ELSEVIER

Expert Systems with Applications 31 (2006) 397–405

Expert Systems with Applications

www.elsevier.com/locate/eswa

Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques

Hsi-Ching Lin [a], Li-Hui Wang [b], Shyi-Ming Chen [a,*]

[a] *Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, 43, Section 4, Keelung Road, Taipei 106, Taiwan, ROC*

[b] *Department of Finance, Chihlee Institute of Technology, Banciao City, Taipei County, Taiwan, ROC*

Abstract

In document retrieval systems, proper query terms significantly affect the performance of document retrieval systems. The performance of the systems can be improved by using query expansion techniques. In this paper, we present a new method for query expansion based on user relevance feedback techniques for mining additional query terms. According to the user's relevance feedback, the proposed query expansion method calculates the degrees of importance of relevant terms of documents in the document database. The relevant terms have higher degrees of importance may become additional query terms. The proposed method uses fuzzy rules to infer the weights of the additional query terms. Then, the weights of the additional query terms and the weights of the original query terms are used to form the new query vector, and we use this new query vector to retrieve documents. The proposed query expansion method increases the precision rates and the recall rates of information retrieval systems for dealing with document retrieval. It gets a higher average recall rate and a higher average precision rate than the method presented in Chang, Y. C., Chen, S. M., & Liau, C. J. (2003). A new query expansion method based on fuzzy rules. *Proceedings of the Seventh Joint Conference on AI, Fuzzy System, and Grey System*, Taipei, Taiwan, Republic of China.

Keywords: Document retrieval; Fuzzy rules; Query terms; Query expansion; User relevance feedback

1. Introduction

Query expansion is one of the important research topics of information retrieval systems. In order to improve the performance of information retrieval systems, some query expansion techniques have been proposed (Billerbeck, Scholer, Williams, & Zobel, 2003), (Berardi, Lapi, Leo, Malerba, Marinelli, & Scioscia, 2004), (Chang et al., 2003), (Chen, Yu, Furuse, & Ohbo, 2001), (Cooper & Byrd, 1998), (Cui, Wen, Nie, & Ma, 2002), (Jin, Zhao, & Xu, 2003), (Kim, Kim, & Kim, 2001), (Latiri, Elloumi, Chevallet, & Jaoua, 2003a,b), (Li & Agrawal, 2000), (Lin, Wang, & Chen, 2005), (Martin-Bautista, Sanches, Chamorro-Martinez, Serrano, & Vila, 2004), (Nakauchi, Ishikawa, Morikawa, & Aoyama, 2003), (Safar and Kefi 2003), (Stojanovic, 2004), (Takagi & Tajima, 2001), (Wei, Bressan, & Ooi, 2000), (Wang, Lin, & Chen, 2005), (Xu & Croft, 1996). Billerbeck et al. (2003) presented a method for query expansion using associated queries. Berardi et al. (2004) used association rules to mine query expansion terms and presented how to filter off redundant association rules. Chang et al. (2003) presented a query expansion method based on fuzzy rules. Chen et al. (2001) used association rules to discover the degrees of similarity between terms and constructed a hierarchical-tree structure to pick out query expansion terms. Cooper and Byrd (1998) constructed a visual interface with graphical relations between items by lexical neighborhoods for prompted query refinement. Cui et al. (2002) presented a method for probabilistic query expansion using query logs. Jin et al. (2003) presented a method for query expansion based on the term similarity tree model. Kim et al. (2001) presented a method for query term expansion and reweighting using the term co-occurrence similarity and fuzzy inference techniques. Latiri et al. (2003) considered the relationship between terms and documents as a fuzzy binary relation, based on the closure of the extended fuzzy Galois connection, and used fuzzy association rules to find out real-correlated terms as query expansion terms. Li and Agrawal (2000) used multi-granularity indexing and query processing for supporting the web query expansion. Lin et al. (2005) presented a method for mining additional query terms for query expansion. Martin-Bautista et al. (2004) presented a method to mine web documents for finding additional query terms.

* Corresponding author. Tel.: +886 2 27376417; fax: +886 2 27301081.
E-mail address: smchen@et.ntust.edu.tw (S.-M. Chen).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.078

Nakauchi et al. (2003) created thesaurus and relationships of terms for query expansion. Safar and Kefi (2003) presented a query expansion method based on the domain ontology and the lattice structure. Stojanovic (2004) used a conceptual schema to query neighbourhood for query expansion. Takagi and Tajima (2001) presented a method for query expansion using conceptual fuzzy sets for search engines. It calculates the degrees of similarity between terms to construct a hierarchical-tree structure and lets terms with higher degrees of similarity be expansion terms of the structure. Wei et al. (2000) presented a method to mine term association rules for automatic global query expansion. Xu and Croft (1996) used the local analysis and the global analysis of documents for query expansion, respectively.

In this paper, we present a new method for query expansion based on user relevance feedback techniques (Baeza-Yates & Ribeiro-Neto, 1999) for mining additional query terms. The proposed query expansion method according to the user's relevance feedback calculates the degrees of importance of relevant terms to find additional query terms. It uses fuzzy rules to infer the weights of additional query terms. Then, the weights of the additional query terms and the weights of the original query terms are used to form a new query vector, which is used to retrieve documents to increase the precision rates and the recall rates of information retrieval systems. The proposed method gets a higher average recall rate and a higher average precision rate than the method presented in Chang et al. (2003).

The rest of this paper is organized as follows. In Section 2, we briefly review Chang–Chen–Liau's method for query expansion from (Chang et al., 2003). In Section 3, we present a new method for query expansion for document retrieval by mining additional query terms. In Section 4, we show the experimental results. The conclusions are discussed in Section 5.

