# Rough Set Theory Approach for Opinion Extraction of the Product from Text

Sumeet V. Vinchurkar
PG Department of Computer Science and Engineering,
G.H. Raisoni College of Engineering, Nagpur.

Smita M. Nirkhi
PG Department of Computer Science and Engineering,
G.H. Raisoni College of Engineering, Nagpur.

## ABSTRACT

In this paper, the fuzzy rough set theory is useful to extract the key sentences and its feature attribute after getting the opinion from the post will be evaluated. Before this operation the pre-processing steps will be discussed for finding the entity and its attribute, on the basis of the output the rough set theory is used for avoiding the ambiguities between the word sense sentences. Fuzzy rough set theory is the main focused of this paper. It generates the result with the help of Fuzzy Rough Set approach and showing the way for reduction of feature.

## Keywords

Fuzzy-rough sets, word sense disambiguation, semantic patterns retrieval, POS Tag, key feature extraction.

## 1. INTRODUCTION

It is now well recognized that the user-generated content (e.g., product reviews, forum discussions and blogs) contains valuable consumer opinions that can be exploited for many applications. There are already many companies that provide opinion mining services. In this process, firstly identifies what entities (e.g., products) each sentence talks about. Most opinion mining researches are based on product reviews [1,2,3,9] because a review usually focuses on a specific product or entity and contains little irrelevant information. However, in forum discussions and blogs, the situation is very different, where the authors often talk about multiple entities [4, 8] (e.g., products), and compare them. This raises two important issues: (1) how to discover the entities that are talked about in a sentence and (2) how to assign entities to each sentence because in many sentences entity names are not explicitly mentioned, but are implied. We term the first problem entity discovery and the second problem entity assignment. Without knowing the entities that a sentence talks about, any opinion mined from the sentence is of little use. For example, if an algorithm finds that a sentence expresses a negative opinion about something, but it cannot determine on what product, then the opinion is meaningless.

The first problem is about the Named Entity Recognition (NER). This will performed with help of some processing steps and sequential pattern matching steps on that we apply the POS tag for getting the entity. After the entity will find we used the Rough set theory on that for removing disambiguates between two attribute and feature reduction.

The rest of this paper is organized as follows, section 2 includes the system work gives the survey about the system, section 3 introduces the Rough Set Theory and

Extraction with Set Theory approach, and section 4 includes Conclusion of this paper.

## 2. PREPOSED ARCHITECTURE

The overall architecture of the system is shown in figure 1. There are four process phases organized as a pipeline in the system.

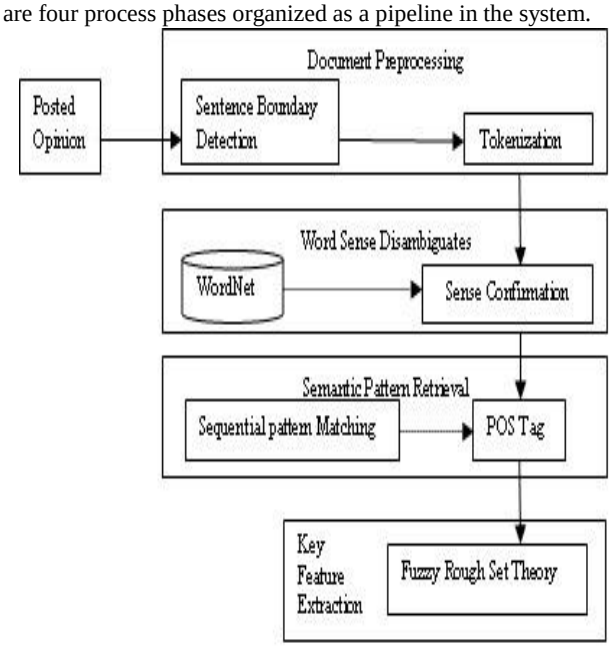


**Figure1: System Architecture**

The functions of these process phases are described in the following subsections.

### 2.1. Document Pre-processing

Currently, input post is of plain text format. By applying some NLP [3,8] steps for making the tokens is as follows.

Statement extraction: firstly, from each post, sentences are individuated, that are parts of text ending with a full stop, comma, question mark, exclamation mark or semicolon. Subsequently, conjunctions are analyzed for dividing sentences into statements, which are parts of text expressing only one meaning.

Anaphora resolution: often, in informal text, subject and predicate could be understood; hence some statements could be incomprehensible. The goal of the anaphora resolution step is to restore a statement by adding understood parts.

