(\sum_{d\in D} imp_D(t,d))} = \frac{\sum_{d\in D} p_d(t) idf_D(t)}{\sum_{t\in V}(\sum_{d\in D} p_d(t) idf_D(t))}.$$

It is easily verified that $P_D(t) > 0$ for every $t \in V$ $(\supseteq V^{\Re})$.

More discussions on the estimation of components and their corresponding experimental results will be given in the next section.

Finally, we would like to point out that in this formal model the postulates of $P_D(t) > 0$ and $P_{\Re}(t) > 0$ for every $t \in V^{\Re} \subseteq V$ are not excessive. They are necessary and the least of conditions for applying the divergence $J(P_{\Re}, P_D)$ to construct a scoring function in this model. Because the vocabulary *V* is a finite tuple, these postulates are not infeasible and are practical in a realistic information retrieval context.

4. PRELIMINARY EXPERIMENTS

The main purpose of this section is to evaluate the retrieval effectiveness of the strategy proposed in this study. We compared the overall retrieval performance of query expansion using our approach, with that of the original query without query expansion, and with that of query expansion using reduced Rocchio's formula.

4.1 Description and Setting

In order to investigate to what extent our approach of query expansion constitutes an improvement in retrieval performance, we have carried out a number of experiments. Our experiments use two collections from the TREC ad hoc data: AP90 (Associated Press newswire, 1990) with 78,321 documents, and FT (the Financial Times) with 210,158 documents. Regarding the queries for the experiments, we use two sets of queries, which were automatically derived from two sets of 50 natural language *topics* (201-250 and 351-400) without any manual intervention. The average length of the topics in the two sets are 16.3 terms and 14.3 terms, respectively.

The classical approach of query expansion due to Rocchio's formula [6] has been shown to achieve good retrieval performance [7]. The reduced formula, which can be used both to rank candidate terms and to re-weight the expansion terms for obtaining *expanded* query *q'*,

$$v_{q'} = \lambda_1 v_q + \frac{\lambda_2}{|\Re|} \sum_{d\in\Re} \frac{v_d}{||v_d||}$$

is employed in our experiments with $\lambda_1 = \lambda_2 = 1$. The strategy of query expansion based on the reduced formula will be adopted as one of the baselines in our study.

In the initial retrieval, we use Okapi's weighting scheme due to Robertson, el at. in [8], which has also been shown to produce good initial retrieval performance. That is, for every $t \in V$ the weights for the document *d* and the query *q* are approximately represented, respectively, by

$$w_d(t) = \begin{cases} \frac{2.2 f_d(t)}{f_d(t)+1.2(0.25+0.75||d||/ave(D))} & \text{when } t \in V^d \\ 0 & \text{when } t \in V - V^d \end{cases}$$

$$w_q(t) = \begin{cases} \frac{1001 f_q(t)}{1000+f_q(t)} \times log\frac{|D|-F_D(t)+0.5}{F_D(t)+0.5} & \text{when } t \in V^q \\ 0 & \text{when } t \in V - V^q. \end{cases}$$

And we simply use the *inner product* between the document vector v_d and the query vector v_q as a *decision* function to compute the similarity between *d* and *q*, that is,

$$sim(d, q) = v_d \cdot v_q = \sum_{t\in V^d \cap V^q} w_d(t) \times w_q(t).$$

Based on the initial retrieval, for each original query *q*, the set $\Re$ of top α pseudo-relevant documents, and the corresponding $V^{\Re}$ can be derived. Then our scoring function score,(t) is applied to score and rank candidate terms $t \in V^{\Re}$, and pick out the top β terms for obtaining expanded query q' from q. The re-weighting scheme, using reduced formula above, is performed over the terms $t \in V^{q'} = V^{e(q)} \cup V^{q}$, and the weights of terms $t \in V^{e(q)} \cap V^{q}$ are probably modified. The whole collection then goes through the second retrieval iteration with respect to the expanded query q', and documents are re-ranked using the similarity $sim(d, q')$. Finally, the results are presented to the user.

All parameters, such as α = 10 and β = 30, are set before the query is submitted to the system, and before going through retrieval

The approach proposed in this study to select the expansion terms is not complex, the time necessary for performing solely the query expansion is negligible, and the time required for the second retrieval iteration is comparable to standard query expansion approaches, e.g., [2].

