JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 17, 1-xxx (2001)

# Short Paper

## A Linguistic Approach to a Chinese Query Model

TYNE LIANG AND TZYY-CHYANG WU
*Institute of Computer and Information Science*
*National Chiao Tung University*
*Hsinchu, Taiwan 300, R.O.C.*
*E-mail: tliang@cis.nctu.edu.tw*

As network and input technologies have improved rapidly, researches in information retrieval have become more important than ever. An user-friendly and efficient query interface will certainly facilitate utilization of an information retrieval system. In this paper, a linguistic approach to Chinese query models is proposed and investigated. It includes Chinese linguistic variables, quantifiers and combination queries. Meanwhile, an improved computation method is proposed to handle general conjunctive queries, and it is proved to be better than traditional fuzzy computation in terms of a higher level of retrieval satisfaction. Experimental results also verify that the proposed quantifiers do not only simplify a multi-criteria request, but also yield a higher level of retrieval satisfaction than do Boolean queries.

***Keywords:*** linguistic variables, quantifier, query model, aggregation model, Chinese texts

### 1. INTRODUCTION

Most existing present retrieval systems employ a binary Boolean model because it is simple to implement and easy to use though it is unable to fully reveal the relevance of retrieved documents with respect to the user's inquiries. To solve this problem, an extended Boolean model was proposed by Salton [1] in which each query is extended and assigned a numerical weight value. However, it is usually difficult for users to give appropriate values to query terms.

In view of this drawback, Bordogna [2] proposed a linguistic query model based on linguistic variables introduced by Zadeh [3-5]. According to their individual fuzziness, linguistic variables can be classified into four types, namely, qualification, modification, the quantification and composition variables. It is qualification type and modification type variables that Bordogna used to define his linguistic query prototype. In the prototype, relevance evaluation between documents and queries is carried out using a membership function proposed by Bosc [6]. Since this function produces continuous values, it

Received November 17, 1998; revised May 21, 1999; accepted July 26, 1999.
Communicated by Wen-Lian Hsu.

can be used to describe, say, the different important levels of a query term.

Another type of linguistic variable applied to query expressions consists of quantifiers with which Yager [7] proposed ordered weight aggregation and Bordogna [8, 9] proposed aggregation operators. They are essentially equivalent, and both of them provide simpler and more natural ways to express queries.

In this paper, a linguistic query model composed of linguistic variables and quantifiers particularly useful for Chinese textual retrieval systems are proposed and investigated. In addition, a new computation model for conjunctive queries is proposed and is proved to be superior to traditional fuzzy computation in terms of the retrieval results. The proposed aggregation model has been tested using real test queries, and the results indicate that it can yield level of higher retrieval satisfaction than a Boolean model can in both weighted and unweighted cases.

## 2. LINGUISTIC VARIABLE MODEL

It is usual for an end-user to issue a query containing several query terms in which some terms are more important than others. To describe different degrees of importance for query terms, the traditional fuzzy approach assigns terms with different values between zero and one in such a way that more important terms receive high values. However it is not easy for users to assign query terms with appropriate numerical values.

Unlike assignment of a single numerical value, the linguistic approach uses a linguistic variable which may be a common adjective in natural language to describe a query term. For example, the linguistic variable '重要' ('important') can be used to describe a query term, say, 'image' as an important search term used by end-users. Therefore, end-users can avoid query term weight assignment. In following subsection, we will describe the proposed linguistic variables as well as their membership functions.

### 2.1 Linguistic Variables

Among numerous adjectives in Chinese, '重要' can be used to describe the importance of a query term to a certain degree. Hence, the proposed linguistic variables are derived from '重要'. The definition of linguistic variables consists of quadruple {L, T(L), U, G }. L is the name of the linguistic variable, T(L) is a linguistic variable set that the linguistic variable L produces, U is the domain of the base variable, and G is the context-free grammar. Therefore, '重要' is defined as follows:

L : 重要;
T(L): the linguistic variable set;
U: the base variable domain [0,1];
G: the context-free grammar consisting of quadruple {T, N, P, S}, where T is the set of terminal symbols, N is the set of non-terminal symbols, P is the production rules and S is the start symbol;
T={重要, 很, 非常, 有點, 不};
N={<linguistic variable>, <primary term>, <hedge>};
P={<linguistic variable>::=[hedge] [不] <primary term> | 不存在

<primary term>::= <重要>
<hedge>::=非常非常 | 很 | 有點 };
S=<linguistic variable>.

There are nine different linguistic variables to be generated, namely, '非常非常重要' ('very very important'), '很重要' ('very important'), '重要' ('important'), '有點重要' 'rather important', 有點不重要 'rather unimportant', '不重要' ('unimportant'), '很不重要' ('very unimportant'), '非常非常不重要' ('very very unimportant'), '不存在' ('non-existent').

