I.J. Information Technology and Computer Science, **2017, 7, 35-41**
Published Online July 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2017.07.04

Determining the Degree of Knowledge Processing in Semantics through Probabilistic Measures

Rashmi S
Department of Computer Science and Applications, Bangalore University, Bangalore, 560056, India
E-mail: rashmi.karthik123@bub.ernet.in

Hanumanthappa M
Department of Computer Science and Applications, Bangalore University, Bangalore, 560056, India
E-mail: hanu6572@bub.ernet.in

***Abstract*—**World Wide Web is a huge repository of information. Retrieving data patterns is facile by using data mining techniques. However identifying the knowledge is tough, tough because the knowledge should be meaningful. Semantics, a branch of linguistics, defines the process of supplying knowledge to the computer system. The underlying idea of semantics is to understand the language model and its correspondence with the meaning associability. Though semantics indicates a crucial ingredient for language processing, the degree of work composition done in this area is minimal. This paper presents an ongoing semantic research problem thereby investigating the theory and rule representation. Probabilistic approach for semantics is demonstrated to address the semantics knowledge representation. The inherit requirement for our system is to have the language syntactically correct. This approach identifies the meaning of the sentence at word-level. The accuracy of the proposed architecture is studied in terms of recall and precision measures. From the experiments conducted, it is clear that the probabilistic model for semantics is able to associate the language model at a preliminary level

***Index Terms*—**Information Retrieval, Knowledge Representation, Language Model, Natural Language Processing, Probabilistic Model, Semantics.

I. Introduction

Information retrieval is the process of obtaining the relevant information from a data repository. The information on internet is huge and one needs the useful and pertinent information pattern. The retrieved information must be meaningful. As a matter of discourse, meaning has to be elevated by the computer system without humanly interactions or knowledge repositories. The degree of accuracy is utmost important and be more efficient than the man power. It is because of this reason that, semantics is considered as one of the challenging fields in Natural language Processing [1]. With the less research done in the semantic area, it gives us lot of opportunities to try and explore the topic more about the subject.

The primary goal of this research paper is to perform semantic analysis for textual data. This is dependent on the factors that influence the user's text-comprehension capability such as subjective relevance, nominal variables, primary data and cognitive association. The practical approach to characterize the word meaning is to statistically check word-word relations, draw a word-to-clause association and then to sentence formation [2]. Hence the concept of semantics exhibits a close resemblance between discourse and pragmatic analysis. However the conceptual theory is more stereotype than reality. The interpretation of verbal statements provides potential access to the knowledge. With this, a hypothesis can be made that the probability of the sentence ratio across the huge corpus will definitely yield the expected result. The probabilistic model has several advantages as mentioned below

- Uncertainty and noise can be considered as the parameters
- Manifests language structure for robustness
- This can be used to scale up for a huge set of data

Probability allows the system to learn from the language structure keeping in count of all the noise and uncertainties existing in the system. The inverse probability theory is deduced by Bayes. This is called as Bayesian probability. The rule is given by (1)

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)P(hypothesis)}{P(data)} \quad (1)$$

This indicates that there can be different ways of solutions for a given problem and every hypothesis will lead to the exactly the correct answer.

In figure [1], we describe the architecture of our proposed system. The model addresses the problem of knowledge representation at a lowest level. The obligatory requirement for our proposed system is to have the sentence in a syntactically correct form – free from

various syntactic errors such as spelling mistakes, grammar mistakes and language discourse. The sentence free from above type of errors is later structured into a trigram model. Here each of the sentences is split into three consecutive word-pairs. The word-pairs are later compared against the chosen corpus to find out the probability of occurrence. It is then collated for the base probability model of the overall sentence by taking the product terms. Finally the sentence is declared either as semantically correct or incorrect.