Tokenization: in this step, each statement is divided into tokens, which are parts of text bounded by a separator (space, tab or end of line).

Stemming and Lemmatization: in order to reduce the number of different terms, each token is transformed reducing its inflectional forms to a common base form. The main difference among stemming and lemmatization's that the former extract "brutally" the root of a word (e.g. bio is the stem of biology,

biocatalyst and biochemical); while the latter uses a vocabulary (often a lexical ontology) for returning the dictionary form of a word, that is the lemma (e.g. be is the lemma of is, are, was, and so on).

Tagging and Stopwords elimination: some word categories are too common to be useful to distinguish among statements. Hence, in this step articles, prepositions and conjunctions are first recognized and then removed. At the same time, we also remove proper nouns, which usually don't have an affective content.

## 2.2. Word Sense Disambiguation

In automatic text summarization, word sense disambiguation is important and many different approaches have been taken [18]. In this process phase, the Lesk [19] approach is adopted and modified for word sense disambiguation. The Lesk approach assumes that words used in a sentence are collaborative in terms of topic and their dictionary definitions, thus must use some common words in their sense definitions. Based upon this assumption, the nouns and the verbs are first extracted from each sentence together with their senses given in WordNet [23] as the input to the following process for sense disambiguation:

1) For a word to be disambiguated, the process first scores the semantic relatedness between any two senses, one for this word and the other for any other word in the same sentence.
2) Note each sense in WordNet is semantically related with a set of similar senses. The score computed in the previous step reflects the direct relatedness between any two senses. It is like a local link. To fully reflect the relatedness between two senses, their indirect relatedness should be taken into account. The indirect semantic relatedness of two senses is the sum of the pair wise relatedness scores of the two set of similar senses.
3) The final score of each sense for the word is the sum of the score given by Step 1.) And the half value of the score given by Step 2.).
4) Among all the senses of the word, the sense with the highest score is selected as the candidate sense of the word. After all words in a sentence are disambiguated, this phase builds and reports the sense representation for the sentence in terms of WordNet senses to indicate what concept the sentence may cover.

## 2.3. Semantic Patterns Retrieval and Pos Tag for Entity Discovery

These pre-processing steps gives the lexicon and for passed to the POS Tag for the iterative steps. These are excellently defined by Xiaowen Ding, Bing Liu, Lei Zhang in [4]. They are as follows.

*Step 1 – Data preparation for sequential pattern mining* [7, 21]

This step perform two tasks, it first finds all sentences that contain anyone of the seed entities, $e_1$, $e_2$, …, $e_n$ in the dataset, and then generate a sequence for each occurrence of $e_i$ for pattern mining. In order to focus patterns on entities and not generate too many patterns, we use only a window of 5 words before each entity name and 5 words after each entity name. Each word of a seed entity name is replaced with a generic name "ENTITYXYZ". The purpose of using this generic word is to ensure that general patterns about any entities are found. Note that each entity name may consist of more than one word. The part-of-speech (POS) tag [11,12] of each word is also used. In the final sequence each element of the sequence is a pair, POS tag of the word and the word.

Example 1: We have the following sentence with POS tags attached. Here n95 is a phone model (an entity).Hiiiiiiiii/NNP SK/NNP -/: ,/, don't/NN be/VB mad/JJ everyone/NN doesn't/NN have/VBP a/DT n95/CD phone/NN fetish/NN ducky/JJ

The window is (n95 has been replaced with ENTITYXYZ): mad/JJ everyone/NN doesn't/NN have/VBP a/DT ENTITYXYZ /CD phone/NN fetish/NN ducky/JJ

The resulting sequence is :<{ JJ, mad}{NN, everyone}{NN, doesn't}{VBP, have}{DT, a}{CD, ENTITYXYZ}{NN, phone}{NN, fetish} {JJ, ducky}>

*Step 2 – Sequential pattern mining* [4, 6, 21]

Given the set of sequences generated from step 1, a sequential pattern mining algorithm is applied to generate sequential patterns [6]. We use 0.01 as the minimum support. We also require that each pattern must contain {POStag, ENTITYXYZ} and its length to be greater than or equal to 2 for obvious reasons. An example pattern is:<{IN}, {DT}, {NNP, ENTITYXYZ }, {is}>Here "IN", "DT", "NNP" are POS tags which can match any words with that tag, and "is" is a concrete word which can only match this particular word.