It should be pointed out that the reason why we propose this set of linguistic variables derived from '重要' is that this variable is essential and relevant to the proposed weighting function. Other linguistic variables can be used so long as appropriate weighting functions can be found and quantified. For example, the weighting function of the set of linguistic variables derived from '有影響力的' ('influential') may be related to the number of times a document is cited for a particular query.

### 2.2 Membership Function Computation

Each linguistic variable employs its corresponding membership function to produce a numerical value between zero and one. This value will be used as a retrieval status value for a retrieved document, and a higher value means higher association between a document and a given linguistic query term. The following function was proposed by Bordogna [2] and can be used for the linguistic variable '重要':

$$\mu_{\text{重要}}(value) = \begin{cases} e^{(value-i)^2 \ln C} & \text{if } 0 \le value < \text{i}; \\ 1 & \text{if i} \le value < \text{j}; \\ e^{(value-j)^2 \ln C} & \text{if j} \le value \le 1; \end{cases} \tag{2.1}$$

in which (i, j) is the input interval for the variable '重要' and $C$ is a control parameter defined as following

$$C = e^{(\ln 10^{-n} / (x_0 - i)^2)} \tag{2.2}$$

In equation-2.2, $n$ is the function value that approaches the $n$th digit below decimal; $C$ is related to the degree of compatibility distribution. The smaller $C$ is, the narrower the distribution is. On the other hand, the larger $C$ is the smoother the distribution. $x_0$ is the minimal number in all input values, and the *value* in equation-2.1 is generally a term weight in a certain document. Since the nine linguistic variables produced are derived from '重要' and are used to describe different levels of importance, one can assign different input intervals for each of them. For example, the input interval can be set to be [0.85, 1.0] for '非常非常重要' to indicate that only when the term weight in a certain document is between 0.85 and 1.0 can this document obtain the maximum retrieval status value. On the other hand, '非常非常不重要' with the input intervals [0.0, 0.03] indicates that a retrieved document will also get the maximal value while its query term really will play a minor role and get quite a small term weight in the document.

### 2.3 Experiments and Analysis

The experiments were conducted using a collection of 2000 Chinese technical reports, and each document was represented using a list of author-given keywords which were used as query terms. Since all the keywords were weighted to reflect their significance in each individual document, a document-to-keyword matrix was used to store the term weights. The weight $M_{ij}$ of term $j$ in a document $i$, was normalized by the maximal weight of term $j$ in all $n$'s documents and was calculated as follows:

$$M_{ij} = \frac{\sqrt{w_{ij}}}{\max(w_{1j}, w_{2j}, ..., w_{nj})}, \tag{2.3}$$

where

$w_{ij}$ = 2 × occurrence frequency of term j in titles + 3 × the occurrence frequency of term j in the keyword set + the occurrence frequency of term j in abstracts.

During retrieval, $M_{ij}$ is used as the input value in Equation-2.1 to calculate the membership function for each linguistic variable.

It is typical that a general query may contain a certain number of query terms connected by AND-connectors and expressed using different or same linguistic variables. Traditionally, the retrieval status value $RSV\_T_i$ of a retrieved document $i$ with respect to a conjunctive query containing $p$'s terms takes the minimum of the membership values as the retrieval status value and is expressed as follows:

$$RSV\_T_i = min\ (u_{i1}, u_{i2}, \ldots, u_{ip}). \tag{2.4}$$

However, it is noted that $RSV\_T_i$ will be dominated by the lowest of various membership values. To avoid this kind of bias, we propose an improved computation model in which various semantic strengths of linguistic variables are taken into account. For example, the proposed nine linguistic variables are divided into the following five classes according to their semantic strength.

The first class : 不存在.
The second class : 非常非常重要, 非常非常不重要.
The third class : 很重要, 很不重要.
The fourth class : 重要, 不重要.
The fifth class : 有點重要, 有點不重要.

To these nine linguistic variables are assigned different control parameters $c_j$. The value of $c_j$ is decided based on two principles. One is that the linguistic variables in low classes will receive larger values of $c_j$ than will those variables in high classes. Another is that linguistic variables with positive meanings in the same class will obtain larger values of $c_j$ than will those variables with negative meanings. Table 1 shows the values of $c_j$ for these nine linguistic variables.

