ICRCS at Intent2: Applying Rough Set and Semantic Relevance for Subtopic Mining

Xiao-Qiang Zhou, Yong-Shuai Hou, Xiao-Long Wang, Bo Yuan, Yao-Yun Zhang
Key Laboratory of Network Oriented Intelligent Computation
Department of Computer Science and Technology
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, P.R. China
Xiaoqiang.jeseph@gmail.com houyongshuai@hitsz.edu.cn

ABSTRACT
The target of the subtopic mining subtask of NTCIR-10 Intent-2 Task is to return a ranked list of subtopics. To this end, this paper proposes a method to apply the rough set theory for redundancy reduction in subtopic mined from webpages. Besides, semantic similarity is used for subtopic relevance measure in the re-ranking process, computed with semantic features extracted by NLP tools and semantic dictionary. By using the reduction concept of rough set, we first construct *r*ough *s*et *b*ased *m*odel (RSBM) for subtopic mining. Next, we combine the *r*ough *s*et *t*heory and *s*emantic *r*elevance into a new model (RS&SRM). Evaluation results show the effectiveness of our approach compared with a baseline *f*requency *t*erm *b*ased *m*odel (FTBM). The best performance is achieved by RS&SRM, with I-rec of 0.4046, *D-nDCG* of 0.4413 and *D#-nDCG* of 0.4229 on the subtask of Chinese subtopic mining.

Categories and Subject Descriptors
H.3 [**INFORMATION STROAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms
Algorithms, Documentation, Experimentation

Keywords
Rough Set Theory, Subtopic Reduction, Semantic Relevance, Subtopic Mining

1. INTRODUCTION
When user is searching information with query in the web, there are too few keywords in the query to express user's true intent entirely, furthermore, the keyword of user' query contain a certain extent of ambiguity [1], namely a query may refer to different interpretations and multiple aspects, which are called subtopics [2] of user' query. This kind of query causes that the search system could not recognize the true aspect associated with the query, so the subtopic mining is an important research issue of analyzing true intent of user's to satisfy the need of user .

Conference'10, Month 1–2, 2010, City, State, Country.

Major of subtopic mining methods commonly select the occurrence frequency of term in given relevant data collection as the primary factor during the mining process, such as cluster-based model and query log based model. In order to mining subtopic for reflecting the intent of user's query, we have to consider the need of user from the semantic factors besides of the frequency. There is a state that one term may have high frequency weight in data collection, but the term has no affection in expressing the intent of user' query. This kind of term or subtopic is redundant subtopic from the point of semantics, whereas the terms are core subtopics. This state is called subtopic redundancy state. Another state is that some terms or subtopic appear in pairs, they are semantic interrelated with each other. Only if considering one single term or subtopic as the mining object, the subtopic mining result may miss semantic related subtopics and reflect the intent of user' query incomplete. This state is called semantic related state.

Rough set theory [3], proposed by Z. Pawlak has been applied to in many fields, such as data mining and knowledge discovery [4]. This theory is a new Mathematical framework, which has the ability to deal with fuzzy and uncertainty problem. A fundamental principle of supporting this ability is the concept of attribute reduction [5], one of the most important research issues in recent years. To deal with subtopic redundancy state, in this paper we apply the rough set theory to eliminate the redundant subtopics and extract the core subtopics, because the subtopic is a kind of intent semantic knowledge of one query, which has uncertainty characteristics. By considering these characteristics and applying the basic concepts of rough set, we propose *r*ough *s*et *b*ased *m*odel (RSBM) to mine the subtopics from webpages of user' query.

Semantic relevance computing has been researched and applied for many years. The methods of semantic relevance computing include Edge-based, Static Analysis based, Node-based and so on. The Edge-based method defines the semantic relevance as the path length function [6] between two topic words in the topological structure of the classification system. In order to mine the semantic related subtopics after *r*ough *s*et *b*ased *m*odel (RSBM), in this paper we choose *N*ormalized *G*oogle *D*istance [7] (NGD), which belongs to The Edge-based method, to calculate the semantic relevance between different candidate subtopics. According to this approach we construct *r*ough *s*et & *s*emantic *r*elevance *m*odel (RS&SRM) to discover the semantic related subtopics from candidate subtopics.

The next sections are arranged as follows: section 2 describes the architecture of subtopic mining system and explains the subtopic mining models in detail, we illustrates how to use semantic similarity computing for the subtopic re-ranking in this section.