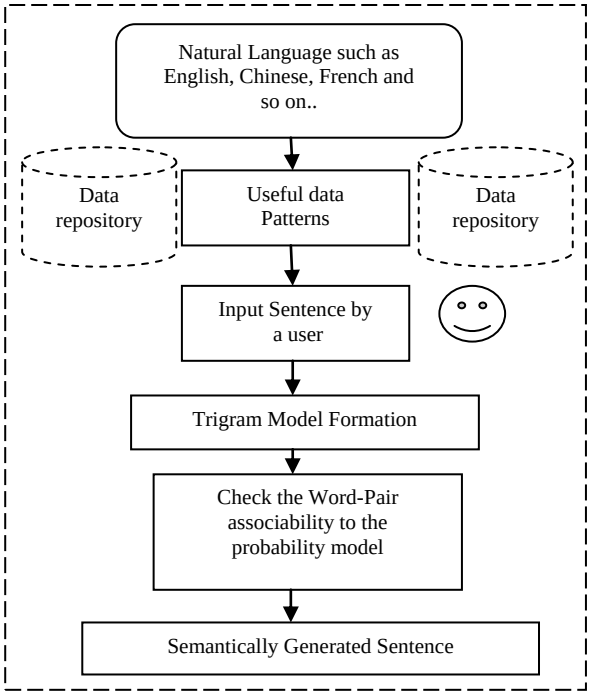

Fig.1. Architecture of the proposed system. The probabilistic Approach model for identifying the sentences either as semantically correct or incorrect

The research aspect of the semantic is relatively less as the semantic field is broad in its arena. In this paper we focus on the semantic through probabilistic approach. The existing methodologies in the chosen field are discussed in the related work section. We then talk about how the coherence of semantics can be achieved through probabilistic approach. In the methodology section, the proposed algorithm is explained. Following this, in the experimental result section the proposed algorithm is explained with examples and we also explain the calculation of term frequency. Finally in the results and discussions the accuracy of the proposed algorithm is put forward. Here the accuracy is calculated using recall, precision and F-measures.

II. Related Works

In the late 1980's there were many computers and various types of processing occurred with the help of computers. But none has the power of Knowledge Based system. Then robotics came into existence and the efficient robots were built with herculean effort. But the knowledge based systems i.e. the computer possessing the knowledge of human were not found. A question arose as to address the issue of knowledge and this branch of study was named as Semantics under Natural Language Processing. In the year 1996, Peter W. Foltz [3] introduced the concept of Latent Semantic Analysis (LSA) for text based search. They showed the comparison to be performed at semantic level rather than surface level as with hand coding. The results of their experiment almost correlated with that of humans. The semantic similarities between the pieces of textual information were exposed with the LSA in this paper.

In the year 2006 [4], Doucet worked on presenting a novel approach for disseminating the occurrences of a disconnected sequence of n words. The Markov model was used for this implementation to reduce the computational complexities. The obtained and expected frequencies through the statistical test were done to assess the results. This approach does not require the human involvement and it application dependent.

G Madhu et al [5] has done a survey on the existing Intelligent Semantic Web Search Engines. The characteristics of these engines were studied. The survey concluded on four distinct differentiations such as designers and users' perceptions, static knowledge structure, low precision and high recall and lack of experimental tests.

Besides research indications, as observed there is a very little work that has been done so far in the field of semantics. The design patterns as perceived can adopt lot of improvisation. A corpus based semantic study can be made to augmentation of traditional aspects with the primitive correlations. The query problem can be elevated to much deeper concerns[6]. Having said, the problem of semantic language representation itself is a huge paradigm. Therefore in this research paper we have concentration one tiny aspect of semantics and have shown the implementation details of its organization.