**Table 1. The control parameter values of the nine linguistic variables.**

| Linguistic variable | 非常非常重要 | 很重要 | 重要 | 有點重要 | 有點不重要 | 不重要 | 很不重要 | 非常非常不重要 | 不存在 |
|---|---|---|---|---|---|---|---|---|---|
| $c_j$ | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 6 |

The computation of the retrieval status value $RSV\_I_i$ for a certain document $i$ in the improved model becomes

$$RSV\_I_i = \frac{\sum_{j=1}^{p} c_j \mu_{ij}}{\sum_{j=1}^{p} c_j}. \tag{2.5}$$

Afterwards, all the retrieved documents will be output in the order of their retrieval status values. To compare the traditional model with the improved model, we use Spearman's rank correlation coefficient [10] to evaluate the linear relation between $x$ serial (the retrieved output order) and $y$ serial (the actual result order decided by an expert). The formula is below

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}. \tag{2.6}$$

In Equation-2.6, $d_i$ represents the difference between serial $x_i$ and serial $y_i$, $n$ is the number of test documents, and $r_s$ is a value between –1 and 1 and is defined as retrieval satisfaction. When $r_s$ is 1, this means that there is a complete relation between $x$ and $y$ (i.e. two series are completely the same); when $r_s$ is –1, this means that the opposite situation exists; and zero means that there is no relation between the two series. Hence, a larger $r_s$ value implies that the model will achieve better performance.

In the experiments, the top five documents in their order of retrieval status values were retrieved. To compare the improved model with the traditional model, an expert first combined the documents retrieved by the two models and decided on a real list of documents (which may have had more than five documents) relevant to the queries. Then, Equation-2.6 was used to calculate the retrieval satisfaction. Fig. 1 shows that the improved model outperformed the traditional model in a test using twenty conjunctive queries containing two terms (e.g. '非常非常重要(平行編譯器)，有點重要 (語法)') and three terms (e.g. '非常非常重要(光纖)，很重要(網路)，有點重要(通訊)'), respectively. Similar results, as shown in Fig. 2, were obtained in a test using thirty conjunctive queries classified according to their linguistic semantic strength. For example, a conjunctive query with linguistic variables from the second class looked like '非常非常重要 (平行編譯器)，非常非常不重要 (語法)', and a query with linguistic variables from the third class looked like '很重要 (網路)，很不重要 (光纖)'. Ten queries from each class were tested in the experiments.

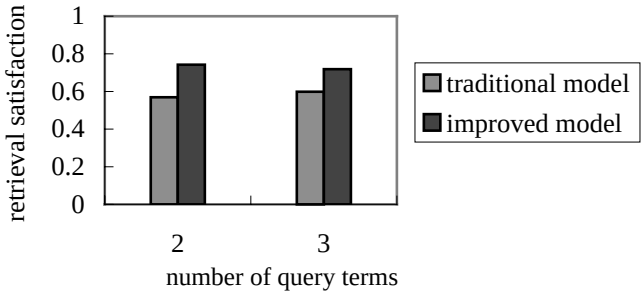


Fig. 1. Traditional model vs. improved model w.r.t. various queries.

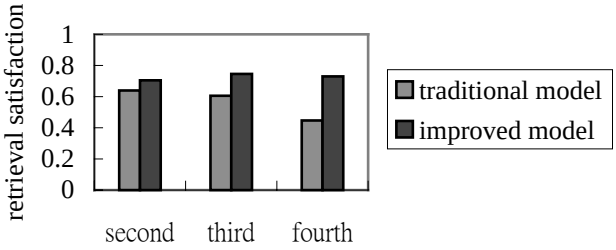


Fig. 2. Traditional model vs. improved model w.r.t. various variables.

## 3. AGGREGATION MODEL

According to [7], the aggregation model provides a more natural and easier way to express multi-criteria retrieval. For example, if a user wants to retrieve documents containing at least three terms with respect to a four-term query, a more natural query can be expressed using a quantifier '至少 3' ('at least three') as follows:
Query = 至少 3(image, digital, analysis, compression).

On the other hand it will be expressed using a Boolean query as follows:

Query = {(image AND digital AND analysis ) OR
(digital AND analysis AND compression ) OR
(analysis AND image AND compression ) OR
(image AND digital AND analysis AND compression ) }.

In the following subsections, we will show that the proposed quantifiers not only facilitate multi-criteria information retrieval, but also provide a higher level of retrieval satisfaction than do traditional Boolean expressions.

### 3.1 Quantifiers

The proposed quantifiers are designed so that they can be practically used in real queries. They are generated according to the following production rules:

<quantifier> ::=<atomic term> |<composite term>;
<atomic term> ::=<全部滿足>|<大部份滿足>|<小部份滿足>|<一半滿足>;
<composite term>::=<quantity adverb><conditional number>;
<quantity adverb>::= 至少 | 至多 | 大於 | 小於 | 剛好;
<conditional number>::=K.

Based on the production rules, nine quantifiers are generated and classified into the following three classes according to their meanings:

(1) Monotone: The retrieval status value of a retrieved document increases as more inquiry criteria (i.e., the number of search terms) are satisfied. The quantifiers are: '至少 K' ('at least K criteria are satisfied'), '大於 K' ('more than K criteria are satisfied'), '小部份滿足'('a small portion of the criteria are satisfied'), '大部份滿足' ('most of the criteria are satisfied'), and '一半滿足' ('half of the criteria are satisfied').
(2) Anti-monotone: The retrieval status value of a retrieved document decreases as more criteria are satisfied. The quantifiers are: '至多 K' ('at most K criteria are satisfied'), and '小於 K' ('less than K criteria are satisfied').
(3) Uni-modal: The retrieval status value of a retrieved document reaches one, the maximum, when the exact number of criteria are satisfied. The quantifiers are: '剛好 K' ('exactly K criteria are satisfied') and '全部滿足' ('all the criteria are satisfied')

### 3.2 Retrieval Status Value and Weighting Value Computation

For an unweighted multi-criteria request, the retrieval status value of a retrieved document will always reach one, the maximum value, whenever exactly the number of request criteria are satisfied; otherwise, it will be zero, the minimum value, in the traditional Boolean model. For example, those documents containing at least K query terms will receive a value of one for the quantifier '至少 K'. However the unweighted aggregation model will compute the retrieval status value $RSV\_U_i$ of a retrieved document $i$ containing $t$-query terms as the summation of the weighting function values $W_p$*('quantifier')* as follows:

$$RSV\_U_i = \begin{cases} \sum_{p=1}^{t} W_p(\text{'}quantifier\text{'}) & \\ 1 & \text{if } \sum_{p=1}^{t} W_p(\text{'}quantifier\text{'}) > 1 \end{cases}, \tag{3.0}$$

where $W_p$*('quantifier')* is the value computed using the weighting formulas specifically designed for each of the proposed quantifiers. For the sake of clarity, the total number of criteria (indicated by $N$) is assumed to be eight, and the conditional number K is set to be four in each of the following quantifiers.

(1) 至少 K:
The weighting function for '至少 K' is designed to yield a large value so that the retrieval status value of a document will be close to one when it matches at least K search terms:

$$W_p(\text{'至少}K\text{'}) = \begin{cases} \dfrac{p+1}{\sum_{s=1}^{K+1} s} & \text{if } 0 < p \le K; \\ \dfrac{1}{\sum_{s=1}^{K+1} s \times (N-K)} & \text{if K} < p \le N; \end{cases} \tag{3.1}$$

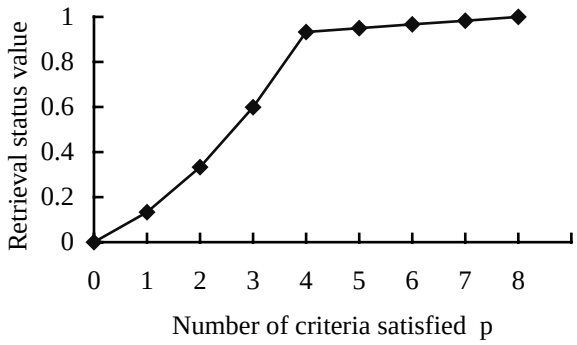


Fig. 3. The retrieval status value for '至少 4'.

(2) 大於 K:
The weighting function is designed to produce a large value so long as a document satisfies more than K criteria:

$$W_p(\text{'大於}K\text{'}) = \begin{cases} \dfrac{1}{\sum_{s=1}^{N-K+1} s \times K} & \text{if } 0 < p \le K; \\ \dfrac{(N-p+2)}{\sum_{s=1}^{N-K+1} s} & \text{if K} < p \le N; \end{cases} \tag{3.2}$$

(3) 小部份滿足:
The weighting function for this quantifier is designed to yield small values when a document contains more than one third of keywords that a request needs:

$$W_p(\text{'小部份滿足'}) = \begin{cases} \dfrac{p+1}{\sum_{s=1}^{\lfloor \frac{N}{3} \rfloor + 1} s} & \text{if } 0 < p \le \lceil N/3 \rceil; \\ \dfrac{1}{\sum_{s=1}^{\lfloor \frac{N}{3} \rfloor + 1} s \times (N - \lceil \frac{N}{3} \rceil)} & \text{if } \lceil N/3 \rceil < p \le N; \end{cases} \tag{3.3}$$

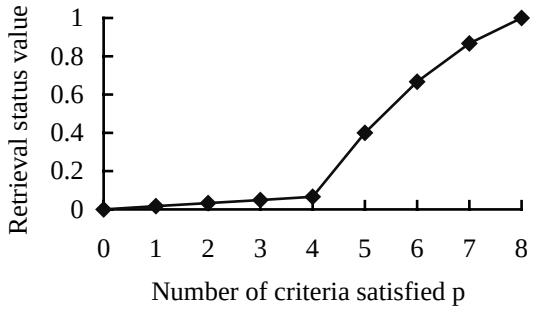


Fig. 4. The retrieval status value for '大於 4'.