2. A review of Chang–Chen–Liau's method for query expansion based on fuzzy rules

Chang et al. (2003) used fuzzy rules to deal with query expansion based on user relevance feedback techniques, where the 'relevant frequency' (RF) and the 'inverse document frequency' (IDF) (Baeza-Yates & Ribeiro-Neto, 1999), (Salton, 1971) of each relevant term t_i will affect the relevance degree (RD) of each relevant term t_i. The relevant frequency (RF) of each relevant term t_i is defined as follows:

$$\mathrm{RF}_i = \frac{m_i}{M}, \tag{1}$$

where m_i denotes the number of retrieved relevant documents containing relevant term t_i and M denotes the number of relevant documents. The inverse document frequency (IDF) of each relevant term t_i is defined as follows:

$$\mathrm{IDF}_i = \log_{10}\frac{N}{n_i}, \tag{2}$$

where N denotes the number of documents in the document database and n_i denotes the number of documents having term t_i. The membership functions of the linguistic variables RF, IDF and RD are shown in Figs. 1–3, respectively.

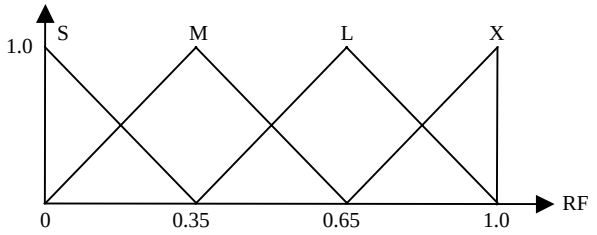

Fig. 1. Membership functions of the linguistic variable RF (Chang et al., 2003).

Then, Chang et al. used 16 fuzzy rules shown in Table 1 and the center of gravity method (Sugeno, 1985) to infer the value r_i of the relevance degree (RD) for each relevant term t_i.

The weights W_{it} of the new query terms in the new user's query vector $\overline{Q_{\mathrm{new}}}$ and the weights W_{iq} of the original query terms in the original user's query vector $\overline{Q_{\mathrm{old}}}$ are defined as follows:

$$W_{it} = \frac{\sum_{k=1}^{m_i} w_{ik}}{m_i}, \tag{3}$$

$$W_{iq} = \frac{\sum_{k=1}^{m_i} w_{ik}}{m_i}, \tag{4}$$

where w_{ik} denotes the weight of relevant term t_i in relevant document d_k, $w_{ik}\in[0,1]$, m_i denotes the number of relevant documents, $w_k\in[0,1]$, and $w_{iq}\in[0,1]$. The degree of similarity $S(\bar{Q},\overline{d_k})$ between a user's query vector $\bar{Q}$ and a document vector $\overline{d_k}$ is calculated as follows:

$$S(\bar{Q},\overline{d_k}) = \frac{\sum_{i=1}^{s}(1-|w_{iq}-w_{ik}|)r_i}{\sum_{i=1}^{s} r_i}, \tag{5}$$

where $S(\bar{Q},\overline{d_k})\in[0,1]$, r_i denotes the relevant degree of query term t_i, w_{iq} denotes the weight of term t_i in the user's query Q,

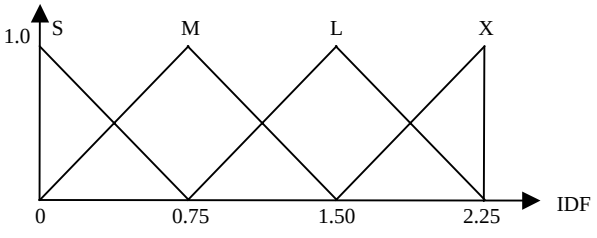

Fig. 2. Membership functions of the linguistic variable IDF (Chang et al., 2003).

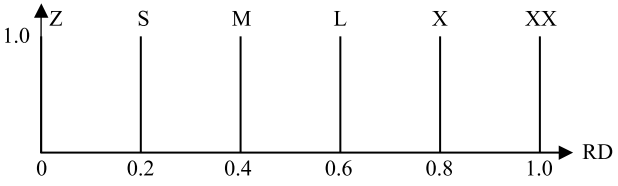

Fig. 3. Membership functions of the linguistic variable RD (Chang et al., 2003).

Table 1
The matrix of sixteen fuzzy rules

IDF RD RF	S	M	L	X
S	Z	Z	S	M
M	Z	S	M	L
L	Z	M	L	X
X	Z	L	X	XX

$w_{iq} \in [0,1]$, $\bar{Q} = <w_{1q}, w_{2q}, \ldots, w_{sq}>$, $1 \le i \le s$, w_{ik} denotes the weight of term t_i in document d_k, $w_{ik} \in [0,1]$, $\overline{d_k} = <w_{1k}, w_{2k}, \ldots, w_{sk}>$, $1 \le i \le s$ and s denotes the number of terms that not only appear in the user's query Q but also appear in document d_k The degree of similarity $DS(\overline{Q*}, \overline{d_k})$ of document d_k with respect to the expanded user's query vector $\overline{Q*}$ can be calculated as follows:

$$DS(\overline{Q*}, \overline{d_k}) = (1 - r_{\text{avg}}) \times S(\overline{Q_{\text{old}}}, \overline{d_k}) + r_{\text{avg}} \times S(\overline{Q_{\text{new}}}, \overline{d_k}), \quad (6)$$

where $\overline{Q*}$ denotes the expanded user's query vector and is formed by the vectors $\overline{Q_{\text{old}}}$ and $\overline{Q_{\text{new}}}$ (i.e., $\overline{Q*}$ is the concatenation of the vectors $\overline{Q_{\text{old}}}$ and $\overline{Q_{\text{new}}}$), $\text{DS}(\overline{Q*}, \overline{d_k}) \in [0,1]$, and r_{avg} denotes the average value of the weights of the terms in the relevant degree vector $\overline{R_{\text{new}}} = <r_1, r_2, \ldots, r_h>$ of the new query term set, which is defined as follows:

$$r_{\text{avg}} = \frac{\sum_{i=1}^{h} r_i}{h}, \quad (7)$$

where $r_{\text{avg}} \in [0,1]$, r_i denotes the relevant degree of query term t_i with respect to the relevant documents of the user's relevance feedback, and h denotes the number of elements in the relevant degree vector $\overline{R_{\text{new}}} = <r_1, r_2, \ldots, r_h>$