4.2 Components and Estimates

As discussed in the last section, estimating the components of $score_q(t)$ is crucial for effectively distinguishing the potentially relevant terms $t \in V^{\Re}$ to the query q from non-relevant ones using our approach. We now show some more examples of the approximate estimates of components, which generate the 'different' scoring functions. The experiments based on these estimates, which are discussed below, have been designed and performed for the comparison of retrieval performance using these scoring functions to obtain expanded query.

Table 1 gives the statistics in terms of standard evaluation measures when 6 scoring functions with different schemes (a - f) of component estimates below are applied to collection AP90 for queries 201-250. The statistics that apply to the original queries can be found in the 2nd column (Ori), and columns a f pertain to the statistics of the different schemes (a - f) for query expansion.

Table 1: Comparison of performance for difl :rent schemes

	Ori	a	b	c	d	e	f
RET	628	696	697	707	702	702	698
at-0.00	.6518	.6395	.6379	.6514	.6392	.6392	.6385
at-0.10	.5288	.5710	.5687	.5797	.5646	.5646	.5627
A-P	.2794	.3241	.3240	.3214	.3191	.3191	.3157
pre-at-5	.3917	.4042	.4042	.4042	.4083	.4083	.4125
pre-at-10	.3396	.3750	.3729	.3667	.3729	.3729	.3688
R-P	.3278	.3663	.3628	.3618	.3678	.3678	.3609

A-P: Average-Precision, R-P: R-Precision.

In which,

a) $agr_d(t,q) = p_d(t) \times p_q(d)$, $imp_D(t,d) = p_d(t) \times idf_D(t)$;

b) $agr_d(t,q) = p_d(t) \times p_q(d)$, $imp_D(t,d) = f_d(t) \times idf_D(t)$;

c) $agr_d(t,q) = p_d(t) \times p_q(d)$, $imp_D(t,d) = f_d(t)$;

d) $agr_d(t,q) = f_d(t) \times P,(d)$, $imp_D(t,d) = f_d(t)$;

e) $agr_d(t,q) = f_d(t) \times sim(v_d, v_q)$, $imp_D(t,d) = f_d(t)$;

f) $agr_d(t, q) = f_d(t)$, $imp_D(t, d) = f_d(t)$;

are used to approximately estimate components

$$P_{\Re}(t) = \frac{\sum_{d \in \Re} agr_d(t,q)}{\sum_{t \in V^{\Re}} (\sum_{d \in \Re} agr_d(t,q))},$$

$$P_D(t) = \frac{\sum_{d \in D} imp_D(t,d)}{\sum_{t \in V} (\sum_{d \in D} imp_D(t,d))}.$$

The results, obtained in Table 1, for retrieval using the different schemes are consistently better than the one for retrieval using the original query. The level of improvement (based on all 7 standard evaluation measures) is substantial at most evaluation points for all of schemes listed above. It seems that scheme c give better performance, but the difference between applying these schemes is inconsiderable.

4.3 Results and Analysis

Another aspect of experiments is the comparison of performance of retrieval using our approach of query expansion, with retrieval using the original query, and with retrieval using the reduced Rocchio's formula of query expansion. The corresponding testing statistics given in the Table 2.

The results, in Table 2, give us a view when the different strategies are applied to collection AP90 for queries 201-250. The statistics that apply to the original query can be found in the 2nd column (Ori). The 3th Column pertains to apply reduced Rocchio's formula (Roc), and 4th column (c) pertains to apply scheme c.

Table 2: Comparison of performance for different strategies

	Ori	Roc	c	Ori	Roc	c
RET	628	675	707	569	614	621
at-0.00	.6518	.6004	.6514	.7242	.7244	.7138
at-0.10	.5288	.5457	.5797	.5955	.5861	.5956
A-P	.2794	.2976	.3214	.2819	.2885	.2898
pre-at-5	.3917	.3875	.4042	.4583	.4667	.4500
pre-at-10	.3396	.3604	.3667	.3708	.3833	.3771
R-P	.3278	.3265	.3618	.3060	.3101	.3110

The results show that the three strategies obtain rather different average retrieval performance, and the differences are significant. On close inspection, the strategy proposed in this study works markedly better than others.

