Intelligent information filtering using rough sets

Tithiwat Ratanapakdee*, Ouen Pinngern**

*Research Center for Communications and Information Technology (ReCCIT), Department of Computer Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand.
(Tel: +66-2-737-3000 ext 3334; E-mail: s3061619@kmitl.ac.th)
**Research Center for Communications and Information Technology (ReCCIT), Department of Computer Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand.
(Tel: +66-2-737-3000 ext 3334; E-mail: kpouen@kmitl.ac.th)

Abstract: This paper proposes a model for information filtering (IF) on the Web. The user information need is described into two levels in this model: profiles on category level, and Boolean queries on document level. To efficiently estimate the relevance between the user information need and documents by fuzzy, the user information need is treated as a rough set on the space of documents. The rough set decision theory is used to classify the new documents according to the user information need. In return for this, the new documents are divided into three parts: positive region, boundary region, and negative region. We modified user profile by the user's relevance feedback and discerning words in the documents. In experimental we compared the results of three methods, firstly is to search documents that are not passed the filtering system. Second, search documents that passed the filtering system. Lastly, search documents after modified user profile. The result from using these techniques can obtain higher precision.

Keywords: Information Filtering, Information Retrieval, Rough sets, User profile

1. INTRODUCTION

The World Wide Web, with its large collection of documents, is a storehouse of information for any user. Search engines help users locate information. But these search engines usually return a huge list of url's which are ordered according to general relevance computation function. Most of the users find a large proportion of these documents to be irrelevant.

For these reasons, many researches in the field try to improve the result from search engine. Phaitoon S, Ouen P [1] presented the method to improve results by using genetic algorithm. Query is encoded into chromosome. The results from search engine are evaluated by fitness functions to find the best chromosome. This chromosome represents the relevant document ranking that will present to users.

In this research, we apply rough set decision theory to classify the retrieval documents according to the user information need. To efficiently estimate the relevance between the user information need and documents by fuzzy [7, 8]. We will modify user profile by the user's relevance feedback and select key-terms in the documents from decision table.

2. THE USER INFORMATION NEED

In this section, we first analyze the general format of information on Web sites, and then we introduce the method of representing user's queries.

In order to manage vast information, people often classify information into several categories. For example, the books in library can be divided into some categories, e.g. Science, Sociology and Sport. Based on this natural way of structuring categories [5], most Web sites use this hierarchy like the example in Figure 1 to store their own information.

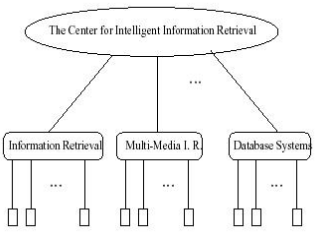

Fig. 1 The structure of storing information of Web sites

In Figure 1, the first level is the home page address (e.g. http://ciir.cs.umass.edu/cdi-bin/w3msql/pulication_database/publications-edit.html) of this Web site, the second level includes all of the categories, and the third level is the documents (all research papers).

From this view point, the concept space might be a hierarchical structure (as in Fig. 2) under a particular theme so that people can easily find certain areas of interest.

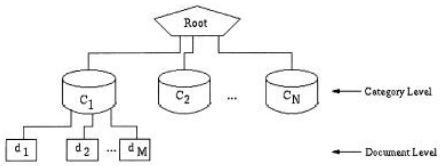

Fig. 2 Concept-based hierarchical structure.

In our research, we allow users to describe their information needs on user concept space. There are two levels in an user concept space: category level and document level. The user information need consists of profiles on the category level and queries on the document level.

Under this strategy, users can firstly describe profile on category level with the weights within a particular field, given within the set:

$$\{CP_{C_1}, CP_{C_2}, CP_{C_3}, \ldots, CP_{C_m}\} \tag{1}$$

where $CP_{C_i} \in [0,1]$ $(1 \le i \le m)$ is the weight that a category C_i is interest to users.