In the following, we review Chang–Chen–Liau's algorithm for query expansion from (Chang et al., 2003) as follows:

Step 1. Get relevant terms in the relevant documents of the user's relevance feedback, and calculate the relevant frequency RF_i and the inverse document frequency IDF_i of each relevant term t_i based on formulas (1) and (2), respectively, where $1 \le i \le n$.

Step 2. Based on the membership functions of the linguistic variables RF and IDF, use the 16 fuzzy rules and the center of gravity method to infer the value r_i of the relevance degree of each relevant term t_i, where $\text{r}_\text{i} \in [0,1]$ and $1 \le i \le n$.

Step 3. Based on the relevant degrees of relevant terms, get the top h terms which have the largest relevant degrees and are different from the original user's query terms, and use formulas (3) and (4) to calculate the weights of the original user's query terms and the weights of the new user's query terms, respectively.

Step 4. Use formula (5) to calculate the degree of similarity $S(\overline{Q_{\text{old}}}, \overline{d_k})$ between the document descriptor vector $\overline{d_k}$ and the original user's query vector $\overline{Q_{\text{new}}}$ and use formula (5) to calculate the degree of similarity $S(\overline{Q_{\text{new}}}, \overline{d_k})$ between the document descriptor vector $\overline{d_k}$ and the new user's query vector $\overline{Q_{\text{new}}}$, respectively. Then, use formula (6) to calculate the degree of similarity $S(\overline{Q*}, \overline{d_k})$ between the document descriptor vector $\overline{d_k}$ and the expanded user's query vector $\overline{Q*}$, where $\overline{Q*}$ is formed by concatenating the vectors $\overline{Q_{\text{old}}}$ and $\overline{Q_{\text{new}}}$.

Step 5. Based on the degrees of similarity of the documents with respect to the expanded user's query vector $\overline{Q*}$, rank the documents in the document database in a descending sequence and retrieve the top l documents which have the largest degrees of similarity with respect to the expanded user's query vector $\overline{Q*}$ for user's browsing, where $1 \le l \le p$.

3. A new query expansion method for document retrieval based on user relevance feedback techniques

In the following, we use the vector space model (Baeza-Yates & Ribeiro-Neto, 1999), (Salton, 1971) to represent the user's query and the documents. The inverse document frequency IDF_i (Baeza-Yates & Ribeiro-Neto, 1999) of term t_i is defined as follows:

$$\text{IDF}_i = \log_{10} \frac{N}{n_i},$$

where IDF_i denotes the inverse document frequency of term t_i, N denotes the number of documents in the document database, and n_i denotes the number of documents having term t_i. The weight w_{ik} of term t_i in document d_k and the weight w_{iq} of term t_i in the user's query Q are calculated as follows (Salton, 1971):

$$w_{ik} = \frac{tf_{ik}}{\max_j tf_{jk}} \times \text{IDF}_i, \quad (8)$$

$$w_{iq} = \left(0.5 + 0.5 \times \frac{tf_{iq}}{\max_j tf_{jq}}\right) \text{IDF}_i, \quad (9)$$

where tf_{ik} denotes the occurrence frequency of term t_i in document d_k and ft_{iq} denotes the occurrence frequency of term t_i in the user's query Q. The calculation of the degree of similarity $S(\bar{Q}, \overline{d_k})$ between the user's query vector $\bar{Q}$ and the document vector $\overline{d_k}$ is as follows (Horng et al., 2003):

$$S(\bar{Q}, \overline{d_k}) = \frac{\sum_{i=1}^{s} 1 - |w_{iq} - w_{ik}|}{s}, \quad (10)$$

where $S(\bar{Q}, \overline{d_k}) \in [0,1]$, $\bar{Q} = <w_{1q}, w_{2q}, \ldots, w_{sq}>$, w_{iq} denotes the weight of term t_i in the user's query Q, $w_{iq} \in [0,1]$, $1 \le l \le s$, $\overline{d_k} = <w_{1k}, w_{2k}, \ldots, w_{sk}>$, w_{ik} denotes the weight of term t_i in document d_k, $w_{ik} \in [0,1]$, $1 \le l \le s$, and s denotes the number of terms appearing in the user's query Q and document d_k, simultaneously. In this paper, the system translates the original query terms into term weights based on formula (2) and formula (9), and then these weights form a query vector $\bar{Q}$. Each document d_k in the document database is represented by a

document vector $\overline{d_k}$. Based on formula, (10) the system calculates the degree of similarity between the query vector $\bar{Q}$ and each document vector $\overline{d_k}$. Then, the system ranks the documents according to their degrees of similarity with respect to the user's query from the largest to the smallest and lets the user browse the top h documents, where the value of h is specified by the user and $h \geq 1$. Then, the user marks each relevant document as relevant or irrelevant as user relevance feedback for local analysis. The system considers each term appearing in any relevant document from the user relevance feedback as a relevant term. The weight of each relevant term in each relevant document is calculated using formula (8). The average weight W_{avg} of each relevant term t_i is calculated as follows:

$$W_{\text{avg}_i} = \frac{\sum_{k=1}^{m} w_{ik}}{m}, \tag{11}$$

where w_{ik} denotes the weight of relevant term t_i in relevant document d_k $w_{ik} \in [0,1]$, m denotes the number of relevant documents, and $w_{\text{avg}_i} \in [0, 1]$.