III. Semantics for Natural Language Procesasing – A Probabilistic Approach

The probabilistic approach for semantics is to predict what words are more likely to arise in a given context of a sentence. The relevance of this prediction will ensure to create a true semantic association that can be matched across various contexts that the user inputs. The distribution model is shown in equation [2],

$$p(w1 \mid w2) = \sum_{n-1}^{S} p[(w1 \mid w2) \mid S_{1,2}] \quad (2)$$

In equation [2], w1 and w2 are any 2 consecutive word-pair. The probability of this occurrence is entrusted by the corpus which is considered at the training stage of the architecture. The corpus (S) must also consist the same word-pair pattern i.e., w1 and w2 in it. If so the sentence is semantically correct. The enumeration of equation [2] can be extended to the generalized pattern as shown in equation [3],

$$p(wi, wj \mid Wn) = \sum_{n-1}^{S} p[(wi, wj \mid wn) \mid S_{i,j...n}] \quad (3)$$

The co-occurrence pattern is repeated for every sentence with different variations in context. However the chain model, Markov chain enables the enumeration of such a pattern. The initial stage of the Markov chain model is drawn sequentially thereby embarking the frequencies of the word-pair occurrence. After composing all the comparison of the word-pair in the given sentence the Markov chain terminates its iterations. Every time the Markov chain is initialized the comparison model is drawn from the equation [2] and [3]. However the speed of comparison is truly dependent on the place of occurrence of each word-pair as in the chosen corpus. The deeper the word-pair is found in the corpus the more time the Markov chain takes to arrive at the result [7]. Consider for example, the length of the corpus is 10000 words. Let us assume the word for comparison is found in 9000^{th} position of the corpus. In probabilistic approach the target word has to be compared 9000 times hence it takes longer time than usual [8]. The distribution of samples is directly related to the length of the corpus. Such a symmetrical correlation is hard to recite as the accuracy is dependent on the chosen corpus in par with the distribution domain.

IV. Methodology

In this paper the semantic representation for textual data is elucidated using the probabilistic models. Semantics bridge the gap between logical form of data representations and deep understanding [9]. Semantics addressed in this research work mainly focuses on three stages; Lexical, Sentential and discourse. The lexical part focuses on the mere meaning of the words whereas sentential observes the meaning of the overall sentence and finally discourse establishes the contextual meaning of the sentence. The probabilistic version of semantics expresses the dependencies among the different word phrases given in a sentence. However one might argue that this is a complex paradigm as the entire meaning of the sentence has to be preserved to know the overall semantic structure. On the other hand, if there are no such dependencies among the word phrases it becomes impossible to derive the meaning [10]. However we have not made an attempt to capture all the word phrases. Instead, a built-in corpus is used for the textual comparison. Observe the examples of semantically incorrect sentences, "Apple eats Alice". This sentence is syntactically correct i.e. there are no grammatical mistakes nor there are any errors in the structure formation of the sentence or the sentence phrase. However the sentence indicates no meaning (Semantics) and makes no sense as Apple cannot eat Alice. The fortune of this paper is to identify these kinds of semantically incorrect sentences. Figure [2], represents the proposed algorithm to identify the whether or not the given sentence is semantically correct. This indicates that the sentence whose probability of occurrence falls within the defined probability then the sentence is categorized as semantically incorrect.

The probabilistic approach adopted in our architecture shows the frequency distribution of the word occurrence in the chosen corpus. The important criteria for a consideration is to have the corpus as big as possible since the word-pair combination will be more and hence the proposed algorithm will have more accurate results during comparison [11]. The generative model of the semantic representation is dependent upon the probability of the word that appear in the corpus and its coherence with the dense patterns. In this section we discuss about the proposed algorithm to identify whether or not a given sentence is semantically correct. Initially the user input the sentence; the sentence has to be free from syntactic errors. Later a trigram model is generated for the syntactically correct input sentence. Each trigram combination is compared with the corpus. If the exact combination is found, then the probability of the combination is calculated. If not, the probability of that particular trigram combination is considered as zero. This is repeated for all the trigram combination in the given input sentence. Final the product term of all the frequencies are calculated. If above zero, then the sentence is semantically correct otherwise the sentence is semantically incorrect.