*Step 3 – Pattern matching to extract candidate entities* [4,6]

For each sentence in the test dataset, the system matches the generated patterns to extract a set of candidate entities. The patterns are sorted based on their supports. In order not to generate too many spurious candidates, the matching process in a sentence terminates after 5 patterns have been matched. We tried several numbers and find that 5 is a good number with respect to results and efficiency.

Example 2: We have the follow sentence with POS tags attached: The/DT misses/VBZ has/VBZ currently/RB got/VBN a/DT Nokia/NNP 7390/CD at/IN the/DT end/NN of/IN the/DT day,/VBG all/DT she/PRP does/VBZ is/VBZ text/NN and/CC make/VB calls,/NN but/CC the/DT reception/NN is/VBZ terrible,/VBG where/WRB my/PRP$ 6233/CD would/MD get/VB full/JJ bars/NNS hers/PRP would/MD only/RB get/VB 1/CD or/CC 2./CD

The pattern, < {DT}, {NNP, ENTITYXYZ}, {CD}>, will match the sentence segment, a/DT Nokia/NNP 7390/CD, to produce the candidate entity: "Nokia". The pattern, <{DT}, {NNP}, {CD, ENTITYXYZ}, {IN}>, will match the sentence segment, a/DT Nokia/NNP 7390/CD at/IN, to produce the candidate entity: 7390

*Step 4 – Candidate pruning* [3,4]

The above pattern matching may extract many wrong entities. A pruning method based on POS check is proposed. It remedies some errors made by the POS tagger system. Since an entity is always associated with a POS tag in our patterns, this method checks in the dataset to see whether the POS tag is the most frequent one for this candidate. If it is not, the candidate entity is eliminated (a possible POS tagging error).

Example 3: Given the sentence: You/PRP can/MD also/RB be/VB sure/JJ it/PRP will/MD work/VB with/IN all/PDT the/DT Sony/NNP Ericsson/NNP walkman/NN phone/NN accessories./CD

The pattern, <{IN}{DT}{CD, ENTITYXYZ}>, matches the sentence segment: with/IN all/PDT the/DT Sony/NNP Ericsson/NNP walkman/NN phone/NN accessories/CD to produce the candidate entity: accessories, which is incorrect. But when the algorithm goes over the sentences in the dataset again, it found that "accessories" appear as "NNS" more often than as "CD". This candidate is deleted. The algorithm so far is generic and applicable to any domain because no assumption was made. The step below is more applicable to manufactured products (which are our main area of applications), which have brands and models. It should not be used for non-manufactured products. This step makes the assumption that a model name has a digit in it. In the experimental section we will show their results separately.

*Step 5 – Pruning using brand and model relation and syntactic patterns* [3,4]

For most manufactured products, brands and models often appear together, e.g., "Moto Razr V3". Here we need to use the above digit assumption. Thus, based on the entities that were found so far (step 4); this step tries to prune entities by using the pattern <Brand Model>. The first task is to discover relationships from the entities discovered so far. This is simple as the example below shows.

Example 4: We have the following sentence: As/RB far/RB as/IN I/PRP heard/VBD Nokia/NNP N95/CD seems/VBZ to/TO be/VB the/DT leader/NN in/IN this/DT sense./CD In this sentence, if both "Nokia" and "N95" are in the entity list, "Nokia" is considered as <Brand>, and "N95" is considered a <Model>. Then using some syntactic patterns can help find competing brands and models. The syntactic patterns exploit conjunctions and comparisons in sentences. The second task is to remove those entities discover in step 4 that never appear together with a <Band> or a <Model>, or never appear with a candidate in the syntactic patterns.

**Table 1 Word Tags for POS Tagging**

| Tag | Word | Tag | Word |
|---|---|---|---|
| CC | Coordinating conjunction | PRP$ | Possessive pronoun |
| CD | Cardinal number | RB | Adverb |
| DT | Determiner | RBR | Adverb, comparative |
| EX | Existential there | RBS | Adverb, superlative |
| FW | Foreign word | RP | Particle |
| IN | Preposition or subordinating conjunction | SYM | Symbol |
| JJ | Adjective | TO | to |
| JJR | Adjective, comparative | UH | Interjection |
| JJS | Adjective, superlative | VB | Verb, base form |
| LS | List item marker | VBD | Verb, past tense |
| MD | Modal | VBG | Verb, gerund or present participle |
| NN | Noun, singular or mass | VBN | Verb, past participle |
| NNS | Noun, plural | VBP | Verb, non-3rd person singular present |
| NNP | Proper noun, singular | VBZ | Verb, 3rd person singular present |
| NNPS | Proper noun, plural | WDT | Wh-determiner |
| PDT | Predeterminer | WP | Wh-pronoun |
| POS | Possessive ending | WP$ | Possessive wh-pronoun |
| PRP | Personal pronoun | WRB | Wh-adverb |