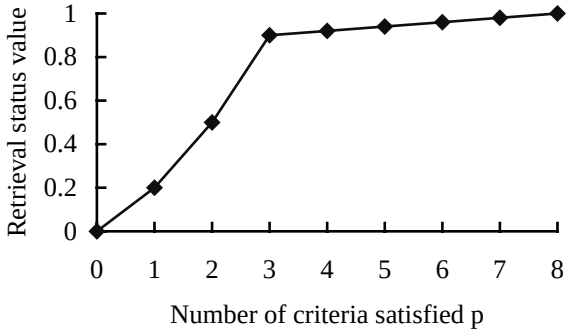


Fig. 5. The retrieval status value for '小部份滿足'.

(4) 一半滿足:

The weighting function is almost the same as that for '小部份滿足' except that the number of satisfied criteria is set to be half of the keywords that the request needs:

$$W_p('\text{一半滿足}') = \begin{cases} \sqrt[p+1]{\sum_{s=1}^{\left\lfloor \frac{N}{2} \right\rfloor+1} s} & \text{if } 0 < p \le \left\lceil N/2 \right\rceil; \\ \sqrt[1]{\sum_{s=1}^{\left\lfloor \frac{n}{2} \right\rfloor+1} s \times \left( N - \left\lceil \frac{N}{2} \right\rceil \right)} & \text{if } \left\lceil N/2 \right\rceil < p \le N; \end{cases} \tag{3.4}$$

(5) 大部份滿足:

The weighting function is almost the same as that for '小部份滿足' except that the number of satisfied criteria is set to be two-third of the keywords that the request needs:

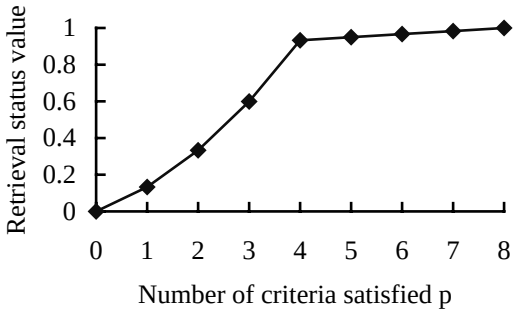


Fig. 6. The retrieval status value for ‘一半滿足’.

$$W_p(\text{'大部份滿足'}) = \begin{cases} \dfrac{1}{\sum_{s=1}^{\left\lfloor \frac{N}{3} \right\rfloor+1} s \times \left\lfloor \frac{2N}{3} \right\rfloor} & \text{if } 0 < p \le 2N/3; \\ \dfrac{(n-p+2)}{\sum_{s=1}^{\left\lfloor \frac{N}{3} \right\rfloor+1} s} & \text{if } 2N/3 < p \le N; \end{cases} \tag{3.5}$$

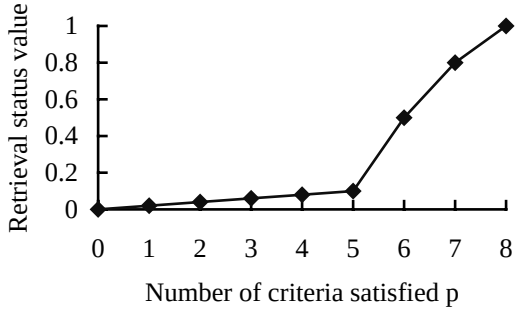


Fig. 7. The retrieval status value for ‘大部份滿足’.

(6) 至多 K:
The function for ‘至多 K’ is designed to yield a negative retrieval status value whenever more than K criteria are satisfied:

$$W_p(\text{'至多}K\text{'}) = \begin{cases} 1 & \text{if } 0 < p \le K; \\ -\left[ \dfrac{(N-p+1)}{\sum_{s=1}^{N-K} s} \right] & \text{if } K < p \le N; \end{cases} \tag{3.6}$$

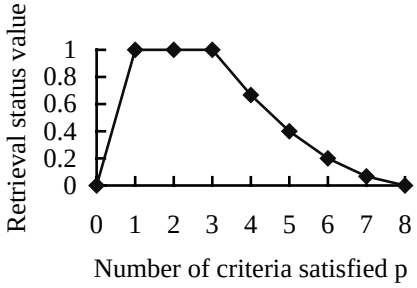


Fig. 8. The retrieval status value for '至多 4'.

(7) 小於 K:
Like the function designed for '至多 K', the function for '小於 K' will produce a negative value and cause the retrieval status value to decline whenever more than K criteria are satisfied:

$$W_p('小於K') = \begin{cases} 1 & \text{if } 0 < p < K; \\ -\left[ (N-p+1) \Big/ \sum_{s=1}^{N-K+1} s \right] & \text{if } K \le p \le N; \end{cases} \tag{3.7}$$

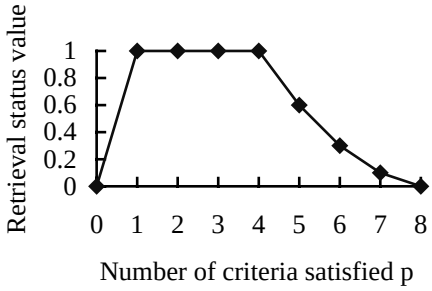


Fig. 9. The retrieval status value for '小於 4'.

(8) 剛好 K:
This function is designed to produce a positive value before K criteria are satisfied, and to produce a negative value when more than K criteria are satisfied:

$$W_p('剛好K') = \begin{cases} p \Big/ \sum_{s=1}^{K} s & \text{if } 0 < p \le K; \\ -(N-p+1) \Big/ \sum_{s=1}^{N-K} s & \text{if } K < p \le N; \end{cases} \tag{3.8}$$

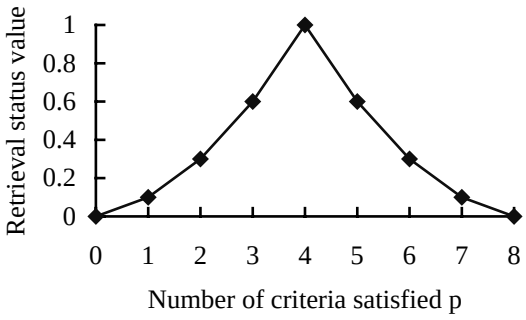


Fig. 10. The retrieval status value for '剛好 4'.

(9) 全部滿足:
The function for '全部滿足' is designed to produce a small value even when a document does not satisfy all the requested criteria:

$$W_p(\text{'全部滿足'}) = \begin{cases} 1/N & \text{if } p \neq N \\ 1 & \text{if } p = N \end{cases} \tag{3.9}$$

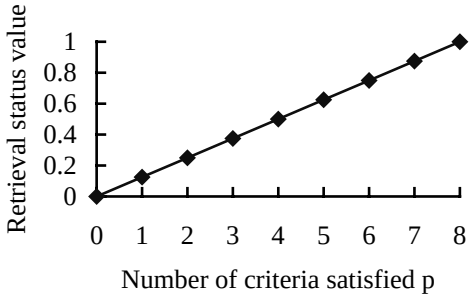


Fig. 11. The retrieval status value for '全部滿足'.

### 3.3 Experiments and Analysis

The experiments were carried out using the same corpus described in section 2, and each of the proposed quantifiers was tested using three conjunctive queries (e.g., 大於 2 (醫學, 影像, 資料庫, 超音波, 斷層掃瞄)). Two comparisons were made. One was made between the unweighted and weighted aggregation models. For the unweighted query model, the retrieval status value $RSV\_U_i$ was defined as in Equation-3.0. For the weighted model, the retrieval status value $RSV\_W_i$ was calculated as the summation of the weight product of the quantifier weight $W_p$*('quantifier')* and the term weight $M_{ij}$ (defined in Equation 2.3):

$$RSV\_W_i = \begin{cases} \sum_{j,p=1}^{t} W_p('quantifier')M_{ij} & \\ 1 & \text{if } \sum_{j,p=1}^{t} W_p('quantifier')M_{ij} > 1 \end{cases} \quad (3.10)$$

The performance was also evaluated in terms of retrieval satisfaction as defined in Equation 2-6. Fig. 12 shows that the weighted model yielded higher retrieval satisfaction than the unweighted model for all three different typed of quantifier.

Another comparison was made between the weighted aggregation model and the weighted Boolean model. Traditionally, a multi-criteria query in a Boolean model is transformed into a query connected with AND/OR. Then, the retrieval status value RSV-$B_i$ of a document i can be computed to take the minimal weight of terms connected by AND-connectors and the maximum weight of terms connected by OR-connectors. Therefore, we let

Q = {$q_1$, …, $q_N$}, where $q_i$ is the $i^{th}$ query term in a query containing *N* terms;
$S^u$(‘*quantifier’*)= the $u^{th}$ subset of Q which satisfies the quantifier;
M($S^u$(‘*quantifier’*)) = the set of weights corresponding to the elements in $S^u$(‘*quantifier’*);
then,

$$RSV\text{-}B_i = \max_{u=1}^{v} \{min\,(M(S^u(\text{‘}quantifier\text{’})))\}, \quad (3.12)$$

where v is the total number of $S^u$(‘*quantifier’*)

For example, if the quantifier is ‘至少 K’, K=3, N=4 and if
Query = (image, digital, analysis, compression) and
$Document_i$ = {(image, 0.4), (digital, 0.3), (analysis, 0.2), (compression, 0.1)},
then we get four subsets of Q satifying the quantifier ‘至少 3’;
$S^1$(‘至少 3’)={ image, digital, analysis};
$S^2$(‘至少 3’)={ image, digital, compression};
$S^3$(‘至少 3’)={ image, analysis, compression};
$S^4$(‘至少 3’)={ image, digital, analysis, compression};
M($S^1$(‘至少 3’))={0.4, 0.3, 0.2} and min(M($S^1$(‘至少 3’)))=0.2;
M($S^2$(‘至少 3’))= {0.4, 0.3, 0.1} and min(M($S^2$(‘至少 3’)))=0.1;
M($S^3$(‘至少 3’))= {0.4, 0.2, 0.1} and min(M($S^3$(‘至少 3’)))=0.1;
M($S^4$(‘至少$_4$3’))= {0.4, 0.3, 0.2, 0.1} and min(M($S^4$(‘至少 3’)))=0.1;
RSV-$B_i$= $\max_{u=1}$ {min(M($S^u$(‘至少 3’)))}=0.2.

Fig. 13 shows that in a test of twenty-seven multi-criteria queries the weighted aggregation model indeed outperforms weighted Boolean model in terms of higher retrieval satisfaction.

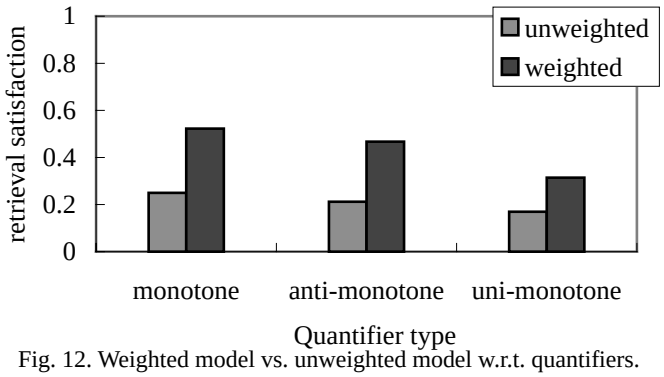


Fig. 12. Weighted model vs. unweighted model w.r.t. quantifiers.

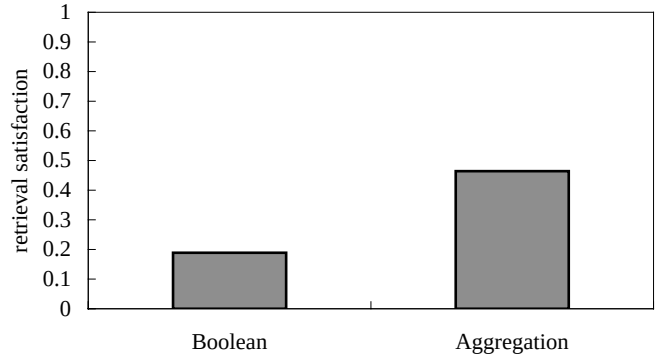


Fig. 13. Boolean model vs. Aggregation model.

## 4. COMBINATION QUERY

### 4.1 Query Grammar Structure

A general linguistic query may contain both a linguistic variable and the quantifier, such as 很重要 (全部滿足 (分散式, 資料庫)) 且 有些重要 (至少 1 (網路, 管理)). Following is the grammar:

<query> ::= < atomic query> | <composite query>
<atomic query> ::= linguistic variable ( query-term ) | quantifier ( query-term set )
| linguistic variable ( quantifier (query-term set ) )
<composite query> ::= <atomic query> *[ <connective> <atomic query>]
<connective> ::= 且

### 4.2 Experiments and Analysis

The experiments are carried out using twenty combination queries, and the intervals for the linguistic variables are shown in Table 2. In the experiments, four tests were car-

ried out to test the performance of the proposed linguistic approaches in different cases. The tests were as follows:

T1: the traditional computation with unweighted data;
T2: the traditional computation with weighted data;
T3: the improved computation with unweighted data;
T4: the improved computation with weighted data.

From Fig. 14 one can observe that the improved computation for conjunctive combination queries still outperforms the traditional computation in terms of higher retrieval satisfaction in both unweighted and weighted cases.

**Table 2. The intervals of the linguistic variables.**

| Linguistic variable | 非常非常重要 | 很重要 | 重要 | 有些重要 |
|---|---|---|---|---|
| Interval | (0.85, 1.0) | (0.71, 1.0) | (0.58, 1.0) | (0.35, 0.58) |
| Linguistic variable | 有些不重要 | 不重要 | 很不重要 | 非常非常不重要 |
| Interval | (0.16, 0.35) | (0.0, 0.16) | (0.0, 0.09) | (0.0, 0.03) |
| Linguistic variable | 不存在 | | | |
| Interval | (0.0, 0.0) | | | |

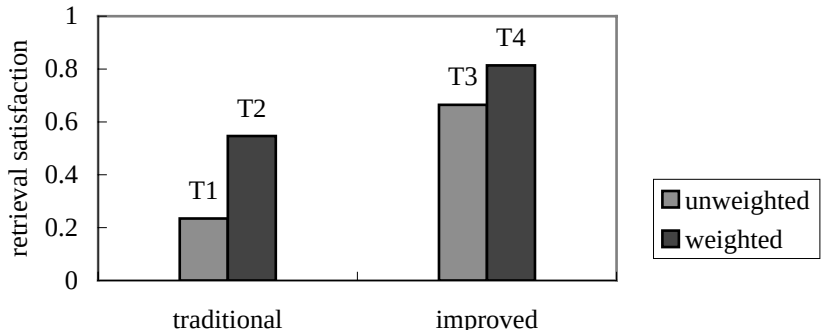


Fig. 14. Traditional model vs. improved model for combination queries.

## 5. CONCLUSIONS

In this paper, a linguistic approach to a Chinese query model has been proposed and investigated. The approach includes Chinese linguistic variables, aggregation model and combination queries. It has been observed that the fuzzy semantics of a general keyword-based query can be easily expressed using the proposed linguistic variables for a casual user without the need to perform numerical weight assignment. Meanwhile, the proposed improved computation for general conjunctive queries does yield higher level of retrieval satisfaction than the traditional computation does based on fuzzy set theory.

As for the Chinese aggregation model, nine quantifiers as well as their corresponding weighting formulas have been proposed. Compared to the Boolean model, the pro-

posed quantifiers not only facilitate multi-criteria information requests but also produce higher levels of retrieval satisfaction.

In principle, the proposed approach can be applied to support natural-language-like queries. However, some technical problems remain to be solved. For example, it is difficult to extract linguistic variables within input query strings. In addition, it may be even more difficult to quantify the relationships between terms (say, '網際網路' and '人格發展') in a query (say, '找出討論到'網際網路對人格發展有重要影響'的文章') using an appropriate weighting function for the relationship (say, '…對…有重要影響'). Future research on linguistic query models could investigate such linguistic variables and their appropriate membership functions.

## ACKNOWLEDGEMENT

This paper was partially supported by the National Science Council, Taiwan, R.O.C., under contract NSC87-2213-E-009-087 (1998).

## REFERENCES

1. G. Salton, G., E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communications of ACM*, Vol. 26, No. 11, 1983, pp. 1022-1036.
2. G. Bordogna, G. Paso, and G. Pasi, "A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation," *Journal of the American Society for Information Science*, Vol. 44, No. 2, 1993, pp. 70-82.
3. L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning (1)," *Information Science*, Vol. 8, No. 3, 1975, pp. 199-249.
4. L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning (2)," *Information Science*, Vol. 8, No. 4, 1975, pp. 301-357.
5. L. A. Zadeh, "Fuzzy set as a basis for a theory of possibility," *Fuzzy Set and Systems*, Vol. 1, No. 1, 1978, pp. 3-28.
6. P. Bosc, M. Galibourg, and G. Hamon, "Fuzzy querying with SQL: Extensions and implementation aspects," *Fuzzy Sets and Systems*, Vol. 28, No. 3, 1988, pp. 333-349.
7. R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Transactions on System, Man, and Cybernetics*, Vol. 18, No. 1, 1988, pp. 183-190.
8. D. H. Kraft, G. Bordogna, and G. Pasi, "An Extended Fuzzy Approach to Generalize Boolean Information Retrieval," *Information Science*, Vol. 2, No. 3, 1994, pp. 119-134.
9. G. Bordogna and G. Pasi, "Linguistic aggregation operators of selection criteria in fuzzy information retrieval," *International Journal of Intelligent System*, Vol. 10, No. 2, 1995, pp. 233-248.
10. M. Kendall, *Rank Correlation Methods*, 4th edition, High Wycombe: Charles Griffin & Company Ltd., London, 1975.

**Tyne Liang, (梁婷)** is currently working as an associate professor in the Department of Computer and Information Science, National Chiao Tung University. Her research interests include Chinese document processing, database design, and information retrieval.

**Tzyy-Chyang Wu (吳子強)** received his master degree in computer science from National Chiao Tung University in 1997. He is currently in military service.