The system finds important relevant terms as additional query term candidates according to the degrees of importance of relevant terms. The 'degree of support' (Lin et al., 2005) of a relevant term t_i, denoted as 'Support(t_i)', is defined as follows:

$$\text{Support}(t_i) = \frac{F_{r_i}}{m}, \tag{12}$$

where F_{r_i} denotes the frequency of relevant term t_i appearing in relevant documents, m denotes the number of relevant documents, and Support(t_i) $\in [0,1]$. From our experiment, a relevant document with a larger value of support(t_i) to be chosen as an additional query term is not always a guarantee to improve the performance of document retrieval systems, and it may even decrease the performance of document retrieval systems (Lin et al., 2005). Thus, we let F_{ir_i} be the frequency of relevant term t_i appearing in irrelevant documents. Then, we can get 'the purity frequency of support' F_{purity_i} of relevant term t_i by subtracting F_{ir_i} from F_{r_i}, shown as follows (Lin et al., 2005):

$$F_{\text{purity}_i} = F_{r_i} - F_{ir_i}, \tag{13}$$

where $1 \leq i \leq n$ and n denotes the number of relevant terms. We also define the 'degree of confidence' of a relevant term t_i, denoted as 'Confidence(t_i)', as follows:

$$\text{Confidence}(t_i) = \frac{F^*_{r_j}}{m}, \tag{14}$$

where $F^*_{r_i}$ denotes the frequency of relevant term t_i and the previous query terms appearing in relevant documents simultaneously and m denotes the number of relevant documents. In our experiment, we also found that a relevant document with a higher value of Confidence(t_i) to be chosen as an additional query term is not always a guarantee to improve the performance of document retrieval systems, and it may even decrease the performance of document retrieval systems. Thus, we let $F^*_{ir_i}$ be the frequency of relevant term t_i and the previous query terms appearing in irrelevant documents simultaneously. Then, we can get the 'purity frequency of confidence' $F^*_{\text{purity}_i}$ of relevant term t_i by subtracting $F^*_{ir_i}$ from $F^*_{r_i}$, shown as follows:

$$F^*_{\text{purity}_i} = F^*_{r_i} - F^*_{ir_i}, \tag{15}$$

Then, we define the 'degree of importance' of relevant term t_i, denoted as Importance$_i$, as follows:

$$\text{Importance}_i = \left(\frac{F_{\text{purity}_i} - \min_{i=1}^{n} F_{\text{purity}_i} + 1}{0.5 + \text{Log}_{10}\left(\frac{M}{F_{r_i}}\right)} \right) \times \log_{10} \left(\frac{\left(F^*_{\text{purity}_i} - \min_{i=1}^{n} F^*_{\text{purity}_i} + 1 \right)^2}{F^*_{r_i}} \right), \tag{16}$$

where M denotes the number of documents including relevant term t_i in the document database; the value 0.5 is used to avoid the denominator to be 0; the square of '$F^*_{\text{purity}_i} - \min_{i=1}^{n} F^*_{\text{purity}_i} + 1$' can strengthen the character of the 'modified purity frequency of confidence'. Then, the system ranks relevant terms according to the degrees of importance of the relevant terms from the largest to the smallest. Because 'the purity frequency of support' F_{purity_i} of relevant term t_i shown in formula (16) might be a negative value, we let '$F^*_{\text{purity}_i} - \min_{i=1}^{n} F^*_{\text{purity}_i} + 1$' be the 'modified purity frequency of support' of relevant term t_i to let its value be larger than zero. We also let '$F^*_{\text{purity}_i} - \min_{i=1}^{n} F^*_{\text{purity}_i} + 1$' be the 'modified purity frequency of confidence' (Lin et al., 2005) of relevant term t_i.

Then, the system uses fuzzy rules to infer the weight of each additional query term. Let the 'combined purity frequency' CPF$_i$ of a term t_i be defined by:

$$\text{CPF}_i = \frac{F_{\text{purity}_i}}{F_{r_i}} \times \frac{F^*_{\text{purity}_i}}{F^*_{r_i}}. \tag{17}$$

The membership functions of the linguistic variables CPF, W_{arg} (i.e., the average weight of a term) and W_{expand} (i.e., the linguistic weight of an additional query term) are shown in Figs. 4–6, respectively.

In this paper, we use the following 25 fuzzy rules to infer the weight of an additional query term t_i:

Rule 1: IF CPF is S AND W_{avg} is S THEN W_{expand} is Z,
Rule 2: IF CPF is M AND W_{avg} is S THEN W_{expand} is Z,
Rule 3: IF CPF is L AND W_{avg} is S THEN W_{expand} is Z,

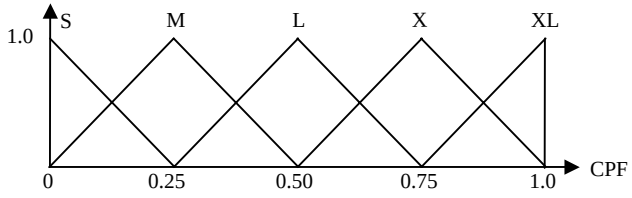

Fig. 4. Membership functions of the linguistic variable CPF.

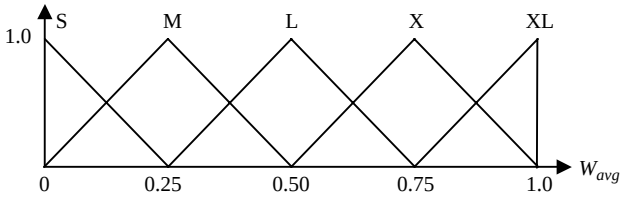

Fig. 5. Membership functions of the linguistic variable W_{avg}.