# 3. ROUGH SET THEORY

Rough set theory was developed by Z. Pawlak [17] on the assumption that with each object of the universe of discourse we associate some information, and the objects can be "seen" only through the accessible information. Hence, the object with the same information cannot be discerned and appear as the same. These results, that indiscernible object of the universe forms clusters of indistinguishable objects which are often called granules or atoms. These granules are called elementary sets or concepts, and can be considered as elementary building blocks of knowledge. Elementary concepts can be combined into compound concepts, i.e. concepts that are uniquely defined in terms of elementary concepts. Any union of elementary sets is called a crisp set, and any other sets are referred to as rough (vague, imprecise). Consequently, each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified as members of the set or its complement. Obviously crisp sets have no boundary-line elements at all. This means that boundary-line cases cannot be properly classified by employing the available knowledge. The main goal of rough set theoretic analysis is to synthesise approximation (upper and lower) of concepts from the acquired data. While fuzzy set theory assigns to each objects a grade of belongingness to represent an imprecise set, the focus of rough set theory is on the ambiguity caused by limited discernibility of objects in the domain of discourse. However, the rough set theory has been successfully applied to solve many real-life problems which involve decision making approaches [14,15,16,20]. The main advantage of rough set theory is that it does not need any preliminary or additional information about data like probability in statistics and the grade of membership or the value of possibility in fuzzy set theory. It has been found by investigation that hybrid systems which consist of different soft computing tools combined into one system often improve the quality of the constructed system. Recently, rough sets and fuzzy sets have been integrated in soft computing framework, the aim being to develop a model of uncertainty stronger. Therefore, Rough-fuzzy Hybridization decision systems have a significant potential.

## 3.1 Key Feature Extraction with Set Theory

Let us present some of the basic concepts of Rough set theory which are related to this paper. For more details, one may refer to [13, 14, 15, 17,20].

An *Information system* [13] can be viewed as a pair $\hat{S}=<U, A>$, or a function $f$: $U \times A \rightarrow V$, where U is a non-empty finite set of objects called the Universe, A is a non-empty finite set of attributes, such that a: $U \rightarrow V_a$ for every $a \in A$. The set $V_a$ is called the value set of a. In many applications, there is an outcome of classification that is known. This is a posterior knowledge is expressed by one distinguished attribute called decision attribute, the process is known as supervised learning. Information systems of this kind are called decision systems. A decision system is any information system of the form $\hat{A}= (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. The elements of A are called conditional attributes or simply conditions. The sentence attribute may take several values though binary outcomes are rather frequent.

*Indiscernibility and Set Approximation:* A post expresses all the knowledge available about a product. This post may be unnecessarily large because it is redundant in at least two ways. The same or the indiscernible [13,14] entity may be represented several times, or some of the attributes may be superfluous. With every subset of attributes $B \subseteq A$, one can easily associate an equivalence relation IB on U: $I_B = \{(x, y) \in U$: for every $a \in B$, $a(x) = a(y)\}$. IB is called B-indiscernibility relation. If $(x, y) \in$ IB, the entity x and y are indiscernible from each other by attributes B. The equivalence classes of the partition induced by the B-Indiscernibility relation are denoted by $[x]_B$. These are also known as granules. The partition induced by the equivalence relation IB can be used to build new subsets of the universe. Subsets that are most often of interest have the same value of the outcome attribute. It is here that the notion of rough set emerges. Although we cannot delineate the concept crisply, it is possible to delineate the entity which definitely "belong" to the concept and those which definitely "do not belong" to the concept.