Secondly, the user can provide term sets:

$$\begin{pmatrix} t_{C_{u1},1} & t_{C_{u1},2} & \cdots & t_{C_{u1},m_{u1}} \\ t_{C_{u2},1} & t_{C_{u2},2} & \cdots & t_{C_{u2},m_{u2}} \\ \vdots & \vdots & \vdots & \vdots \\ t_{C_{um},1} & t_{C_{um},2} & \cdots & t_{C_{um},m_{um}} \end{pmatrix} \tag{2}$$

to different interesting categories, where $t_{C_{ui},1}, t_{C_{ui},2}, \ldots, t_{C_{ui},m_{ui}}$ are the terms for category C_i $(1 \le i \le m)$

3. ROUGH SET BASED IF MODEL

The rough set model was proposed by Pawlak [6] in the early 1980s. It is an extension of standard set theory that supports approximations in decision making. The main goal of rough set analysis is to synthesize approximation of concepts from the acquired data [3]. That is partition objects into disjoint equivalence classes. Objects with in the same equivalence class are indistinguishable with regard to the relation.

Let U denote a finite and non-empty set called the universe and let $R \subseteq U \times U$ and R be an equivalence relation the partitions U , that create approximation $apr_R = (U, R)$. Let the partition be denote as $U / R = \{C_1, C_2, ..., C_n\}$, where C_n is an equivalence class of R.

Given an arbitrary set $A \subseteq U$, it may be describe A precisely in the approximation space $apr_R = (U, R)$ by a pair of lower and upper approximation. Then we can interpreted by three ordinary sets:

Reference set:
$A \subseteq U$
Lower approximation:
$\underline{apr}_R(A) = \{x \in U | [x]_R \subseteq A\}$
Upper approximation:
$\overline{apr}_R(A) = \{x \in U | [x]_R \cap A \neq \phi\}$

By definition, $\underline{apr}_R(A) \subseteq A \subseteq \overline{arp}_R(A)$. The pair $(\underline{apr}_R(A), \overline{apr}_R(A))$ is called a rough set with reference set A.

From the above description, we use rough set to derive decision rules [4] that classify incoming documents into three regions: the positive region POS(X_R) in which all documents are relevant, the boundary region BND(X_R) which includes some relevant documents, and the negative region NEG(X_R) in which all documents are irrelevant:

$$POS(X_R) = \underline{apr}(X_R)$$
$$BND(X_R) = \overline{apr}(X_R) - \underline{apr}(X_R)$$
$$NEG(X_R) = D - \overline{apr}(X_R)$$

Figure 3 illustrates the set of relevant documents and the positive, boundary and negative regions.

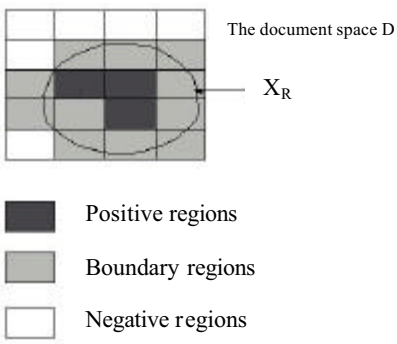

Fig. 3 The set of relevant documents and its POS, BND and NEG regions

In terms of decision-theoretic [9] language, we have a set of states $\Omega = \{X_R, \neg X_R\}$, indicating that a document is relevant or not based on whether it belongs to X_R or $\neg X_R$. A set of action $A_d = \{a_1, a_2, a_3\}$, representing three actions,

$a_1 \equiv$ deciding $d \in POS_A(X_R)$
$a_2 \equiv$ deciding $d \in BND_A(X_R)$
$a_3 \equiv$ deciding $d \in NEG_A(X_R)$

where, d is the current document under consideration.

Now let $\lambda(a_i | d \in X_R)$ denote the loss incurred for taking action a_i when document d in fact belongs to X_R, and let $\lambda(a_i | d \in \neg X_R)$ denote the loss incurred for taking action a_i when document d in fact belongs to $\neg X_R$.

Thus based on the above definitions, we can express the expected loss $L(a_i | d)$, associated with taking the three individual actions as:

$$\begin{aligned} L(a_1|d) &= \lambda_{11} P(X_R, d) + \lambda_{12} P(\neg X_R, d) \\ L(a_2|d) &= \lambda_{21} P(X_R, d) + \lambda_{22} P(\neg X_R, d) \\ L(a_3|d) &= \lambda_{31} P(X_R, d) + \lambda_{32} P(\neg X_R, d) \end{aligned} \quad (3)$$

where $P(X_R, d)$ is the probability that d belongs to X_R, $P(\neg X_R, d)$ is the probability that d does not belongs to $\neg X_R$, $\lambda_{i1} = \lambda(a_i | d \in X_R)$, and $\lambda_{i2} = \lambda(a_i | d \in \neg X_R)$, where $i = 1,2,3$

From Eq. (3), the Bayesian decision procedure leads to the following minimum-risk decision rules (RP), (RB) and (RN):

(RP)
$L(a_1|d) \leq L(a_2|d) \,\&\&\, L(a_1|d) \leq L(a_3|d) \mapsto a_1$
(RB)
$L(a_2|d) \leq L(a_1|d) \,\&\&\, L(a_2|d) \leq L(a_3|d) \mapsto a_2$
(RN)
$L(a_3|d) \leq L(a_1|d) \,\&\&\, L(a_3|d) \leq L(a_2|d) \mapsto a_3$

Consider a special kind of loss functions with $\lambda_{11} = 0$ $0 < \lambda_{21} < 1$, $\lambda_{31} = 1$ and $\lambda_{12} = 1$ $0 < \lambda_{22} < 1$ $\lambda_{32} = 0$. The loss of classifying document d belonging to X_R into the positive region $POS_A(X_R)$ is zero; the loss of classifying document d belonging to X_R into the negative region $NEG_A(X_R)$ is 1, the biggest loss incurred; and the loss of classifying document d belonging to X_R into the boundary region $BND_A(X_R)$ is strictly less than 1 and strictly greater than zero. We obtain the reverse order of losses by classifying document d that does not belonging to X_R. For these type of loss functions, the minimum-risk decision rules (RP), (RB) and (RN) can be written as:

(RP) Deciding $d \in POS(X_R)$ if

$$P(X_R, d) > \max\left\{1, \frac{1 - \lambda_{22}}{\lambda_{21}}\right\};$$

(RB) Deciding $d \in BND(X_R)$ if

$$\frac{\lambda_{22}}{1-\lambda_{21}} \le P(X_R d) \le \frac{1-\lambda_{22}}{\lambda_{21}};$$

(RN) Deciding $d \in NEG(X_R)$ if

$$P(X_R, d) < \min\left\{1, \frac{\lambda_{22}}{1-\lambda_{21}}\right\}.$$

If we use 1 to represent the biggest loss, then we can consider $\lambda_{21} + \lambda_{22} \le 1$, which implies

$$\frac{\lambda_{22}}{(1-\lambda_{21})} \le 1 \le \frac{1-\lambda_{22}}{\lambda_{21}}$$

Under these assumptions, (RP), (RB) and (RN) can be simplified into:

(RP) Deciding $d \in POS(X_R)$

if $P(X_R, d) > \frac{1-\lambda_{22}}{\lambda_{21}}$

(RB) Deciding $d \in BND(X_R)$

if $\frac{\lambda_{22}}{1-\lambda_{21}} \le P(X_R d) \le \frac{1-\lambda_{22}}{\lambda_{21}}$ (4)

(RN) Deciding $d \in NEG(X_R)$

if $P(X_R, d) < \frac{\lambda_{22}}{1-\lambda_{21}}$

The process of how a classification is made on a particular document begins by using the user profile and Boolean queries. Based on the user information need and the document under consideration, the process decides whether that document belongs to the positive, boundary or negative region.

In order to use rough set base decision rules, the problem is how to estimate the value of $P(X_R, d)$. We use fuzzy to do this estimation by following formula:

$$P(X_R, d) = \sum \mu_Q(t) + \sum \left(CP_{C_i} \times \max(\mu_P(t_C), \mu_{PD}(t_C))\right) \quad (5)$$

in which:

- $\mu_Q(t)$ is member value of each terms in query Q belonging to document d which can be define as: $\mu_Q(t)$ = $\frac{w_{k,j}}{\max w_j}$; where $w_{k,j}$ is the weight of term k in query Q belonging to document d_j, and $\max w_j$ is the maximum weight of term in query Q belonging to document d_j;
- $\mu_P(t_C)$ is member value of each terms in term sets $t_{C_{ui}}$ of user profile which can be define as: $\mu_P(t_C)$ = $\frac{iff_i}{\max iff}$; where iff_i is the influence factor of term $t_{C_{ui}, m_{ui}}$, and $\max iff$ is the maximum influence factor of term $t_{C_{ui}}$;
- $\mu_{PD}(t_C)$ is member value of each terms in term sets $t_{C_{ui}}$ belonging to document d which can be define as: $\mu_{PD}(t_C)$ = $\frac{w_{i,j}}{\max w_j}$; where $w_{i,j}$ is the weight of term $t_{C_{ui}, m_{ui}}$ in user profile belonging to document d_j.

4. MODIFICATION OF USER PROFILE WITH MOST DISCERNING WORDS

From user's relevant feedback which is rated by the user is converted to a weighted vector of words. The most unique aspect of our system is the introduction of the rough set theoretic concept of discernibility to identify words which help in distinguishing between relevant and irrelevant documents. To calculate the weights of words, we use the HTML source code of the pages. Since each tag like <TITLE>, <B> etc. in an HTML document has a special significance [2]. We have given tag weights in the range of 1-10. The plain words have a weighing factor 1. The weight of a particular word s is then computed as follows:

$$W(s) = \sum_{i=1}^{m} w_i \times n_i \quad (6)$$

where w_i represents the weight of tag i and n_i represents the number of times the word s appears within tag i, and m is the total number of tags considered. The weights are normalized by dividing the weight of each word in a document by the maximum weight of a word in that document. We now take a look at the decision table constructed with the weighted word vectors as shown in Table 1.

Table 1 A decision table D_T for documents

Documents	W_1	W_2	W_3	W_4	Decision
D_1	1.0	0.1	0.9	0.2	1 (good)
D_2	0.5	0.5	0.9	0.75	0 (average)
D_3	0.0	1.0	1.0	0.9	-1 (bad)
D_4	0.2	0.9	0.9	0.9	-1 (bad)

Let us suppose the number of distinct documents in user's relevant feedback is N and the number of distinct words in user's relevant feedback is k. We will now show how the discernibility table for this set is constructed. For each distinct word in the domain, its weights are arranged in ascending order. An interval set P_s is the constructed for the word s, which is defined as

$$P_s = \{[I_0, I_1), [I_1, I_2), \ldots, [I_r, I_{r+1})\},$$
$$\text{where } I_s = I_0 < I_1 < I_2 < \cdots < I_r < I_{r+1} \quad (7)$$

For each interval in the interval set, the mid point of the interval is called a *cut*. Each distinct word s is thus associated with a set of cuts.

$$\{(s, c_1), (s, c_2), (s, c_3), \ldots, (s, c_r)\} \subset A \times R \text{ where } c_i \text{ is the mid point of } [I_{i-1}, I_i] \quad (8)$$

Let D_T^* denote the discernibility table. D_T^* is constructed with help of decision Table 1 and the cuts. D_T^* has one column for each cut induced over D_T, and one row for each

pair of documents (D_i, D_j) where D_i and D_j have different decisions. An entry v_{ij}^k in D_T^* is decided as follows:

$v_{ij}^k = 0$ in D_T^*, if the document pair D_i and D_j have different decisions but the weight of the word k in both the document are on the same side of the cut.

$v_{ij}^k = d_i - d_j$, if the weight of the work k in document i is more than the cut and the weight of the word in document j is less than the cut, and the documents have different decisions d_i and d_j respectively.

Otherwise, $v_{ij}^k = d_j - d_i$.

Finally, we will now analyze D_T^* to get the most discerning words

However, since the original MD-Heuristic algorithm works with a discernibility table in which all decision differences were considered as identical, we have modified this algorithm to find the most discerning words

The steps in the modified MD-Heuristic algorithm followed by us are:

Step 1: Let W denote the set of most discerning words. Initialize W to NULL. Initialize $T = r$, where r is the maximum difference in decision possible.

Step 2: For each entry in D_T^* consider the absolute value of the decision-difference stored there. If none of the absolute values are equal to T, then set $T = T - 1$, if $T = 0$ then stop else go to step 3.

Step 3: Considering the absolute values of decision difference, choose a column with the maximal number of occurrences of Ts.

Step 4: Select the word w^* and cut c^* corresponding to this column. Delete the column from D_T^*. Delete all the rows marked in this column by T. Delete all columns for w^* from D_T^*.

Step 5: If majority of the decision differences for this column are negative, than the world is tagged with a (-) sign to indicate that is a negatively discerning word. Otherwise it is tagged with a positive sign (+) to indicate that it is a positively discerning word.