Rule 4: IF CPF is X AND W_{avg} is S THEN W_{expand} is S,
Rule 5: IF CPF is XL AND W_{avg} is S THEN W_{expand} is S,
Rule 6: IF CPF is S AND W_{avg} is M THEN W_{expand} is S,
Rule 7: IF CPF is M AND W_{avg} is M THEN W_{expand} is S,
Rule 8: IF CPF is L AND W_{avg} is M THEN W_{expand} is S,
Rule 9: IF CPF is X AND W_{avg} is M THEN W_{expand} is M,
Rule 10: IF CPF is XL AND W_{avg} is M THEN W_{expand} is M,
Rule 11: IF CPF is S AND W_{avg} is L THEN W_{expand} is M,
Rule 12: IF CPF is M AND W_{avg} is L THEN W_{expand} is M,
Rule 13: IF CPF is L AND W_{avg} is L THEN W_{expand} is M,
Rule 14: IF CPF is X AND W_{avg} is L THEN W_{expand} is L,
Rule 15: IF CPF is XL AND W_{avg} is L THEN W_{expand} is L,
Rule 16: IF CPF is S AND W_{avg} is X THEN W_{expand} is L,
Rule 17: IF CPF is M AND W_{avg} is X THEN W_{expand} is L,
Rule 18: IF CPF is L AND W_{avg} is X THEN W_{expand} is X,
Rule 19: IF CPF is X AND W_{avg} is X THEN W_{expand} is X,
Rule 20: IF CPF is XL AND W_{avg} is X THEN W_{expand} is X,
Rule 21: IF CPF is S AND W_{avg} is XL THEN W_{expand} is X,
Rule 22: IF CPF is M AND W_{avg} is XL THEN W_{expand} is X,
Rule 23: IF CPF is L AND W_{avg} is XL THEN W_{expand} is XL,
Rule 24: IF CPF is X AND W_{avg} is XL THEN W_{expand} is XL,
Rule 25: IF CPF is XL AND W_{avg} is XL THEN W_{expand} is XL.

Based on the above 25 fuzzy rules, the system can infer the weight of an additional query term using the center of gravity method (Sugeno, 1985) to defuzzify a fuzzy set for deriving the weight of an additional query term. In the following, we use an example to illustrate how to use fuzzy rules to infer the weight of an additional query term. Assume that there is a relevant term t_i to be used as an additional query term, whose $F_{r_i} = 6$, $F_{\text{purity}_i} = 5$, $F^*_{r_i} = 3$, $F^*_{\text{purity}_i} = 1$ and $W_{\text{avg}_i} = 0.43$, then based on Figs. 1–3, we can get $\text{CPF}_i = 0.27$, $W_{\text{avg}_i} = 0.43$, $\mu_M(\text{CPF}) = 0.92$, $\mu_L(\text{CPF}) = 0.08$, $\mu_M(W_{\text{avg}_i}) = 0.28$, and $\mu_L(W_{\text{avg}_i}) = 0.72$. Then, we can infer the weight W_{expand} of the additional query term by using Rule 7, Rule 8, Rule 12, Rule 13 and the center of gravity method (Sugeno, 1985), shown as follows:

$$W_{\text{expand}} = \frac{0.28 \times 0.2 + 0.72 \times 0.4 + 0.08 \times 0.2 + 0.08 \times 0.4}{0.28 + 0.72 + 0.08 + 0.08} = 0.33.$$

That is, the weight of the additional query term t_i is 0.33.

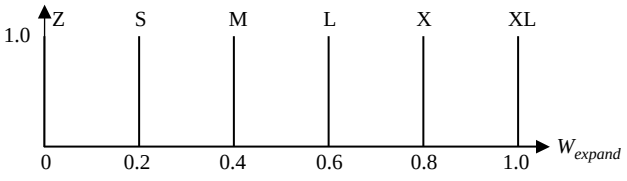

Fig. 6. Membership functions of the linguistic variable W_{expand}.

The proposed query expansion method is now presented as follows:

Step 1: After the user submits the original query terms, translate the original query terms into query term weights based on formula (2) and formula (9), and these query term weights are used to form the query vector. Then, the system translates the terms in each document into weights based on formula (2) and formula (8) to form a document vector.
Step 2: Based on formula (10), calculate the degrees of similarity between the query vector and each document vector, respectively. Rank the documents in the document database based on their degrees of similarity from the largest to the smallest. Retrieve and display the top h documents for the user to browse, where the value of h is defined by the user and $h \geq 1$.
Step 3: Let the user can mark each retrieved document as a relevant document or an irrelevant document for the user relevance feedback for dealing with the process of local analysis.
Step 4: Let each term appearing in relevant documents be a relevant term, except the stopwords. Based on formulas (11), (13), (15) and (16), calculate the values of F_{r_i}, F_{purity_i}, $F^*_{r_i}$, $F^*_{\text{purity}_i}$, W_{avg_i} and Importance_i of each relevant term t_i, respectively. If the relevant term t_i has the largest value of Importance_i, then let it be an additional query term.
Step 5: Use the 25 fuzzy rules shown in Table 1 to infer the weight of the additional query term. The additional query term and the existing query terms are used to form a new set of query terms, then the weight of the additional query term and the weights of the existing query terms are used to form a new query vector.
Step 6: Based on formula (10), calculate the degree of similarity between the new query vector and each document vector, respectively. Rank the documents according to the degrees of similarity from the largest to the smallest. Retrieve and display the top h documents for the user to browse, where the value of h is defined by the user and $h \geq 1$.