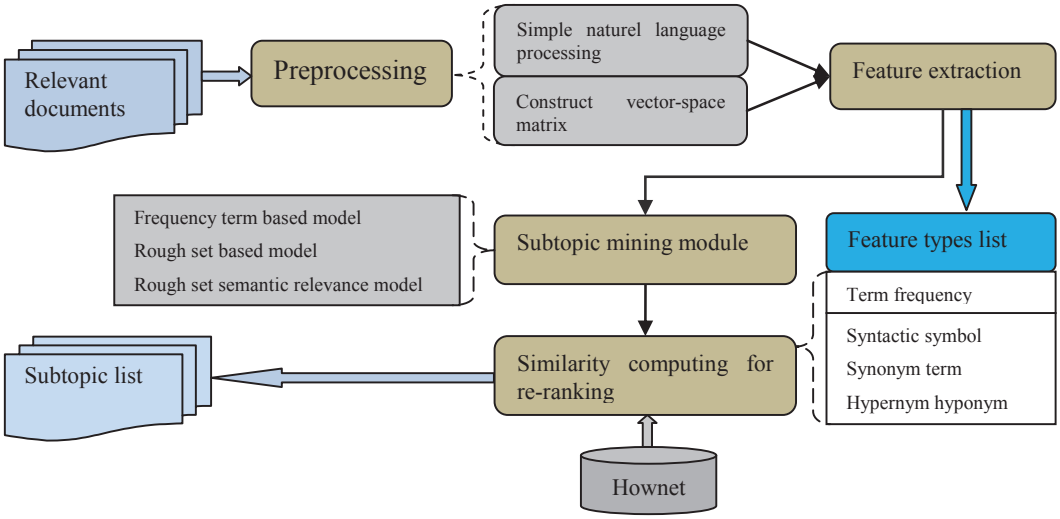

Figure 1. Subtopic mining system architecture of ICRCS

The section 3 presents the experimental results and discussion; and the section 4 concludes the paper.

2. SYSTEM DESCRIPTION

Fig 1 shows the architecture of our subtopic mining system. The modules in this system and implementation details of each module are described as follows.

2.1 Preprocessing Module

In order to mining the subtopics of textual contents which are extracted from relevant document collection of each target query, this module achieves to translate relevant documents into one formal representation which is convenient for subtopic mining. For our mining models, documents are represented using the *v*ector-*s*pace *m*odel, which is also called *f*ull *t*ext *i*ndexing [8]. In the vector space model, a vector is used to represent each item or document in a collection. Each component of the vector reflects a particular word, or term, associated with the given document.

To construct the *v*ector-*s*pace *m*odel for given relevant document collection of each query, this module purifies the relevant document collection of each query and extracts textual content. Secondly, this module uses the LTP [1] tool to finish NLP procedures for document collection of each query, include sentence and word segmentation, POS tagging, syntactic parsing and *n*amed *e*ntity *r*ecognition (NER). At last this module eliminates stop words and some high frequency words. After all above of steps, all remaining words are called candidate terms of each document. In the *v*ector-*s*pace *m*odel, the weight value assigned to each component of the vector reflects the importance of the term in representing the semantics of the document. How to compute the weight of each component will be explained in detail in section 2.3.1.

2.2 Feature Extraction Module

In order to semantic computing during subtopic mining and ranking, according to the vector-space model generated after preprocessing, we extract frequency distribution feature and semantic features of each candidate term in this module. Table 1 shows feature types and concrete descriptions.

Term Frequency Distribution Feature (TF): this type of feature illustrates one term or word distribution state in the relevant document collection of one query, which is equal to the occurrence time in each document. We extract the TF feature from *v*ector-*s*pace *m*odel which constructed in the preprocessing module. The formal representation of the *j*th term is a one-dimensional vector as followed:

$$TF_j = \{\mathrm{tf}_{1j}, \mathrm{tf}_{2j} .. \mathrm{tf}_{ij} ..\} \quad 0 \le i \le N \tag{1}$$

Where the N is equal to the total number of relevant document collection of one query.

Syntactic Symbol Feature (SSF): this feature illustrates the syntactic role which one term or word plays in one sentence. We can extract the SSF feature from the results of syntactic parsing by LTP tool. We know that all symbols of syntactic component are finite set which includes "ADV", "SUB" and other syntactic component symbols. This formal representation also is a one-dimensional vector like TF feature, each value in vector is equal to the occurrence time which the syntactic component symbol of one term appears of finite symbol set in relevant document collection, the vector size is equal to the size of symbol set. The value of vector can be count by parsing process. For example, the occurrence time which the term "战争" plays the role of "SUB" in Webpages is *M times*, so the value corresponding to SUB syntactic component symbol for the term "战争" is *M* in this feature vector. The concrete feature definition of the *j*th term is determined as followed:

$$SSF_j = \{\mathrm{ss}_{j1}, \mathrm{ss}_{j2} .. \mathrm{ss}_{jk} ..\} \quad 0 \le k \le K \tag{2}$$

The K is equal to the total number of syntactic symbol class in the formula 2.

Synonym Term Feature (STF): we extract the feature also from Hownet. The features contain the similar words with term in semantic, so the feature is a word list for term. The formal representation is defined as followed:

[1] http://www.oschina.net/p/hit-ltp.

$$STF_j=\{st_{j1},st_{j2}..st_{ji}..\} \quad 0\le i\le C \tag{3}$$

Where the C is the number of synonym terms of the jth term.

Hypernym Hyponym Feature (HHF): we extract the feature from Hownet which is a Chinese semantic dictionary. The feature illustrates Hypernym-relation and Hyponym-relation between term and words of Hownet,[2] these two kind of relations are built according to the sememe hierarchy in Hownet. This kind of feature contains the level ID of Hypernym, Hyponym and the term self.

The SSF, STF and HHF are semantic features of one term. These features include the semantic information of term and will be used for semantic computing in subtopic mining module and ranking module. TF is distribution feature which reflects the degree of affection between term and relevant document.

Table 1. Description list of term feature sets

Feature	Description
TF	Reflect the degree of affection the document
SSF	Reflect the syntactic component distribution
STF	Contain the similar semantic information
HHF	Contain hierarchic semantic information

2.3 Subtopic Mining

This section introduces three subtopic mining models during mining process, including frequency term based model, rough set based model and rough set & semantic relevance model.

2.3.1 Frequency Term Based Model

For discovering the subtopics from Webpages of one query, the basic method is extracting the most important terms in the document, and the frequency value of one term can reflect the degree which the term will be latent subtopic in one documents. After preprocessing module, for one query, the given relevant documents have been represented in a *v*ector-*s*pace *m*odel. The frequency term based model mainly apply the term-weighting technique which is commonly called *TF-IDF* and used in Information Retrieval. The original *TF-IDF* equation for the weight of term *j* in document *i* is determined as followed:

$$w_{ij}=tf_{ij}*\log N/\mathrm{df}_j \tag{4}$$

Where tf_{ij} is the frequency of occurrence of term *j* in document *i;* $idf=\log N/df_j$ is the inverse document frequency; N is the total number of relevant documents for the query; df_j is the document frequency where term j occurs.

But in the long document, the frequency of occurrence of some terms can be very high, which produce relatively higher weight for these terms. So we need modified the representation of tf_{ij} to smooth the weight value distribution, the modified calculation equation is determined as followed:

$$w_{ij}=\mu_{ij}*\lfloor \log N-\log df_j+1\rfloor \tag{5}$$

In this formula, the values of μ_{ij} are calculated as followed:

$$\left\{\begin{array}{l}\mu_{ij}=1;tf_{ij}=1\\ \mu_{ij}=1.5;1<tf_{ij}\le\frac{tf_i}{4}\\ \mu_{ij}=2;\frac{tf_i}{4}<tf_{ij}\le\frac{tf_i}{2}\\ \mu_{ij}=2.5;\frac{tf_i}{2}\le tf_{ij}\end{array}\right\} \tag{6}$$

Where tf_i is the total number of terms in document i.

For each term in the vector- space, we calculation the total weight of the term in the given relevant document collection of one query by the following equation:

$$W_j=\sum_{i=1}^{N}w_{ij} \tag{7}$$

Where N is the total number of relevant documents for the query. We rank the term list by the value of W_j, and select the top terms as the subtopics of the query.

2.3.2 Rough Set Based Model

Constructing *r*ough *s*et *b*ased *m*odel (RSBM) mainly is based on the attributes reduction method of the rough set theory. According to the uncertainty characteristics between subtopic and query, *r*ough *s*et *b*ased *m*odel (RSBM) applies the subtopic reduction method to mine the core subtopics, which have a great of affection in expressing the intent of user' query.

As the basic concepts of rough set theory, the information system about the relationship between subtopic and document can be represented as: *RS = {U, S, V, f}*; where *U* is a nonempty finite set of documents, *S* is a nonempty finite set of candidate subtopic, *V* is the set of attribute values, and *f* is an information function which decide the subtopic value of the each document *x* of *U*. The formal expression of *V* and *f* can be determined as followed:

$$V=\bigcup\nolimits_{a\in A}V_a \tag{8}$$

$$f:U\times S\to V \tag{9}$$

With every subset *B* of candidate subtopics set *S*, We associate a binary relation *I(B)*, called B-indiscernibility relation, the concrete defined as:

$$I(B)=\{(x,y)\in U\times U|f(x,b)=f(y,b),\forall b\in B\} \tag{10}$$

The *I(B)* is an equivalence relation and We donate the equivalence class of *I(B)* including the document *x* as $[x]_B$. And for any document subset *X* of *U*, the *B-lower* and *B-upper* approximation of *X* in *S* respectively are determined as followed:

$$\underline{L}(BX)=\{x\in U|[x]_B\subseteq X\} \tag{11}$$

$$\overline{U}(BX)=\{x\in U|[x]_B\cap X\neq\phi\} \tag{12}$$

If a subtopic *b* of subset *B*, which is subset of candidate subtopic *set S*, is superfluous in subset *B* if *I(B)=I(B-{b})*, otherwise the subtopic *b* is indispensable in subset *B*. The collection of all indispensable subtopics in candidate subtopic *set S* is called the core subtopics set. we say that subset *B* is independent in candidate subtopic *set S* if every subtopic in subset B is indispensable in subset *B*. the subtopic subset *B* of *S* is called a reduction if subset *B* is independent and *I(B)=I(S)*. The subtopics from all the reductions in candidate subtopic *set S* is denoted as

[2] http://www.keenage.com/

the core subtopic of the candidate subtopic set *S*. all the above of illustrated are the concept of the subtopic reduction.

The subtopic reduction process of this model applies the attribute reduction algorithm [9] which is based on attribute frequency to our mining model. This algorithm uses the properties of frequency as the attribute importance to reduce the attribute set and get the core attribute list. In *r*ough *s*et *b*ased *m*odel, we select the occurrence frequency of candidate subtopic in webpages as the most important factor for mining the core subtopic of user' query.

2.3.3 Rough Set & Semantic Relevance Model

After mining the subtopic mining by *r*ough *s*et *b*ased *m*odel (RSBM), we can get the most interrelated subtopics with user' query, but the result list of subtopic only coverage a part of subtopic, some semantic relevance subtopic has been reduced by the reduction algorithm. Furthermore, there are some subtopic pairs like $<S_i,S_j>$, they are relevant in semantic. In *r*ough *s*et *t*heory and *s*emantic *r*elevance *m*odel, we use the NGD [7] method computing the semantic relevance between the different subtopics to make the subtopic result express more complete intent of user' query. The NGD formula [7] is showed as followed:

$$NGD(S_i,S_j)=\frac{\max\{\log f(S_1),f(S_2)\}-f(S_1,S_2)}{\log M-\min\{\log f(S_1),\log f(S_2)\}} \quad (13)$$

Where M is the total number of document collection; the $f(S_1)$ and $f(S_2)$ are respectively frequency in the Webpages; $f(S_1,S_2)$ is the time of co-occurrence in the document set. The value of NGD is in the range of [0,1]. The smaller the value of NGD is, the more the degree of semantic relevance is. By integrating the semantic relevance computing with *r*ough *s*et *b*ased *m*odel (RSBM), we construct *r*ough *s*et and *s*emantic *r*elevance *m*odel (RS&SRM).

2.4 Semantic Similarity Re-ranking

Generally, the ranking of subtopic list is accord to one weigh value from subtopic mining processing. But the subtopic list usually exacts redundant and similar subtopics in semantic, this is one problem in the result; another problem is that some subtopics are not relevant to the query, although with high weight value in the list. In order to solve these two problems and improve the quality of ranking, we carry on Re-ranking to the subtopic list by the semantic similarity computing. During the processing of semantic similarity computing, we apply the four types of term feature which are described in section 2.2 to calculate the similarity value. The concrete similarity equation is determined as followed:

$$SIM(term_p,term_q)=\alpha\mathrm{sim}(\mathrm{TF}_p,\mathrm{TF}_q)+\beta\mathrm{sim}(SSF_p,SSF_q)+\gamma\mathrm{sim}(STF_p,STF_q)+\lambda\mathrm{sim}(\mathrm{HHF}_p,\mathrm{HHF}_q) \quad (14)$$

In this formula, the weights of computing similarity of the different types of feature have the condition as followed and each weight cannot be zero:

$$\alpha+\beta+\gamma+\lambda=1 \quad (15)$$

Where the values of alpha, beta, gamma and lambda are decided by computing the similarity of small sample data set with different values, such as <0.25 0.25 0.25 0.25> , <0.25 0.3 0.3 0.15> and so on. At last, we choose the values of ***<0.3 0.3 0.15 0.25>*** as the final weights for similarity calculation in our submitted runs. The similarity of feature TF reflects the distribution relationship between different terms. The similarity of SSF reflects the syntactic symbol distribution of one term. TF and SSF are all distribution features for one term in different units, text and sentence. By using these distributional features, we can put the similar distributional terms together to rank. The similarity of STF feature reflects same or similar semantic degree between two terms, by using this feature, we can delete the reduction terms with same semantic. By using the HHF feature, we can rank the terms under same or similar classification together.

For the feature vector of TF, SSF and STF, the similarity is calculated by the cosine of the angle between the feature vectors v_p and v_q. The calculation formula is as followed:

$$\mathrm{sim}(v_p,v_q)=\frac{\sum_{k=1}^{t}v_{pk}\cdot v_{qk}}{\sqrt{\sum_{k=1}^{t}(v_{pk}^2)\cdot\sum_{k=1}^{t}(v_{qk}^2)}} \quad (16)$$

And for the similarity of the HHF features HHF_p and HHF_q is computed based on Hownet API.

At last, if the semantic similarity between $term_p$ and $term_q$ is greater with the similarity threshold SV, we merge $term_p$ and $term_q$ into one term, and the corresponding weight in the subtopic list is the sum of $weight_p$ and $weight_q$. In our paper, the value of the similarity threshold SV is 0.75.

3. EXPERIMENTS AND DISCUSSION

In the task of subtopic mining, there are totally 200 test queries, including 100 queries for INTENT-2. We use those queries do different runs to compare the different subtopic mining models of our system.

Table 2 shows the run name list, the corresponding subtopic mining model; Table 3 shows the official result of our subtopic mining system by different mining models. According the mining result, we can know *r*ough *s*et and *s*emantic *r*elevance *m*odel (RS&SRM) has the best performance comparing with other two models. And *f*requency *t*erm *b*ased *m*odel (FTBM) performs better than *r*ough *s*et *b*ased *m*odel (RSBM). The *f*requency *t*erm *b*ased *m*odel (FTBM) is the baseline to mine the subtopic by using the distribution feature of the candidate terms in the given document. This method mainly selects the frequency of the term as the subtopic weight during mining process.

Table 2. Subtopic mining subtask run

Run name	Model name	Description
ICRCS-S-C-1A	FTBM	*F*requency *T*erm *B*ased *M*odel
ICRCS-S-C-2A	RSBM	*R*ough *S*et *B*ased *M*odel
ICRCS-S-C-3A	RS&SRM	*R*ough *s*et & *S*emantic *R*elevance *M*odel

Table 3. Results of Chinese subtopic mining

Run name	I-rec	*D-nDCG*	*D#-nDCG*
ICRCS-S-C-1A	0.3821	0.4219	0.4020
ICRCS-S-C-2A	0.3704	0.4024	0.3864
ICRCS-S-C-3A	0.4046	0.4413	0.4229

According to the result in Table 3, the core subtopics mined by *r*ough *s*et *b*ased *m*odel (RSBM) has lower *I-rec,* because the core subtopics set is just a few part of candidate subtopic set. Besides, the core subtopics can reflect the intent of user' query very well, but the remainder ones may be not relevant to user' query. For example in Table 4, a user query "花样年华" which is a famous movie name, the topmost subtopics in the list are about multiple aspects of the query and reflect some attributes of this movie query to some extent, whereas the ones at the end of list are not associated with "花样年华" query about movie topic, just include the query keywords in the subtopic string. This state may be caused by the excessive reduction during the algorithm of *r*ough *s*et *b*ased *m*odel.

Table 4. Subtopic list of query "花样年华"

Subtopic list
花样年华电视剧
花样年华电影
重庆花样年华
花样年华下载
花样年华插曲
……..
……
花样年华健身
花样年华地产

Because of considering *r*ough *s*et *b*ased *m*odel and semantic relevance computing, the RS&SRM deal with the subtopic with another one as a subtopic pair $<S_i, S_k>$, not just one single subtopic. The RS&SRM can mine semantic related subtopics and put them together in the result list, so the subtopic from this model can express more complete intent of user' query on the semantic aspect and cluster relevant subtopics together to some extent. For example, a user query "日俄战争" which is a famous historic war between Japan and Russia. Table 5 shows semantic related subtopics of this query, topmost subtopics are all semantic attributes of this war such as the time and place; the next two subtopics are about the historic aspect of this war; the last subtopics of subtopic list are about investigative aspects of this war.

Table 5. Semantic related subtopic pairs of query "日俄战场"

Subtopic list
....
日俄战争战场
日俄战争时间
旅顺日俄战争
1904 年日俄战争
....
日俄战争资料
日俄战争史
....
日俄战争研究
日俄战争原因
日俄战争的影响

The *r*ough *s*et and *s*emantic *r*elevance *m*odel (RS&SRM) reflects better performance than FTBM, it proves that applying the rough set and semantic relevance together are have more effectiveness in subtopic mining.

4. CONCLUSIONS

This paper presents the ICRCS system for subtopic mining subtask. We apply the knowledge reduction concept of rough set theory to construct *r*ough *s*et *b*ased *m*odel (RSBM) to extract the core subtopics from relevant webpages of user' query. By incorporating semantic relevance and RSBM, we propose the rough *s*et *t*heory and *s*emantic *r*elevance *m*odel (RS&SRM) to mine semantic related subtopics from the candidate subtopics. For subtopic ranking, our system achieves subtopic re-ranking by computing semantic similarity with semantic features. Evaluation results prove that our models, based on *r*ough *s*et *t*heory and *s*emantic *r*elevance, have better performance comparing with the baseline method *f*requency *t*erm *b*ased *m*odel (FTBM).

Our future work includes changing term-document representation into the *t*opic-*s*pace *m*odel. Besides, we will add clustering processing to classify the candidate subtopics and try to train the similarity weight value and the threshold by machine learning.

5. ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (No. 61272383).

6. REFERENCES

[1] Spa¨rck-Jones, K., Robertson, S. E., and Sanderson, M. 2007. Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum. 41, 2 (Dec. 2007), ACM, New York, NY, USA, 8-17. DOI= http://dl.acm.org/citation.cfm?id=1328964.1328965

[2] Wang, C. J., Lin, Y. W., Tsai, M. F., and Chen, H. H., 2012. Mining subtopics from different aspects for diversifying search results. Information Retrieval. (Dec. 2012). 1-32. DOI=http://link.springer.com/article/10.1007/s10791-012-9215-y

[3] Pawlak, Z. 1995. Rough sets. In Proceedings of the 1995 ACM 23rd annual conference on Computer science (Nashville, USA, February 28 - March 2, 1995). CSC '95. ACM, New York, NY, 262-264. DOI= http://dl.acm.org/citation.cfm?id=277421

[4] Xie, G., Zhang, J., Lai, K., and Yu, L 2008. Variable precision rough set for group decision-making: An application, International Journal of Approximate Reasoning. 49, 2 (Oct. 2008), 331-343. DOI= http://dx.doi.org/10.1016/j.ijar.2007.04.005

[5] Yao, Y.Y., and Zhao, Y. 2008. Attribute reduction in decision-theoretic rough set models, Information Sciences. 178, 1, (Sep. 2008), 3356-3373. DOI= http://dx.doi.org/10.1016/j.ins.2008.05.010

[6] Rada, R., Mili, H., Bicknell, E., et al. 1989. Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics. 19, 1, (Feb. 1989), 17-30.

[7] Cilibrasi, R. L., and Vitanyi, P. M. B. 2007. The Google Similarity Distance. IEEE Trans.on Knowl. and Data Eng. 19, 3 (Mar. 2007), 370–383.

[8] Merkl, D. 1998. Text classification with self-organizing maps: Some lessons learned. Neurocomputing. 21, 1-3, (Nov. 1998), 61-77. DOI= http://dx.doi.org/10.1016/S0925-2312(98)00032-0

[9] Chen, Y. Y., Qiu, J. L., and et al. 2012. A parallel rough set attribute reduction algorithm based on attribute frequency. In Proceedings of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery. (Chongqing, China, May 29-31, 2012). FSKD'2012. IEEE, 211-215.