Step 6: Add the tagged word w^* and cut c^* to W.

Step 7: If there are more rows left, then go to step 2. Else stop.

This algorithm outputs a list of words, which collectively discern all pairs of documents rate by the user.

5. EXPERIMENTAL

To evaluate the performance, we use traditional evaluation methods precision. For an online Web retrieval, however, it is difficult to obtain all relevant documents. The resource of this trial is "http://google.co.th" and 650 documents were downloaded which are classified as "Information retrieval and Database systems". For various document cutoffs (i.e. 10, 15, 20, 25) the precision has the similar trends, so we select cutoff value 15 to show the evaluation performance. The loss incurred in the trial are $\lambda_{22} = 0.45$ and $\lambda_{21} = 0.50$. The query we use is query Q = {Fuzzy, Genetic} and user profile can be described as the follows:

The interesting categories
(Information Retrieval, 0.9) (Database Systems, 0.6)

The term sets
{retrieval, filtering, intelligent} {query, uncertainty}

In this trial, first, we searched documents that not pass the filtering system by using query Q and term sets in user profile. The system returns 324 documents.

Second, we searched documents that passed the filtering system. The system returns 93 documents into the positive and boundary region and 68 documents into the positive region.

Lastly, we searched documents after modified user profile. The system returns 51 documents into the positive and boundary region and 37 documents into the positive region.

The results of the experiments on precision are displayed in Figs. 4, where the document cutoff is 15.

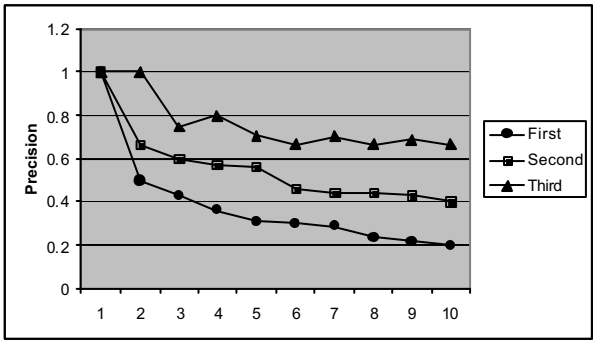

Fig. 4 Precision curves.

This experiment shows that rough set based hierarchical IF models can obtain higher precision.

6. CONCLUSIONS

The very difficult problem in the topic of information gathering from the Web relates to the methods used to describe user information needs. The approach allows users to describe their information needs on user concept space rather than on the space of documents. Therefore, the task of IF models is to build the relationships between user concept spaces and the spaces of documents. A rough set base IF model has been viewed the user information needs as an approximate concept over the space of documents. The other problem was to estimate relevant documents, so we introduced fuzzy set to solve this problem with promising results.

REFERENCES

[1] Phaitoon S, Ouen P., Intelligent Web Search Using Genetic Algorithms and User's Profile, iCAiET 2002, pp 522-527.

[2] A Molinari, G Pasi. A Fuzzy Representation of HTML documents for information retrieval systems, IEEE Fuzzy Systems, Vol.1, pp 107-112, Sep 1996

[3] S K Pal, A Skowron. Rough Fuzzy Hybridization A New Trend in Decision-Marking, Springer; 1999
[4] Y.Y.Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, Int. J. Man-Machine Stud. 37 (1992) 793-809.
[5] Y. Li and C. Zhang, Rough set based decision model in information retrieval and filtering, Third World Multiconference on Systemics, Cybernetics and Informatics (SCI'99) and Fifth International Conference on Information Systems Analysis and Synthesis (ISAS'99), **5**, pp. 398-403, Orlando, USA, July 31 to August 4, 1999.
[6] Z. Pawlak , Rough Sets, International Journal of Information and Computer Sciences , Vol.11, pp.341-356, 1982.
[7] Chakraborty, K., Biswas, R. and Nanda,S., Fuzziness in rough sets, Fuzzy Sets and Systems. 110 (2000) 247-251.
[8] L.A. Zadeh, Fuzzy sets. *Inform. and Control* **8** (1965), pp. 338–353.
[9] Yao, Y.Y., Wong, S.K.M. & Lingras, P. A decision theoretic rough set model. In Z.W. Ras, M. Zemankova & M.L. Emrich, Eds. Methodologies for Intelligent Systems, 5, 17-24, New York: North-Holland.