Table 2
Query terms of the ten user's queries

User's queries	Query terms
Q_1	Artificial and intelligence
Q_2	Face and recognition
Q_3	Fault and tolerance
Q_4	Graph and theory
Q_5	Image and processing
Q_6	Image and restoration
Q_7	Machine and learning
Q_8	Multimedia and database
Q_9	Relational and database
Q_{10}	Speech and recognition

Step 7. If the number of additional query terms reaches a predefined number determined by the user, then Stop. Otherwise, go to Step 3.

In this paper, the system uses the additional query terms together with the original query terms to retrieve documents for improving the recall rate and the precision rate of an information retrieval system, where the recall rate and the precision rate are defined as follows (Baeza-Yates & Ribeiro-Neto, 1999):

$$\text{Recall Rate} = \frac{|R_a|}{|R|}, \tag{18}$$

$$\text{Precision Rate} = \frac{|R_a|}{|A|}, \tag{19}$$

where $|R_a|$ denotes the number of relevant documents retrieved by the system; $|R|$ denotes the number of relevant documents; $|A|$ denotes the number of documents retrieved by the system.

4. Experimental results

Based on the proposed method, we have implemented a document retrieval system using Delphi Version 5.0 on a Pentium 4 PC. The NSC document database (http://fuzzylab.et.ntust.edu.tw/NSC_Report_Database/520documents.html; Data Source: http://sticnet.stic.gov.tw) obtained from a subset of the collection of the research reports of the National Science Council, Taiwan, Republic of China, is used for the experiment. The NSC document database consists of 520 documents in computer science related fields, divided into 28 categories. The number of documents in each category is between 16 and 25. Each document includes title, index number, Chinese abstract and English abstract. First, the system automatically sifts through the documents and chooses index number and English abstract of each document. Then, it uses the stem method (Baeza-Yates & Ribeiro-Neto, 1999) to filter out the term root of each English term and lets each term root be an index term in the term database. In this paper, we use the query terms of the 10 user queries shown in Table 2 to make the experiment, and we assume that the number of the additional query terms determined by the user is ten. Figs. 7 and 8 show the recall rates and the precision rates of the top 10 retrieved documents with respect to the ten user's queries

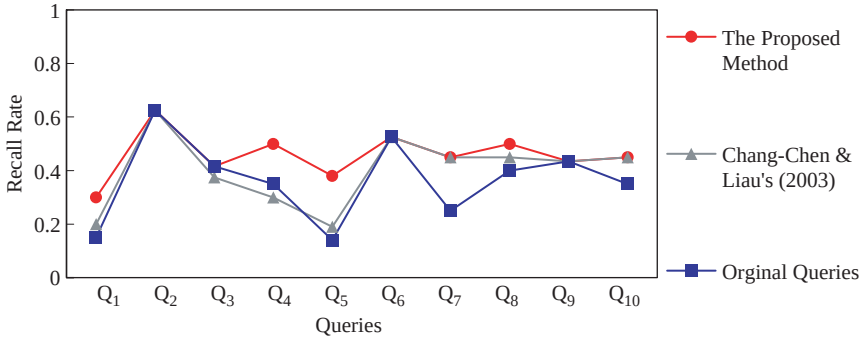

Fig. 7. Recall rates of the top 10 retrieved documents with respect to the ten user's queries for different methods.

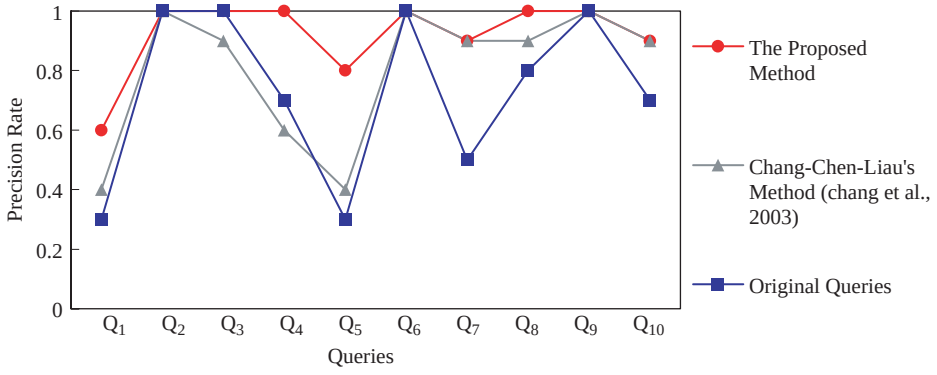

Fig. 8. Precision rates of the top 10 retrieved documents with respect to the ten user's queries for different methods.

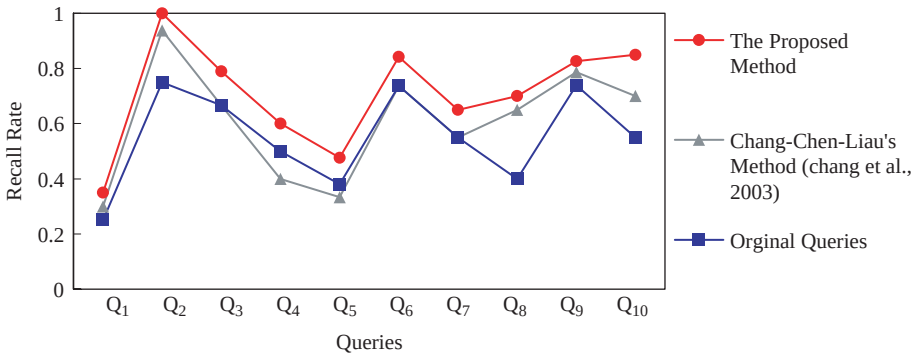

Fig. 9. Recall rates of the top 20 retrieved documents with respect to the ten user's queries for different methods.

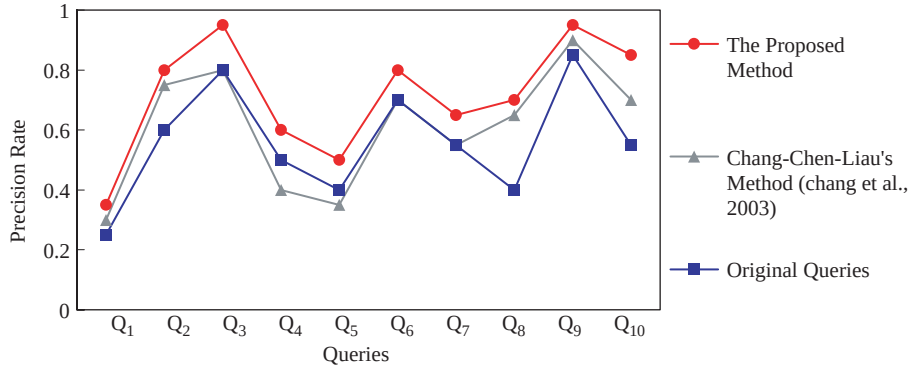

Fig. 10. Precision rates of the top 20 retrieved documents with respect to the ten user's queries for different methods.

Table 3
A comparison of the average recall rate and the average precision rate of the proposed method with Chang–Chen–Liau's method (Chang et al., 2003)

	Top 10 retrieved documents		Top 20 retrieved documents		Top 30 retrieved documents	
	Average recall rate	Average precision rate	Average recall rate	Average precision rate	Average recall rate	Average precision rate
Original user's queries	0.36	0.73	0.55	0.56	0.67	0.46
Chang–Chen–Liau's method (Chang et al., 2003)	0.40	0.81	0.60	0.61	0.70	0.48
The proposed method	0.45	0.92	0.70	0.71	0.75	0.50

shown in Table 2 for different methods, respectively. Figs. 9 and 10 show the recall rates and the precision rates of the top 20 retrieved documents with respect to the ten user's queries shown in Table 3 for different methods, respectively. Figs. 11 and 12 show the improvement of the recall rates and the precision rates of the top 30 retrieved documents with respect to the ten user's queries shown in Table 3, respectively. From Figs. 7–12, we can see that the proposed query expansion

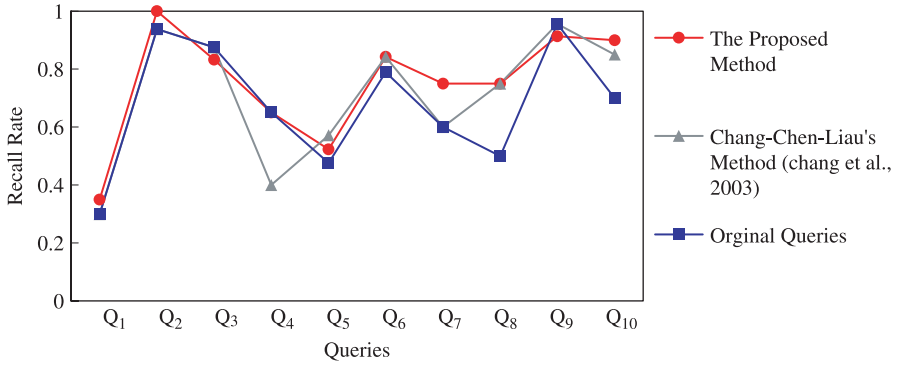

Fig. 11. Recall rates of the top 30 retrieved documents with respect to the ten user's queries for different methods.

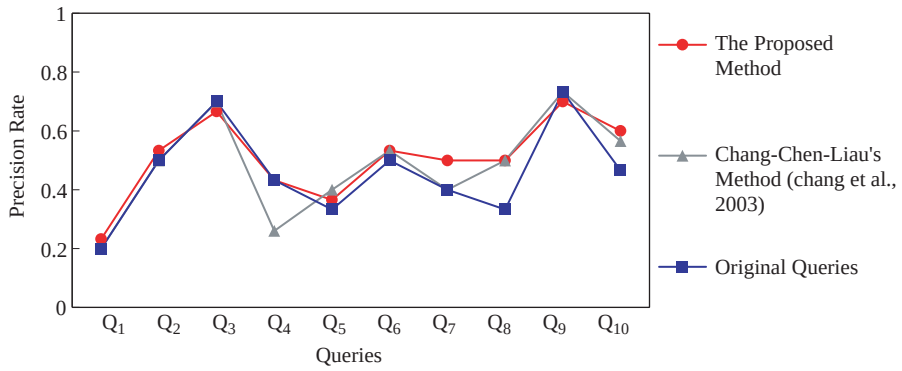

Fig. 12. Precision rates of the top 30 retrieved documents with respect to the ten user's queries for different methods.

method increases the precision rates and the recall rates of the system for dealing with document retrieval. Table 3 makes a comparison of the average recall rate and the average precision rate of the top 10, 20 and 30 retrieved documents of the proposed method with the ones by using Chang–Chen–Liau's method (Chang et al., 2003), respectively. From Table 3, we can see that the proposed method gets a higher average recall rate and a higher average precision rate than the method presented in Chang et al. (2003).