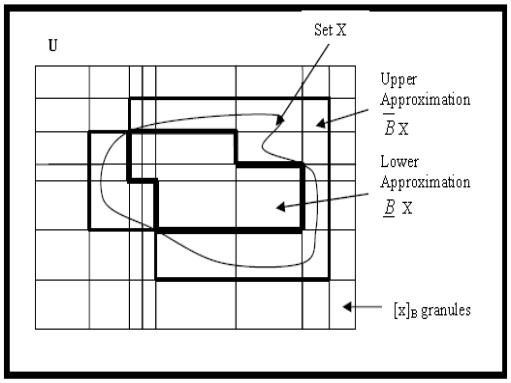


**Figure 2.Rough Representation of a set with upper and lower approximations.**

Let Â= (U, A) be an information system and let B⊆A and X⊆U. We can approximate X using only the information contained in B by constructing the B-lower and B-upper approximations of X, denoted as $\underline{B}$ X and $\overline{B}$ X respectively, where $\underline{B}X=\{x||x|_E \sqsubseteq X\}$ and$\overline{B}X \quad \{x||x|_E \cap X \neq \emptyset\}$. The objects in $\underline{B}$ X can be certainly classified as the members of X on the basis of knowledge in B and the objects in $\overline{B}X$ can only be classified as possible members of X on the basis of B. This is illustrated in Fig 2. The set $B_NB(X) = \overline{B}X \Leftrightarrow \underline{B}X$ is called the B-boundary region of X and thus consists of those objects that we cannot decisively classify into X on the basis of knowledge of B. Thus, a set is said to be rough if the boundary region is non-empty otherwise crisp (boundary region is empty).

## 3.2 Feature Reduction

Indiscernibility relation reduces the data by identifying equivalence feature [17,20], i.e. entity that are indiscernible, using the available attributes. Only one element of the equivalence feature is needed to represent the entire set. Reduction can also be done by keeping only those attributes that preserve the Indiscernibility relation and, consequently, set approximation. So, one is, in effect, looking for minimal set of attributes taken from the initial set A, so that the minimal set induce the same partition on the domain of A. In other words, the essence of the information remains intact and the superfluous attributes are removed.

The below sets of attributes are called reducts. Intersection of all reducts is called the core. Reducts have been clearly characterized in [13] by discernibility matrices and discernibility functions. Let us consider U={$x_1$,…,$x_n$}and A={$a_1$,…..,$a_n$} in the information system Ŝ=<U, A>.By the discernibility matrix M (Ŝ) of Ŝ is meant an n×n-matrix (symmetrical with empty diagonal) with entries $C_{ij}$ Ŝ as follows: $C_{ij}$ = {a∈A: a ($x_i$) ≠a ($x_j$)}. A discernibility function $f$Ŝ is a function of m Boolean variables ā 1,…, ā m corresponding to the attributes a1,….,am respectively, and defined as follows: $f$Ŝ ( ā 1,…., ā m) = ∧{∨(Cij): 1≤i , j ≤n, j<i ,$C_{ij}$ ≠Ø} Where ∨Cij is the disjunction of all variables ā with a∈Cij . It is seen in [15] that {$a_1$,….,$a_m$} is a Reducts in Ŝ if and only $a_{i1}$∩…..∩$a_{ij}$ is a prime implicate (constituent of the disjunctive normal form) of$f$Ŝ. The algorithm for feature reduction is as follows [22, 24]-

**FeatureReduct (C, D)**

C, the set of all conditional attributes;

D, the set of defined attributes;

(1) R ← { }
(2) do
(3) T ← R
(4) ∀x ∈ (C − R)
(5) if γR∪{x}(D) > γT(D)
(6) T ← R ∪ {x}
(7) R ← T
(8) until γR(D) = γT(D)
(9) return R

Suppose the customer enter the statement like " I have a Nokia N95 having good battery backup and good sound and high camera resolution." The above blog relates with an entity and it is already finds in section 2.3. The features are the "battery backup", "sound", "camera resolution". Figure 3 shows the feature extraction of an entity in GUI frame.

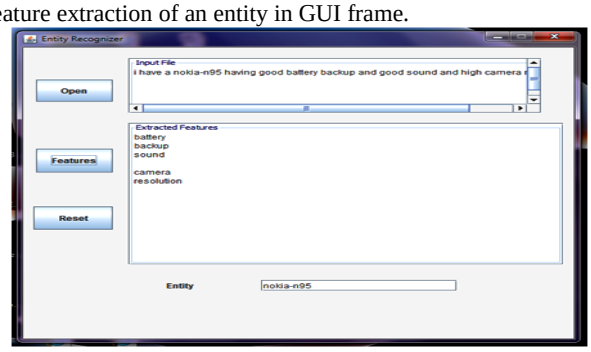


**Figure3. Feature Extracted by Entity Recognizer**

## 4. CONCLUSION

In this paper, we have presented a fuzzy-rough set aided extractive summarizer with better information coverage and redundancy reducing. The rough set theory approach gives the exact result with the help of Reducts and granules and boundary extraction and also there are spaces for improvement if the tools and methods used for document pre-processing, word sense disambiguation and sequential pattern of sentences are perfected. In addition, the method implementation is in process and is not necessarily adequate. Yet, this work does show some directions of further investigations in this work.