The statistics from the 5th to 7th column, in Table 2, however, give us the dissimilar view when the different strategies are applied to collection FT for the queries 351-400. The results show that it seems that the retrieval with two strategies of query expansion does not achieve markedly better performance than the retrieval without query expansion.

It is clear that any query expansion approach may behave very differently depending on the quality of the initial retrieval. The negative effects of using poor pseudo-relevant documents for the query expansion are well-known. As some studies in earlier literature have shown, query expansion often has a negative effect in retrieval effectiveness regardless of the source of the candidate terms unless (and sometimes even if) relevance feedback is employed [9] [10]. There is ample evidence to indicate that improvements in retrieval effectiveness do not occur during query expansion unless the terms added are good terms.

In fact, all the results shown above are obtained averaging over the set of queries $q \in$ Q. In what follows, we will examine carefully the results in Table 2 for the collection FT, and give an alternative way to further investigate how the retrieval effectiveness varies as the variations of the quality of the query by means of a query by query analysis.

4.4 Query Quality and Query expansion

Query quality is quantitatively characterised in this study by the evaluation measure $qq(q)$ = $prec\text{-}at\text{-}\alpha\text{-}docs(q)$ obtained from the initial retrieval with respect to the given query $q \in$ Q. The greater the value $qq(q)$ is, the higher quality the query q has. In fact, the query quality $qq(q)$ highly depends on the specific retrieval system. In our experiments, the parameters a is set to 10, it hence takes

Table 3: Comparison of performance query by query

	0	1	2	3	4	5	6
q-id	qq(q)	A-P	R-P	A-P	R-P	A-P	R-P
351	.9000	.5891	.6429	.6774	.6786	.6828	.6786
353	1.000	.3287	.3673	.4361	.4694	.4194	.4286
365	.8000	.8002	.7273	.8974	.7273	.8609	.7273
366	.8000	.7297	.7895	.7931	.7895	.7900	.7895
391	.8000	.2051	.2892	.2112	.3012	.2200	.3193
392	.8000	.4016	.4600	.4527	.5000	.4365	.4800
354	.6000	.2314	.3944	.2102	.3521	.2334	.3521
357	.6000	.2361	.3750	.2769	.4062	.2791	.4062
367	.7000	.2696	.3243	.2725	.3243	.2866	.3243
372	.5000	.4419	.4000	.5441	.5333	.5034	.5333
373	.5000	.4097	.5000	.4665	.5000	.4675	.5000
374	.7000	.3110	.4110	.3332	.4110	.3151	.3973
377	.5000	.5350	.4000	.5479	.4000	.5501	.4000
385	.5000	.1966	.2647	.3000	.3529	.2894	.3235
395	.5000	.1183	.2796	.1334	.3011	.1335	.3118
396	.6000	.6640	.6000	.6742	.6000	.6800	.6000
398	.6000	.4233	.4211	.4193	.4211	.4198	.4211
400	.6000	.3830	.4600	.4310	.5600	.4051	.5200
352	.3000	.0237	.0795	.0649	.1464	.0588	.1381
356	.3000	.2058	.2000	.2065	.2000	.2112	.2000
360	.3000	.1732	.2857	.1760	.2857	.1789	.3571
364	.3000	.7667	.6667	.3611	.3333	.3611	.3333
368	.4000	.5833	.6000	.5482	.4000	.5603	.4000
370	.4000	.2365	.2800	.2439	.3200	.2628	.3600
375	.4000	.3492	.3077	.3578	.3077	.3311	.3077
382	.4000	.5952	.5000	.6429	.6667	.6429	.6667
383	.4000	.0677	.1860	.1009	.1977	.0717	.1860
387	.3000	.2557	.2308	.2074	.2308	.2101	.2308
388	.4000	.3169	.4000	.2560	.2667	.2617	.2667
390	.3000	.1629	.2353	.2145	.3725	.2090	.3333
399	.3000	.2213	.3000	.2319	.3000	.2194	.3000
355	.0000	.0068	.0000	.0077	.0000	.0079	.0000
358	.0000	.0000	.0000	.0000	.0000	.0000	.0000
359	.1000	.0293	.0909	.0369	.0909	.0398	.0909
361	.0000	.0000	.0000	.0000	.0000	.0000	.0000
362	.2000	.6667	.6667	.6808	.6667	.5689	.6667
363	.0000	.0164	.0000	.0121	.0000	.0147	.0000
369	.0000	.0250	.0000	.0000	.0000	.0139	.0000
371	.0000	.0269	.0000	.0093	.0000	.0102	.0000
376	.1000	.0375	.1071	.0677	.1786	.0728	.1786
378	.0000	.0060	.0909	.0022	.0568	.0015	.0455
379	.0000	.0000	.0000	.0000	.0000	.0000	.0000
380	.1000	1.000	1.000	1.000	1.000	1.000	1.000
381	.0000	.0041	.0000	.0049	.0000	.0049	.0000
384	.0000	.0195	.0000	.0209	.0000	.0294	.0000
386	.1000	.0417	.0000	.0500	.0000	.0500	.0000
389	.1000	.0024	.0071	.0073	.0143	.0024	.0071
393	.2000	.1481	.2222	.0802	.2222	.1429	.2222
394	.2000	.0786	.0000	.0650	.0000	.0786	.0000
397	.2000	.1916	.1250	.1135	.0000	.1401	.1250

$qq(q) = \textit{prec-at-}10\textit{-docs}(q)$.

Table 3 gives the statistics, as discussed in Table 2 for the collection FT, but query by query. The statistics that express the query qualities $qq(q)$ can be found in the column 0. The statistics that apply to the original queries can be found in the column 1 and 2. Column 3 and 4 pertain to apply reduced formula, and column 5 and 6 pertain to apply scheme c of component estimate proposed in this study.

The results in Table 3 show distinct levels of improvement in retrieval performance query by query. In fact, if we split intervals [0, 1] for the domain of the evaluation measure $qq(q)$, and compute the performance of each query within corresponding sub-interval, then it is easily seen that the performance improvement is stable and significant in the interval $(0.7, 1.0]$. The performance improvement is relatively stable and relative significant in the interval $(0.4, 0.7]$. Some exceptions (i.e., queries are harmed because of expansion) in this range emerge indeed, but it is infrequent. Whereas in the interval $(0.2, 0.4]$, it seems that there is no clear pattern to emerge. There is a relatively greater fluctuation compared with ones in the intervals $(0.4, 0.7]$ and $(0.7, 1.0]$, but the performance improvement still tends to be positive, even though the improvement for some queries might be insignificant, or even be zero or negative. In the interval $[0.0,\ 0.2]$, however, the performance improvement almost tends to zero, even tends to negative, even though the improvement might be emerged for some queries, it is in general negligible.

The results in Table 3 also give us a view that two strategies of query expansion are consistently better in the improvement of performance than the initial retrieval for all queries with the qualities in the interval $(0.7, 1.0]$, and for most queries with the qualities in the interval $(0.4, 0.7]$. The level of improvement is substantial at most evaluation points. On close inspection, the performance when applying reduced formula seems slightly better than the one applying our approach, but the differences are inconsiderable.

Table 4: Comparison of performance query by query

	0	1	2	3	4	5	6
q-id	qq(q)	A-P	R-P	A-P	R-P	A-P	R-P
202	1.000	.7324	.7115	.8021	.8077	.7821	.7692
215	.8000	.5594	.5763	.5657	.6441	.5981	.6949
237	.8000	.3548	.4516	.4382	.4355	.4884	.4839
244	.8000	.3831	.5069	.3895	.4792	.3930	.5000
247	.8000	.7088	.7273	.8086	.6364	.8378	.7273

Similarly, Table 4 give the statistics, as discussed in Table 3 for the collection AP90, but query by query. The results, on the relation between measure $qq(q)$ and improvement in retrieval performance, give us rather similar view to the results of FT. The only difference is that the performance applying our approach seems better than one applying reduced formula for those queries with the qualities in the interval $(0.7, 1.0]$.

In addition, from Table 3 we see that it seems that, for those queries that are harmed due to applying the techniques of query expansion, the degree of average harm by using our approach is less than the one by using reduced formula.

In fact, in order to avoid the insignificant, zero and negative effects of query expansion, one should attempt to produce a higher query quality by improvement of the initial retrieval. Finally, we conclude that the approach studied in this study for query expansion is practical and effective, provided that it is used in conjunction with an efficient retrieval system.

5. AN ALTERNATIVE

We also investigate an alternative approach which applies the concept of relative entropy studied in information theory to the technique of automatic query expansion.

Let the distributions $P_D(t)$ and $P_{\Re}(t)$ be defined as in section 3.3. The relative *entropy* due to Kullback & Leibler [5] is defined by

$$I(P_{\Re}(t); P_D(t)) = \sum_{t \in V} P_{\Re}(t) \log \frac{P_{\Re}(t)}{P_D(t)},$$

which can be used to measure the average difference of the information contained in $P_{\Re}(t)$ and that contained in $P_D(t)$ about $P_{\Re}(t)$.

Obviously, the relative entropy is not necessarily symmetric, i.e., $I(P_{\Re} : P_D) = I(P_D : P_{\Re})$ does not hold always, even though, in practical, it is perhaps useful to think of relative entropy as a 'distance' between distributions.

Taking advantage of the relative entropy, the scoring function can be given by

$$score'_q(t) = P_{\Re}(t) \log \frac{P_{\Re}(t)}{P_D(t)} \qquad (t \in V^{\Re}),$$

which has the same interpretation as scoring function discussed in section 3.3 and 3.4. Namely, because the $I(P_{\Re} : P_D)$ can be used to measure the difference of the information contained in $P_{\Re}(t)$ and that contained in $P_D(t)$ about $P_{\Re}(t)$, while the difference is assumed probably to reveal some semantic relations between terms $t \in V$ $(\supseteq V^{\Re})$ and terms $t \in V^q$, therefore, the expansion terms selected (from $t \in V^{\Re}$) should be those, which mostly contribute to the difference.

It is worth mentioning that the item f_1 (t) = $P_\Re(t) \log \frac{P_\Re(t)}{P_D(t)}$ in the expression of $I(P_\Re : P_D)$ is well-defined even when $P_\Re(t)$ = 0 $(t \in V - V^\Re)$. Noted that, in this case, we implicitly use the conventions: 0 . log($\frac{0}{0}$) = 0 when $P_D(t)$ = 0 and $0 \cdot \log(\frac{0}{P_D(t)})$ = 0 when $P_D(t) \neq$ 0 for $t \in$ *V-V'.* Therefore, we can directly apply $I(P_\Re : P_D)$ as a dissimilarity measure for the distributions $P_\Re(t)$ and $P_D(t)$ *(t* $\in$ *V).* There is no need to relocate the probability densities as discussed for the divergence $J(P_\Re, P_D)$ in the sections 3.3 and 3.4, and the necessary arguments could be much simpler.

The duplicate experiments are also carried out by using $score'_q(t)$ to expand two sets of queries (201-250 and 351-400), and applying it to two collections AP and FT. A very interesting fact in our experiments is that using two approaches, applying the relative entropy $I(P_\Re : P_D)$ and the divergence $J(P_\Re, P_D)$ to construct scoring functions, to expand queries arrives at very consistent results of improvement in retrieval performance.

6. REMARK

We would like to point out that the arguments in sections 3.3 and 3.4 are theoretically necessary. First, our discussions provide a theoretical basis to be able to directly apply the concept of divergence as a dissimilarity measure to construct score functions for query expansion, and in some similar applications of divergence, as long as $P_\Re(t)$ and $P_D(t)$ at least satisfy the postulates: $P_\Re(t)$ > 0 and $P_D(t)$ > 0 when $t \in V^\Re \subseteq$ *V,* in which *V* is a finite tuple.

Furthermore, noted that the equality

$$J(P_\Re(t), P_D(t)) = I(P_\Re(t) : P_D(t)) + I(P_D(t) : P_\Re(t))$$

does not hold for the distributions $P_\Re(t)$ and $P_D(t)$ over the same probability space $(V, 2^V)$. This is because the corresponding items

$$f(t) = (P_\Re(t) - P_D(t)) \log \frac{P_\Re(t)}{P_D(t)},$$

$$f_1(t) = P_\Re(t) \log \frac{P_\Re(t)}{P_D(t)}, \quad f_2(t) = P_D(t) \log \frac{P_D(t)}{P_\Re(t)}$$

do not always satisfy $f(t) = f_1(t) + f_2(t)$ since both $f(t)$ and $f_2(t)$ are undefined when $P_\Re(t)$ = 0 (but $P_D(t) \neq$ 0) for $t \in V - V^\Re$.

However, based on the arguments in sections 3.3 and 3.4, the equality below is perfectly meaningful. That is,

$$J(P_\Re(t, \mathrm{to}, t^*), P_D(t, \mathrm{to}, t^*))$$

$$= I(P_\Re(t, t_0, t^*) : P_D(t, t_0, t^*)) + I(P_D(t, t_0, t^*) : P_\Re(t, t_0, t^*))$$

does ***hold*** for the distributions $P_\Re($*t, to,* $t^*)$ and $P_D(t,\ t_0,\ t^*)$ over the same probability space $(V_{t^*}^\Re, 2^{V_{t^*}^\Re})$.

Consequently, we immediately elicit that

$$\textit{score,(t)} = score'_q(t) + f_2(t) \qquad (t \in V^\Re = V_{t^*}^\Re - \{t^*\}).$$

It is interesting to note that the values of $f_2(t)$ are in our experiments inconsiderable compared with the values $score'_q(t)$ when $t \in$ *V”.* This is because the densities $P_D(t)$ are relatively much smaller then the densities $P_\Re(t)$ when $t \in V^\Re$ if we take the set $\Re$ very small (such as α = 10). It is probably the best explanation of the reason why the scoring functions ***score,(t)*** and $score'_q(t)$ achieve a very consistent effect on the improvement of retrieval performance.

7. CONCLUSIONS

This study focuses on the subject of query expansion within a theoretical framework. An important theoretical problem remaining is discussed, which directly forms the basis of our approaches. A formal model is then designed based on the strategies of applying divergence and relative entropy as dissimilarity measures in a practical retrieval environment. By means of such measures, a methodology to assign scores to candidate terms for query expansion is proposed. The query expansion procedure has been designed, implemented, and evaluated on two standard TREC collections. The preliminary results obtained so far are encouraging and further development and evaluation of the model are in progress.

The problems, however, of how to approximately estimate the information distribution by the term distribution for a certain population, and of what is a better way to generate a good sample population for such kind of estimate, remain open problem. It will be impossible to build a effective retrieval system with the mechanism of query expansion without the satisfactory solutions of these problems.

In addition, in this study, we only report the experimental results on two seemingly homogeneous TREC collections. Further work will be to apply the proposed approaches to heterogeneous collections such as a full TREC volume.

Acknowledgements

The authors are very grateful to the anonymous reviewers for their detailed comments. Special thanks are to Ian Ruthven and Gianni Amati for their valuable advice on this study. Hearty thanks are extended to Juliet van Rijsbergen for her help in improving the readability of the paper.

This research is supported in part by EPSRC.

References

[1] Efthimiadis, E. N. Query Expansion. *Annual Review of Information Systems and Technology,* 31, pp.121-187, 1996.

[2] Robertson, S. E. On Terms Selection for Query Expansion. *Journal Of* Documentation, 46(4), pp.359-364, 1990.

[3] Carpineto, C., Mori, R. and Romano, G. Informative Term Selection for Automatic Querv Expansion. In *The 7th Text REtrieval Conference,* pp.363-369, 1998.

[4] Carpineto, C. and Romano, G. (1999). TREC-8 Automatic Ad-Hoc Experiments at Fondazione Ugo Bordoni. In *The 8th Text REtrieval* Conference, pp.377-380, 1999.

[5] Kullback, S. and Leibler, R. A. On information and sufficiency. *Annual Mathematical Statistics, 22,* pp.79-86, 1951.

[6] Rocchio, J. Relevance Feedback in Information Retrieval. *The SMART Retrieval System Experiments* in *Automatic Document Processing,* Prentice Hall Inc., pp.313-323, 1971.

[7] Buckley, C. and Salton, G. Optimisation of Relevance Feedback Weights. In *Proceedings of the 18th ACM-SIGIR Conference,* pp.351-357, 1995.

[8] Robertson, S. E., Walker, S. and Beaulieu, M. Okapi at TREC-7: Automatic ad hoc, Filtering, VLC, and Interactive Track. In *The 7th Text REtrieval* Conference, pp.363-369, 1998.

[9] Van Rijsbergen, C. J., Harper, D. and Porter, M. The Selection of Good Search Terms. Information Processing and *Management,* 17(2), pp.77-91, 1981.

[10] Smeaton, A. and Van Rijsbergen C. J. The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. *The Computer Journal,* 26(3), pp.10-18, 1983.