5. Conclusions

In this paper, we have presented a new query expansion method based on the user relevance feedback techniques for document retrieval by mining additional query terms, where the user can mark each retrieved document as a relevant document or an irrelevant document for the user's relevance feedback. The proposed query expansion method uses the vector space model to represent documents and queries. Each term appearing in any relevant document, except the stopwords, is regarded as a relevant term. According to the degrees of importance of the relevant terms, the proposed method calculates the degrees of importance of relevant terms for finding additional query terms, where the user can determine the number of relevant terms to be chosen as additional query terms. The higher the degree of importance of a relevant term, the higher the chance that the relevant term will be chosen as an additional query term. The proposed method uses fuzzy rules to infer the weights of the additional query terms, and then uses these additional query terms together with the original query terms to retrieve documents for improving the performance of information retrieval systems. The proposed query expansion method increases the precision rates and the recall rates of information retrieval systems for dealing with document retrieval. It gets a higher average recall rate and a higher average precision rate than the method presented in Chang et al. (2003).

Acknowledgements

This work was supported in part by the National Science Council, Republic of China, under Grant NSC 94-2213-E-011-003.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Addison Wesley.

Berardi, M., Lapi, M., Leo, P., Malerba, D., Marinelli, C., & Scioscia, G. (2004). A data mining approach to PubMed query refinement. *Proceedings of the 15th international workshop on database and expert systems applications, Zaragoza, Spain,* pp. 401–405.

Billerbeck, B., Scholer, F., Williams, H. E., & Zobel, J. (2003). Query expansion using associated queries. *Proceedings of the 12th international conference on information and knowledge management, New Orleans, LA,* pp. 2–9.

Chang, Y. C., Chen, S. M., & Liau, C. J. (2003). A new query expansion method based on fuzzy rules. *Proceedings of the seventh joint conference on AI, Fuzzy system, and Grey system, Taipei, Taiwan, Republic of China*.

Chen, H., Yu, J. X., Furuse, K., & Ohbo, N. (2001). Support IR query refinement by partial keyword set. *Proceedings of the second international conference on web information systems engineering, Singapore*, Vol. 1, pp. 245–253.

Cooper, J. W., & Byrd, R. J. (1998). OBIWAN—a visual interface for prompted query refinement. *Proceedings of the 31st Hawaii international conference on system sciences, Hawaii*, Vol. 2, pp. 277–285.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002). Probabilistic query expansion using query logs. *Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii,* pp. 325–332.

Horng, Y. J., Chen, S. M., & Lee, C. H. (2003). A new fuzzy information retrieval method based on document terms reweighting techniques. *International Journal of Information and Management Sciences*, *14*(4), 63–82.

Jin, Q., Zhao, J., & Xu, B. (2003). Query expansion based on term similarity tree model. *Proceedings of the 2003 international conference on natural language processing and knowledge engineering, Beijing, China,* pp. 400–406.

Kim, B. M., Kim, J. Y., & Kim, J. (2001). Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference. *Proceedings of the joint ninth IFSA world congress and 20th NAFIPS international conference, Vancouver, Canada*, Vol. 2, pp. 715–720.

Latiri, C. C., Elloumi, S., Chevallet, J. P., & Jaoua, A. (2003). Extension of fuzzy Galois connection for information retrieval using a fuzzy quantifier. *Proceedings of the 2003 ACS/IEEE international conference on computer systems and applications, Tunis, Tunisia*.

Latiri, C. C., Yahia, S. B., Chevallet, J. P., & Jaoua, A. (2003). Query expansion using fuzzy association rules between terms. *Proceedings of the 2003 fourth JIM international conference on knowledge discovery and discrete mathematics, Mets, France*.

Li, W. S., & Agrawal, D. (2000). Supporting web query expansion efficiently using multi-granularity indexing and query processing. *Journal of Data and Knowledge Engineering*, *35*(3), 239–257.

Lin, H. C., Wang, L. H., & Chen, S. M. (2005). A new query expansion method for document retrieval by mining additional query terms. *Proceedings of the 2005 international conference on business and information, Hong Kong, China*.

Martin-Bautista, M. J., Sanches, D., Chamorro-Martinez, J., Serrano, J. M., & Vila, M. A. (2004). Mining web documents to find additional query terms using fuzzy association rules. *Fuzzy Sets and Systems*, *148*(1), 85–104.

Nakauchi, K., Ishikawa, Y., Morikawa, H., & Aoyama, T. (2003). Peer-to-peer keyword search using keyword relationship. *Proceedings of the third IEEE/ACM international symposium on cluster computing and the grid, Tokyo, Japan,* pp. 359–366.

Safar, B., & Kefi, H. (2003). Domain ontology and Galois lattice structure for query refinement. *Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, California,* pp. 597–601.

Salton, G. (1971). *The smart retrieval system—experiments in automatic document processing*. New Jersey: Prentice Hall.

Stojanovic, N. (2004). On using query neighbourhood for better navigation through a product catalog: SMART approach. *Proceedings of the 2004 IEEE international conference on e-Technology, e-Commerce and e-Service, Taipei, Taiwan, Republic of China,* pp. 405–412.

Sugeno, M. (1985). An introductory survey of fuzzy control. *Information Sciences*, *36*(1), 59–83.

Takagi, T., & Tajima, M. (2001). Query expansion using conceptual fuzzy sets for search engine. *Proceedings of the 10th IEEE international conference on fuzzy systems, Melbourne, Australia,* pp. 1303–1308.

Wang, L. H., Lin, H. C., & Chen, S. M. (2005). A new method for query expansion based on uses relevance feedbach techniques. *Proceedings of the Sixth International Symposium on Advanced Intelligent Systems*, Neosn Korea, pp. 679–684.

Wei, J., Bressan, S., & Ooi, B. C. (2000). Mining term association rules for automatic global query expansion: Methodology and preliminary results. *Proceedings of the first international conference on web information systems engineering, Hong Kong, China*, Vol. 1, pp. 366–373.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, Zurich, Switzerland,* pp. 4–11.