Algorithm_for_Probablistic_approach_in_Semantic_representation

Step:1:–*Consider_a_major_English_Corpus_such_as*
**GoogleN–gram_Corpus*
**American_National_Corpus*
Step:2:–*Input_the_sentence*
Step:3:–*Scan_the_corpus_across_the_given_sentence*
Step:4:–*Perform_the_word–to–word_analysis_on_the_corpus*
Step:6:–*If_match_is_found_mark_the_sentence_Semantically–Correct*
else
if_no_match_found_mark_the_sentence_Semantically–Incorrect
Step:7:–*END*

Fig.2. Proposed Algorithm, the probabilistic approach to identify whether a given sentence is semantically correct or not

The corpus considered in our study is Google N gram. The algorithm works only if the chosen training corpus contains huge collection of words both at lexical and sentential level. This is because when the corpus is huge the samples for the test data will also be more.

V. Experimental Results

In this section we test the proposed algorithm on a chosen sentence to evaluate its working. The proposed algorithm was implemented on ASP.NET platform. We have used the Google N gram corpus for our algorithm since this corpus is huge and it contains around 520 million words [12] with various annotated sentences of approximately 12 – 13 million. It is better when compared to other corpus as it includes part-of-Speech (POS), lemma combinations, tokens levels are at 3, 2 and 1. Hence we can impose the trigram language model as well. The corpus also contains minimum number of repetition words. The number of unique words is proportionately big as compared to other corpus such as American National Corpus [13], COCA and so on [14]. The interface was built to take the input from the user. The given input is subjected to trigram model which groups each word in one section. This section was then compared against the chosen corpus. The existence of word occurrence in the POS form is marked and then the probability is calculated. Any trigram with the probability equal to zero will obviously mark the sentence as semantically incorrect. This is because we consider the product of all the trigram combinations. The evaluation is the input sentence brought into various phases are explained in the below example.

Consider an example, **"Stone eats Dog every Sunday"**

The above example then goes to the classification of trigrams. The overall trigram combination probability is set as per the formulae [2] and [3]. The probability of each trigram combination is recorded and finally the values are multiplied to know the probability of the sentence in the corpus. This becomes the lower threshold [15]. If the probability of the word-pair falls below this threshold, then the word is error prone i.e., it will be declared as semantically incorrect. In the trigram language model, the sentence will go through a series of bifurcations. This is indicated in the below explanation,

The term frequency of a word pair is calculated as, Consider there are N words. $\{A_1, A_2 \ldots A_n\}$ In this the TF can be calculated as,

$$Term\_Frequency = \frac{Number\_of\_occurrences\_of\_particular\_word}{Total\_number\_of\_words\_in\_the\_corpus} \quad (4)$$

In the first iteration, "* * Stone" is compared against the chosen corpus. The probability of this trigram combination in the corpus is indicated in the table 1 – (i)

In the second iteration, "* stone eats" is scanned in the corpus – (ii)

In the third iteration, "Stone eats Dog" – (iii)

In the fourth iteration, "eats dog everyday" – (iv)

In the fifth iteration, "dog every Sunday" – (v)

Finally the probability of all the trigram model is combined in the following way, (i) * (ii) * (iii) * (iv) * (v) = 0 as (ii) and (iii) are zero. Therefore the above sentence is semantically wrong.

Table 1. Trigram Model for the example sentence. The probability values associated for each trigram word-pairs of the given sentence. It proves the given sentence is semantically incorrect through the proposed probabilistic approach.

Corpus Name	Probability in terms of term frequency
For (i)	0.23
For (ii)	0
For (iii)	0
For (iv)	0.02
For (v)	0.31

Example 2: Dog eats stone every Sunday

In the first iteration, "* * Dog" is compared against the chosen corpus. The probability of this trigram combination in the corpus is indicated in the table 1 – (i)

In the second iteration, "* Dog eats" is scanned in the corpus – (ii)

In the third iteration, "Dog eats stone" – (iii)

In the fourth iteration, "eats stone every" – (iv)

In the fifth iteration, "stone every Sunday" – (v)

Finally the probability of all the trigram model is combined in the following way, (i) * (ii) * (iii) * (iv) * (v) = 0 as (ii) and (iii) > zero. Therefore the above sentence is semantically correct.