## 5. REFERENCES

[1] Sumeet V.Vinchurkar, Smita M. Nirkhi. Feature Extraction of Product from Customer Feedback through Blog in International Journal of Emerging Technology and Advanced Engineering (IJETAE), Volume 2, Issue 1, January 2012 (ISSN 2250 – 2459) page No.314-323.

[2] Lingyan Ji, Hanxiao Shi, Mengli Li, Mengxia Cai, Peiqi Feng. Opinion Mining of Product Reviews Based on Semantic Role Labeling in The 5th International Conference on Computer Science Education Hefei, China. August 24–27, 2010,pp.1450-1453.

[3] Domenico Potena, Claudia Diamantini. Mining Opinions on the Basis of Their Affectivity in 2010 IEEE, pp.245-254.

[4] Xiaowen Ding, Bing Liu, Lei Zhang. Entity Discovery and Assignment for Opinion Mining Applications KDD'09, June 28– July 1, 2009, Paris, France.

[5] Wei Jin and Hung Hay Ho. A Novel Lexicalized HMM-based Learning framework for Web Opinion Mining in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

[6] Themis P. Exarchos · Markos G. Tsipouras ·Costas Papaloukas · Dimitrios I. Fotiadis .An optimized sequential

pattern matching methodology for sequence classification in Knowl Inf Syst, 12 April 2008© Springer-Verlag London Limited 2008

[7] Xiaowen Ding, Bing Liu and Philip S. Yu A Holistic Lexicon-Based Approach to Opinion Mining WSDM'08, February 11-12, 2008, Palo Alto, California, USA.

[8] Han, J., and Kamber M. Data Mining: Concepts and Techniques, 2006.

[9] Minqing Hu and Bing Liu Mining and Summarizing Customer Reviews KDD'04, August 22–25, 2004, Seattle, Washington, USA.

[10] Murthy Ganapathibhotla and Bing Liu Mining Opinions in Comparative Sentences © 2008IEEE.

[11] Qiuye Zhao and Mitch Marcus. A Simple Unsupervised Learner for POS Disambiguation Rules Given Only a Minimal Lexicon in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 688–697,Singapore, 6-7 August 2009.

[12] Asif Ekbal and Sivaji Bandyopadhyay. Web-based Bengali News Corpus for Lexicon Development and POS Tagging in SPSAL-Proceedings June12, 2008.

[13] Yiyuan Cheng, Ruiling Zhang, Xiufeng Wang, Qiushuang Chen. Text Feature Extraction Based on Rough Set in Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008 IEEE, pp.310-314.

[14] Hsun-Hui Huang, Yau-Hwang Kuo and Horng-Chang Yang. Fuzzy-Rough Set Aided Sentence Extraction Summarization in Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06).

[15] Qiang Li, Jjan-Hua Li, Gong- Shen liu, Sheng-Hong Li. A Rough Set based Hybrid Feature Selection Method For Topic Specific Text Filtering in Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004, pp.1464-1468.

[16] Richard Jensen and Qiang Shen . Semantics-Preserving Dimensionality eduction: Rough and Fuzzy-Rough-Based Approaches in IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 12, December 2004,pp.1457-1471.

[17] Zdzisław Pawlak. Rough set theory and its applications in journal of Telecommunication and Information Technology 3/2002, pp.7-10.

[18] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation" In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2003.

[19] Satanjeev Banerjee and Ted Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet" In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics,Mexico City, pp. 136–145, 2002.

[20] Richard Jensen and Qiang Shen .Fuzzy-Rough Sets Assisted Attribute Selection in IEEE Transactions on Fuzzy Systems, Vol 15, No.1,February 2007.

[21] Durga Toshniwal and Rishiraj Saha Roy:"Clustering Unstructured Text Documents Using Naïve Bayesian Concept and Shape Pattern Matching" in International Journal of Advancements in Computing Technology Volume 1, Number 1, September 2009, pp.52-63.

[22] Hans-Peter Storr. A compact fuzzy extension of the Naive Bayesian classification algorithm. AI Institute, Dept. Computer Science Technische Universit¨ at Dresden.

[23] Miller, G., Beckwith, R, Fellbaum, C., Gross, D., and Miler, K. 1990. Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue),