TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.1, January 2014, pp. 844 ~ 851
DOI: http://dx.doi.org/10.11591/telkomnika.v12i1.3541

Information Retrieval System Based on Rough Fuzzy Set

Wang Yuemin
Computer Science Department, Suqian College, Suqian, Jiangsu, 223800, China
e-mail: wym07@163.com

Abstract

Based on rough sets, fuzzy set theory gives a model of information retrieval. "Contains" relationship reflects a match between the set of documents and user queries using fuzzy set theory, and its inclusion degree to achieve the sort of document sets of search results. The use of rough set equivalence relation reflects the correlation between keywords, achieve synonyms retrieve. Compared with the traditional rough set model, the model represented the right weight of the document sets. Through user query mode, it gives the interest of each keyword. It improved information recall and precision.

Keywords*: Rough Set, Fuzzy Set, Information Retrieval*

1. Introduction

Information has become a hub for human strengthen ties between themselves, as important connotation in many areas, and has a significant impact on all aspects of society. Large variety of rich information over the Internet, greatly exceeds the human cognition carrying capacity, mixed with a variety of information, thereby reducing their overall value embodied [1]. How do we dig out from such a large number and a wide variety of information, useful information that we need? Many scholars have been studying the subject over the years. Therefore, the importance of information retrieval is increasing. Rough set theory and fuzzy set theory can look as the extension and expansion of the classical set theory, they can deal with uncertain and imprecise information and knowledge [2]. The two effective combinations will be more widely used, and the greater enhanced its ability to process information.

In the traditional computer information retrieval, several indexing words are extracted from each document and are used to represent the semantic of the original document, thus one achieves the retrieval according to the original semantic. However, indexing terms from each document shows different meaning. Some show the central content of the document, but some indexing words simply reflects the field involved in the documents [3]. We use the fuzzy concept to describe the general term, can well describe the degree of the semantics of each term, and apply rough set theory to it, therefore construct a retrieval method based on rough fuzzy sets. This method can more accurately describe the semantic of the document and also can retrieve the useful information and can sort it by the similarity measure.

Information retrieval is one of the attentions in the field of intelligent information processing; it is committed to the information in accordance with a certain way of organizing and storing, and needed to find specific information. Rough set theory and fuzzy set theory is effectively together, used in the field of information retrieval, to become one of the research directions in the field of intelligent information processing.

2. Traditional Research Methods of the Information Retrieval Model

2.1. Rough Set-Based Information Retrieval Research

Rough Set can effectively deal with uncertain and imprecise information; it is mainly the knowledge reduction in classification ability to maintain the same premise, export decisions or rules [4].

Definition 1 (rough): for concept X, $X \subseteq U$, R is an equivalence relation on U, if X is R-definable, some basic areas of R and collection to express the concept X or R accurately

Received June 17, 2013; Revised August 26, 2013; Accepted September 18, 2013

describe the set of attributes X, or X is R can not be defined. R defines the collection can also be referred to as R accurate set, R can not be defined in a collection called the R-rough set.

Definition 2 (rough set upper and lower approximate sets): introduced two precise set to describe the rough set, that rough set upper approximation set and the lower approximation set. Lower approximation: $R_-(X)=\{x\in U: [X]_R\subseteq X\}$, upper approximation: $R^-(X)=\{x\in U: [X]_R\cap X\neq\varphi\}$.

Lower approximated by a domain U of R or some division belongs to X, the collection of elements contained in these divisions is the lower approximation. Lower approximation is based on knowledge of R and is certainly belong to the set X's U elements collection. On the approximation of a domain U of R or some division X intersection is not empty, the elements contained in these divided collection is an upper approximation, can be seen on the approximation based on knowledge of R, certainly, and may belong to the set X composed of a collection of the U-element in [5].

negR(X)=U—R-(X) is called X's R negative domain, expressed according to the knowledge R certainly that does not belong to the set X's U element composes of a collection.

The difference on the Upper and Lower Approximation $bn_R(X)= R^-(X) - R_-(X)$ is called the boundary of X's R domain. Based on knowledge of R it can not be sure that they belong to the collection X or belong to a collection -X. It shows the collection of the uncertainty caused by the boundary of the domain.

Definition 3 (topological characteristics of rough set)

(1) If the R-(X) ≠φ and R-(X) ≠U, then X is R rough definable. Certain elements in U may be determined in X or-X.

(2) If the R-(X) =φ and R-(X) ≠U, X is R, non-defined, that can not determine whether the element in U belongs to the X, but can be determined whether some elements belonging -X.

(3) If the R-(X) ≠φ and R-(X) =U, then X is R outside and can not be defined, that can not determine whether the U-elements in -X, but can determine whether some elements in U belong X.

(4) If the R-(X) =φ and R-(X) =U, X is R 100 can not be defined, it can not determine whether the elements in U belong X or -X.

Information retrieval based on rough set theory, can be cited with standard word space relation to the formation of the concept of class, thus taking into account the relationship between indexing terms [6]. In fact, the rough set model is extended Boolean model. Let D be a set of documents, the set of documents D cited the words extracted form indexing word set V. For R is an equivalence relation, it can be indexing V classification of the word set. Given a specific document d_i, indexing terms set x_i, $x_i\subseteq V$, rough set theory that d_i can use the approximation space A = (V, R) is defined, or d_i can be approximated by $A^-(x_i)$ and $A_-(x_i)$. Here x_i may be a document, may also query. Can be taken one way to measure the difference of the queries and documents, the closeness degree that is calculated between the query and the document, that document and query the closer, the greater the degree of association between the two. Therefore, it is possible according to the size of the degree of similarity to the search results sorted.

2.2. Research on Information Retrieval Based on Fuzzy Set

Definition 4 (fuzzy sets): For a set A on the domain U, so that U exists not absolutely the elements belonging to the set A. There are different degrees of element that belong or do not belong to the set A. The set A is called a fuzzy set.

Definition 5 (fuzzy membership function): On the domain U, a fuzzy set A, u ∈U given a map μ_A:U→[0,1], u |→$\mu_A(u)$ ∈[0,1] ,the mapping μA is called fuzzy set A corresponds to a $\mu_A(u)$ for any u, $\mu_A(u)$ to called u set A membership. Membership and function:

$$M(A)=\sum_{i=1}^{n}\mu_A(x_i) \tag{1}$$

Definition 6 (The inclusion relation between fuzzy sets--Inclusion degree theorem) Dominated membership function relationship: Suppose A and B for the theory on the domain U two fuzzy sets, arbitrary u, u ∈ U, if $\mu_A(u) \le \mu_B(u)$, we call that fuzzy set A is included in the fuzzy set B, referred to as $A\subseteq B$, that is, $A\subseteq B$ if $\mu_A(x) \leq \mu_B(x)$ which is for any x. Contained contributions S(A, B) is as follows:

$$S(A,B)=1-\frac{\sum_{i}\max(0,\mu_{\mathrm{A}}(\mathrm{x_i})\text{-}\mu_{\mathrm{B}}(\mathrm{x_i}))}{M(A)} \tag{2}$$

When $\mu_A(x) \leq \mu_B(x)$, S(A, B) = 1, i.e. for all of x, A is entirely contained in the B. On the contrary, when B is the empty set, or for all x, $x \in A$ and $x \notin B$, S(A, B) = 0, because the empty set is unable to contain any collection. Between these two extremes, the size of the inclusion degree: $0\leq$ S(A,B)≤ 1.

Information Retrieval mainly includes three aspects: how to represent the user's query, the set of documents, as well as matching its sort of the user's query and the document set [7-10]. First of all, for the first and second aspects, expressed in the user query and the document set using a single point of law, said: $A=\{\mu_A(u_1)/ u_1,\ \mu_A(u_2)/ u_2,\ \ldots,\ \mu_A(u_n)/ u_n\}$, documentation set u_i is able to represent the meaning of the entire document retrieval set of words extracted from the search term domain U, $\mu_A(u_i)$ is extracted from the search terms belong to set A membership can be understood as per the search word u_i weight. Documentation set in for the explanation of the user query u_i, where $\mu_A(u_i)$ also can be understood as the weight or the search term interest. Secondly, dominated membership function relationship based on the definition of 5 given query search term membership is less than the degree of membership of the search term in the document, then query retrieves the set of words contained in the document set, through it we know, able to find all documents that contain a query set of search terms, this principle is used in the matching process documents and queries [11,12]. That is when we are given a query set of search terms, calculated by including the degree theorem included in the document the extent of the retrieved documents are sorted according to the size of the inclusion degree.

3. Information Retrieval Model Based on Rough Fuzzy Set

The combination of rough and fuzzy sets is able to deal effectively with complex uncertain. The main idea of rough fuzzy sets is approximate calculation.

Definition 7: Let U be the domain of a given domain U, a fuzzy set A, and R is an equivalence relation on U, R is divided $U/R=\{U_1,U_2,\ldots,U_n\}$ in R, fuzzy set A on lower approximation:

$$\mu_{R^{-}(A)}(U_i)=\sup\{\mu_A(u):u\in U_i\} \tag{3}$$

$$\mu_{R\text{-}(A)}(U_i)=\inf\{\mu_A(u):u\in U_i\} \tag{4}$$

Where, sup {} denotes upper bound, inf{} denotes the greatest lower bound. We can see from the formula: the elements of the same equivalence class of fuzzy sets A have rough fuzzy sets membership on R. μR-(A)(Ui) refers to degree of membership that the universe U's u may belong to a fuzzy set A. $\mu_{R\text{-}(A)}(U_i)$ denotes u belongs to at least the degree of membership of the fuzzy set A. The $\mu_{R^{-}(A)}(U_i)$ and the $\mu_{R\text{-}(A)}(U_i)$, when A is an ordinary collection u's membership degree is either 0, or is 1, then $R^{-}(A)$, $R_{-}(A)$ degenerate into classic rough set's upper and lower approximation.

Information retrieval model based on rough set relation based Indexing word space of the model is the formation of the concept, taking into account the relationship between indexing terms. Information retrieval model based on fuzzy sets, allowing the weighted index terms, and the search results are sorted. Taking into account the advantages of these two methods, both combined, form the basic idea of this paper, rough fuzzy set-based information retrieval model, it not only reflects the relationship between indexing terms, but also capable of indexing terms weighting.

For this method, users' queries and a representation of the document set is still expressed with a single point of law: $A=\{\mu_A(u_1)/ u_1,\ \mu_A(u_2)/ u_2,\ \ldots,\ \mu_A(u_n)/ u_n\}$, where, ui is to represent the meaning of the entire document set of search terms, μA(ui) belongs to the set A membership and is extracted from the search terms, the search term domain U. Matching process for the user query and the document set comes from dominated membership function relationship given by Definition 5. Able to retrieve a synonym for the retrieval process, synonymous equivalence class ideas into Definition 6 in this rough fuzzy sets on approximate synonym information to the user query and the document set, and then take contain a degree of thought, in order to achieve the retrieval of synonyms.

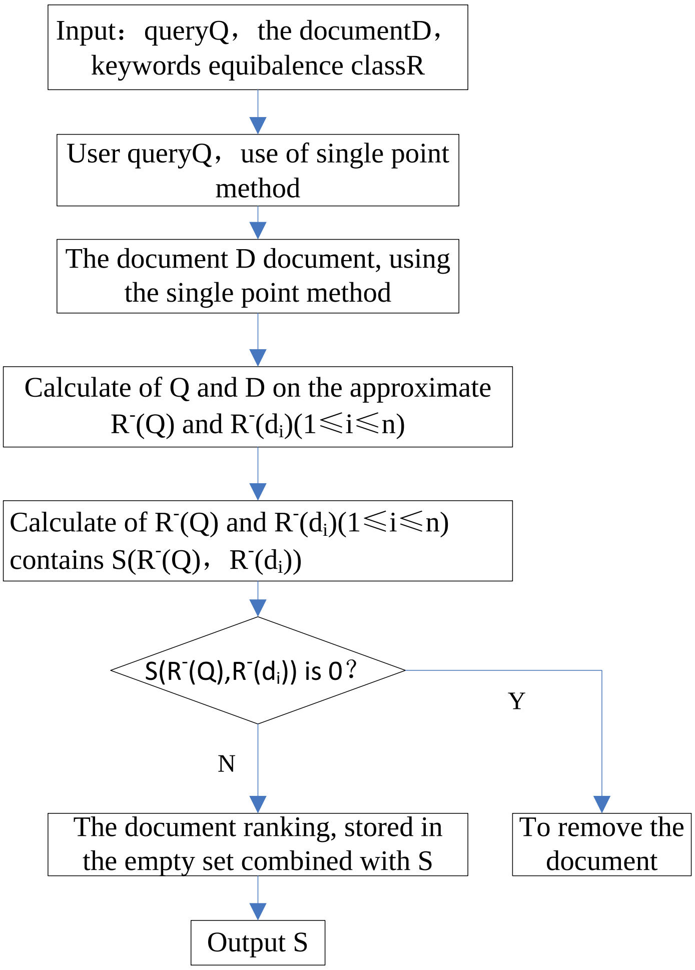
Input：queryQ，the documentD，keywords equibalence classR
User queryQ，use of single point method
The document D document, using the single point method
Calculate of Q and D on the approximate R⁻(Q) and R⁻(dᵢ)(1≤i≤n)
Calculate of R⁻(Q) and R⁻(dᵢ)(1≤i≤n) contains S(R⁻(Q)，R⁻(dᵢ))
S(R⁻(Q),R⁻(dᵢ)) is 0?
Y
N
The document ranking, stored in the empty set combined with S
To remove the document
Output S

Figure 1. Flowchart of algorithm based on rough fuzzy set

By the proposed rough fuzzy set-based information retrieval, you can get the basic algorithm as follows:

Input: user query Q, the document set D={d_1，d_2，d_3，…，d_n}, keywords equivalence class R.

Output: in accordance with the user queries to re-sort the query result the documentation set S.

Step 1: the user inputs weights of each search term query Q or degree of interest, it uses single-point method to get the user's query Q representation.

Step 2: for $d_i(1\leq i\leq n)$ document word in D, take a certain indexing words on behalf of the original document, with the collection of these indexing terms, and WIDF function of indexing terms weighted, using the single point method representation of all documents.

Step 3: initialize empty set S is used to store the query results.

Step 4: Calculate the approximate R-(Q) and R-(di)($1\leq i\leq n$) of query Q and set of documents D

Step 5: Calculate S(R-(Q)，R-(di)) of the formula R-(Q) and R-(di)(1≤i≤n).

Step 6: If S(R-(Q), $R^-(d_i)$) is 0, then the removal of the document; if not 0, then the comparison of the document set containing contributions S(R^-(Q),$R^-(d_i)$), arranged in descending order of these set of documents, and placed in S.

Step 7: Output S

The flow chart of the algorithm is shown in Figure 1

4. Information Retrieval Comparative Case Analysis

4.1. Case Analysis of Information Retrieval Based on Fuzzy Set

Provide information such as Table 1. Indexing is included in the word set V in 9 indexing words, respectively, $t_1,t_2,\ldots,t_9$ to represent. Given the instance analysis method based on rough fuzzy sets. Including 11 documents in the document set D is assumed, respectively, d_1, d_2, ..., d_{11}, said user query Q={$0.5/t_1,0.3/t_2,0.2/t_5,0.5/t_9$} given indexing word set V = { $t_1,t_2,\ldots,t_9$}, the keywords in V synonymous equivalence class is divided, after division the equivalence classes are: T_1 = { t_1, t_3, t_4}, T = { t_2, t_5}, T_3 = {t_6, t_7, t_8}, T_4 = {t_9}, the document set information with table 1.

Table 1. The weight of the document set D

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
d_1	0.6	0.5	0	0	0.7	0	0	0	0.7
d_2	0.7	0.6	0	0	0.8	0	0	0	0.9
d_3	0.4	0.1	0	0	0.1	0	0	0	0.3
d_4	0.3	0.1	0	0	0.1	0	0	0	0.1
d_5	0.6	0.5	0	0.6	0	0.3	0	0	0
d_6	0.5	0.4	0	0.2	0	0.1	0	0	0
d_7	0.3	0.1	0	0.6	0	0.3	0	0	0
d_8	0.3	0.1	0	0.9	0	0.5	0	0	0
d_9	0.2	0.1	0	0.6	0	0.3	0	0	0
d_{10}	0	0	0.1	0	0	0.4	0.5	0.2	0
d_{11}	0	0	0	0.5	0	0.2	0.6	0.3	0

Given in Table 1, Documentation Set includes all: completely contains query Q, section contains Q ,also does not contain Q. The results are shown in Table 2.

Table 2. Query results based on fuzzy sets

S(Q,d_i)	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
Q	1	1	9/15	6/15	8/15	8/15	4/15	4/15	1/5	0	0

Document results sorted by the table to get: d_1, d_2 → d_3 → d_5, d_6 → d_4 → d_7, d_8 → d_9. The important retrieval based on fuzzy set method, the introduction of the importance of the search term, and thus be able to distinguish between a search term and sort according to its search results, the investigation was 81.8%, and the retrieved documents are also useful document.

4.2. Case Analysis of Information Retrieval Based on Rough Fuzzy Set

The inclusion degree obtained by Q document, according to the case study, shows that this approach is consistent with the actual situation, but did not take into account the relationship between keywords, such as synonyms, so in the improved method we add synonym considerations, i.e., when to retrieve a keyword, it is necessary to consider their synonymous keywords, and this can be done by setting the equivalence relationship R synonymy, according to the equivalence relation R, on the query Q using the formula (3) is obtained:

R^-(Q)= {$0.5/t_1,0.3/t_2$, $0.5/t_3$, $0.5/t_4,0.3/t_5,0.5/t_9$}, in this model, the user query will use the R^-(Q) instead of the original query Q, For the keywords t1, t1 and t3 is synonym, t1 and t3 mutually independent search terms to retrieve a document does not reflect the synonym relationship. However, the improved method, t_1 and t_3 are synonyms, obtained $R^-(Q)$ by the equation (3) can be seen, making the original t_3, t_3 of the user query endowed with t_1 synonym relation R the same weight, making the document contains keywords t_3, i.e. retrieved by the keywords in the user query, it is possible to retrieve the document contains the keywords of the keyword synonyms.

For d_i, $R^-(d_l)$= {0.6/t_1, 0.7/t_2, 0.6/t_3, 0.6/t_4, 0.7/t_5, 0.7/t_9} (5)

Using Equation (2) to get:

$$S(R^-(Q),R^-(d_1)) = 1 - \frac{\sum_i \max(0, \mu_{R^-(Q)}(t_i) - \mu_{R^-(d_i)}(t_i))}{M(R^-(Q))} \tag{6}$$

And so, the results are shown in Table 3.

Table 3. Query results based on rough fuzzy sets

S	*d₁*	*d₂*	*d₃*	*d₄*	*d₅*	*d₆*	*d₇*	*d₈*	*d₉*	*d₁₀*	*d₁₁*
Q	*1*	*1*	17/26	20/26	21/26	21/26	11/26	11/26	17/26	3/26	15/26

The document results sequence: d_1, d_2, d_3, d_4→d_5, d_6, d_7, d_8, d_9→d_{10}, d_{11} is obtained by the table 3. It must be pointed for document d_{11}, information retrieval method based on fuzzy sets; the document will not be retrieved because the d_{11} does not contain the key words contained in the Q.

Now, this paper compares based on rough fuzzy sets method with based on rough set to prove its feasibility. Method based on rough sets keywords using Boolean, Either 1 or 0, the proposed method Keywords not using Boolean said, and therefore should be proposed method Keywords conducted to quantify the unified model based on rough sets.

The information given in Table 1 documentation set, select a threshold λ (0 <λ≤1) Boolean quantify this information, first select λ = 0.5, as shown in Table 4, which based on rough set method to retrieve.

Table 4. Boolean documentation set table

	t₁	**t₂**	**t₃**	**t₄**	**t₅**	**t₆**	**t₇**	**t₈**	**t₉**
d₁	1	1	0	0	1	0	0	0	1
d₂	1	1	0	0	1	0	0	0	1
d₃	0	0	0	0	0	0	0	0	0
d₄	0	0	0	0	0	0	0	0	0
d₅	1	1	0	1	0	0	0	0	0
d₆	1	0	0	0	0	0	0	0	0
d₇	0	0	0	1	0	0	0	0	0
d₈	0	0	0	1	0	1	0	0	0
d₉	0	0	0	1	0	0	0	0	0
d₁₀	0	0	0	0	0	0	1	0	0
d₁₁	0	0	0	1	0	0	1	0	0

According to Table 4 rough set information retrieval strategies, with the document and query similarity calculation method, Table 5 shows search results based on rough set method.

Table 5. Search results based on rough set method

SIM(Q,dᵢ)	d₁	d₂	d₃	d₄	d₅	d₆	d₇	d₈	d₉	d₁₀	d₁₁
	1	1	0	0	5/6	1/2	1/2	1/3	1/2	0	1/3

Table 5 shows that the retrieved documents: d_1, $d_2 \to d_5 \to d_6$, d_7, $d_9 \to d_8$, d_{11}, recall rate is 72.7%, the rate is 100% precision, the recall rate is not very high, the reason we can see, the threshold λ values affect the search results. λ the greater the value, the document Keywords take "0", the greater the probability that the document does not include the greater the probability of the keyword, so that the retrieved documents is less. With reference to Figure 2 shows, when λ, with the relationship between the number of documents retrieved.

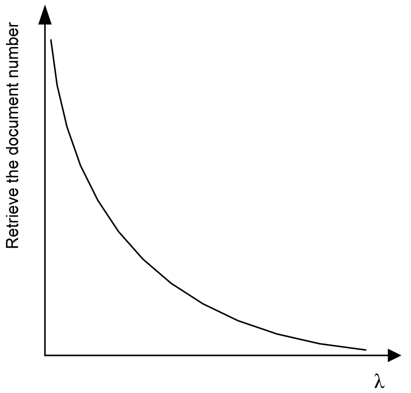

Figure 2. The relationship between λ and the number of retrieved documents

Table 6. Comparison of three methods

	Query results	recall	precision
Method based on fuzzy sets	$d_1 \to d_2 \to d_3 \to d_4 \to d_5 \to d_6 \to d_7 \to d_8 \to d_9$	81.8%	100%
Method based on rough set (λ = 0.5)	$d_1, d_2 \to d_5 \to d_6, d_7, d_9 \to d_8, d_{11}$	72.7%	100%
Method based on Rough fuzzy set	$d_1 \to d_2 \to d_5 \to d_6 \to d_4 \to d_3 \to d_9 \to d_{11} \to d_7 \to d_8 \to d_{10}$	100%	100%

Case analysis results show that the method based on fuzzy sets in the set of users' queries and information is the same set of documents, recall rate of 81.8%, based on the traditional rough when we set λ = 0.5 the recall rate is 72.7%, and when λ to take a very small number of the best, the recall rate is 100%, and this article based on rough fuzzy sets recall rate is 100%, which can be seen the given method advantage.

5. Conclusion

Based on rough fuzzy set theory, given a rough fuzzy set information retrieval model, the model combines the traditional rough set-based methods, as well as the advantages of the method based on fuzzy sets; considers the weight information and keywords relationship. The approximation, synonymous equivalence relation of rough fuzzy set theory, synonymous with information, the user query and the document set to expand provide a foundation to improve the retrieval recall rate. After the analysis and comparison of based on fuzzy set methods and based on rough fuzzy sets method, the paper embodies the advantages where based on rough fuzzy set method.

References

[1] Fu Xuefeng, Liu Chiu cloud, Ming wen. Mutual information-based rough set information retrieval model. *Shandong University of Science and Technology (Natural Science), Journal Publish*. 2006; 41(3): 21-24.

[2] LL Wei, Qiu Tao-rong, Chen Xia. Retrieval based on similarity of Rough relational database. *The Computer Engineering and Design, Journal Publish*. 2007; 28(17): 13-17.

[3] Dai Jun, Wu Chen, Zhang Huan. Rough sets based on the fuzzy tolerance relation decomposition and its applications in information retrieval. Microcomputer information, Journal Publish. 2006; 22(8): 36-39.

[4] Tang Minan, Wang Xiaoming, Yuan Shuang. Site Selection of Mechanical Parking System Based on GIS with AFRARBMI. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(7): 3935-3944.

[5] Sun WC, Chen YMA. Rough set approach for automatic key attributes identification of zero-day polymorphic worms. *Expert Systems with Applications, Journal Publish*. 2009; 36(3): 4672-4679.

[6] Shao XY, Chu XZ, Qiu HB. An expert system using rough set theory for aided conceptual design of ship's engine room automation*, Expert Systems with Applications, Journal Publish*. 2009: 36(2): 3223-3233.

[7] Hu QH, Liu JF, Yu DR. Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems, Journal Publish*. 2008; 21(4): 294-304.

[8] Parthalain NM, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection. *Pattern Recognition, Journal Publish*. 2009; 42(5): 655-667.

[9] Wen Jun-Qin. Design of a Multi-Source Information Collection and Retrieval System. *Journal of Convergence Information Technology, Advanced Institute of Convergence Information Technology*. 2012; 7(3): 292-298.

[10] Zhu Shuxin, Xie Zhonghong, Chen Yuehong. Information Extraction from Research Papers based on Conditional Random Field Model. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(3): 1213-1220

[11] Che-Yu Yang, Shih-Jung Wu. Semantic Web Information Retrieval Based on the Wordnet. *International Journal of Digital Content Technology and its Applications, Advanced Institute of Convergence Information Technology*. 2012; 6(6): 294-302.

[12] Jing Luo, Bo Meng, Xinhui Tu. Enhancing Document Modeling for Information Retrieval Using Wikipedia. *International Journal of Advancements in Computing Technology, Advanced Institute of Convergence Information Technology*. 2012; 4(23): 266-273.