Table 2. Trigram Model for the example sentence. The probability values associated for each trigram word-pairs of the given sentence. It proves the given sentence is semantically correct through the proposed probabilistic approach.

Corpus Name	Probability in terms of term frequency
For (i)	0.32
For (ii)	0.42
For (iii)	0.03
For (iv)	0.10
For (v)	0.28

4.1 Why the comparison happens based on POS of the language?

Most of us would think that rather than having the word pair in the comparison we can assign the POS tags of the word pair to be compared. But, here is why this approach is not feasible; a single word can have multiple POS tags example, Stone. Stone can be noun or verb. If we assign noun, then according to the example, eat becomes verb. Obviously noun is followed by verb in many cases. Therefore the example sentence may be considered as semantically correct. Suppose we build a rule saying if a sentence has noun followed by verb then the sentence is incorrect. This logic works for "Stone eats dog every Sunday" will be identified as semantically incorrect (true) but it will also identify "TajMahal looks beautiful' as incorrect because "TajMahal" is noun and "looks" is verb. However the sentence is actually correct. Therefore it is always ideal to perform the comparison on word-word basis rather than POS tags. A key point to note in the proposed approach is that the semantic

relationship between word-pair is determined as a function of syntactic values and syntactic component structures. The linear distribution of the word-sets across the corpus is enumerated by the probability of occurrence of the exact word-pair in the input sentence. However the proposed model does not represent semantics stochastically as it does not eliminate the risk factor when an unidentified word-pair is encountered. A knowledge generation through semantic analysis is characterized by the domain of meanings associated with the word-pair. Perhaps it is still unclear how the domain has to be relinquished as a primary form of knowledge representation. Henceforth semantics at a preliminary level concentrates on the word-to-word meaning identification as composed to arrive at the overall meaning of the sentence.

VI. Results and Discussions

In this section, the accuracy of the semantic analysis of the proposed system is evaluated. While evaluating the results, one major component to consider is the intended meaning as achieved through the process of semantics. Furthermore the study of semantics is challenging and demanding. This is because of the following facts

- The study of meaning differs from speaker and listeners. When one defines meaning of certain words what are the measures used to rate the accuracy of the meaning
- A single word might have many meaning. In contrast many words might have single meaning. This is purely related to Word Sense Disambiguation (WSD). Should there be any relation between WSD and Semantics?
- Whenever there is revolution in English language what happens if a word gets a new meaning or overwrites the existing meaning?

Hence addressing semantics is very tricky because the approach cannot be straightforward. The realm of semantics somehow depends on the theoretical and arbitrary concerns of the language. The linguistic processing of semantics should be perceived with a sense of concise otherwise the meaning evaluation becomes a major consideration for any natural language. Therefore in this paper, semantics is evaluated only in the primary meaning of discrete senses. The syntactic entailment is not considered in the present evaluation system. The hypothesis made during this research work is to impart the sentences that are syntactically correct. In order to assess the efficiency of the proposed system, the evaluation study is divided into three phases. First, the training phase; here the corpus used is Google N Gram consisting of 520 million words with contribute texts and annotated data is considered as the training corpora dataset. The system is trained with this corpus so that the test data is checked against the training corpus. Second, testing phase is induced with the proposed algorithm on the given input sentence. Third and lastly, evaluation phase; in this stage, the proposed algorithm works on the input sentence and classifies it either as semantically correct or incorrect sentence. To measure the efficiency of the proposed system the accuracy measures are used. The test was conducted on 1000 sentences that were manually built during the research work. Out of 1000 sentences, 500 were semantically incorrect sentences and 500 were semantically correct sentences. The sentences were exposed to all the three phases of evaluation. The check measures obtained from the proposed system is indicated in the table [3].

Table 3. Result evaluation of the proposed system in terms of TP, TN, FP and FN

	Identified as Semantically Wrong [I]	Identified as Semantically Correct [II]
Semantically (500) Wrong Sentences [I]	410 [True Negative]	90 [False Positive]
Semantically (500) Correct Sentences [II]	53 [False Negative]	447 [True Positive]

The result tabulated in table [3] has four possible cases as discussed below:

[I] [I] – The system identified the semantically wrong sentences as semantically wrong – This is True Negative (TN) case.

[I] [II] – The system identified the semantically wrong sentences as semantically correct – This is False Positive (FP) case.

[I] [I] – The system identified the semantically correct sentences as semantically wrong – This is False Negative (FN) case.

[I] [I] – The system identified the semantically correct sentences as semantically correct – This is True Positive (TP) case.

Therefore with these four cases, we can arrive at recall [5] and precision [6] formulae as,

$$\text{Re}\,call = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Pr}\,ecision = \frac{TP}{TP + FP} \tag{6}$$

By substituting the values from table [3] in equation [5] and [6] we have, Recall as 89% and Precision as 83.24%.

Overall accuracy of the system is given by [7],

$$Accuracy = \frac{TN + TP}{Total\_Number\_of\_word - set} \tag{7}$$

$$Accuracy = \frac{410 + 447}{1000} = 85.7\%$$

Therefore the proposed algorithm achieves the accuracy of 85.7% with recall and precision being 89% and 83.24% respectively.

Though the proposed approach attains a good accuracy it has certain drawbacks. Those are:

- The efficiency of the system is studied for only semantic aspect. The system fails to identify a sentence if it is semantically correct and syntactically wrong
- The lexical organization of the sentence is not evaluated since the focus is on the training data and the test data. For example: “My uncle was stoned after hearing to the news”. The sentence is semantically correct however the abstract meaning of the “stoned” is not observed as it is not present in the training data henceforth the system identifies the sentence as semantically incorrect.
- The system fails to study the pragmatic observations such as “Sky is green in color” or “Milk is black in color”

These observations can be marked for the future study of this work. It should also be noted that there are certain phrases in any natural language that is difficult to put across in a correct meaning. This becomes a linguistic challenge as well. The concept of semantics varies from element to element mapping which limits the cognitive simulation of a language.

VII. Conclusion

Semantics for a language is very important. The Semantics can be described in many patterns. The one observed in this paper is the meaning of a sentence. We have used the probabilistic approach to address semantics. The frequency of word-occurrence distribution is shown across a large scale of word-pairs that is encountered in a given sentence. Each word-pair is cleaved into a trigram model as converged into three phases of study; training, testing and evaluation. A corpus is considered in the training phase to arrive at the datasets. The testing phase consists of the proposed algorithm under lower dimensional meaning networks. Finally the evaluation stage indicates the accuracy of the proposed system. The evaluation study is made in terms of Recall, Precision and Accuracy which is proven to be 89%, 83.24% and 85.7% respectively. The study of present system is explicitly made by probabilistic approach. In future, a dictionary based approach can be used to address the semantics and extended towards building a robust knowledge processing system.

References

[1] Eetu M ̈akel ̈a “Survey of Semantic Search Research”, Research gate 2008.
[2] James R Hurford, “Semantics, a course book” Cambridge University Press 2007
[3] Peter W. Foltz, “Latent semantic analysis for text-based research” Behavior Research Methods, Instruments, & Computers 1996, 28 (2), 197-202
[4] Antonie Doucet, “A method to calculate probability and expected document frequency of discontinued word sequences” Traitement Automatique des Langues (TAL) 46, 13-37 (2006)
[5] G Madhu, “Intelligent Semantic Web Search Engines: A Brief Survey” International Journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011
[6] Masao Yokota, “ Aware computing guided by Lmdexpression and direct knowledge in spatial language understanding”, 2011 3rd International Conference on Awareness Science and Technology (iCAST), DOI: 10.1109/ICAwST.2011.6163164 Publisher: IEEE
[7] Radziah Mohamad, “ Similarity algorithm for evaluating the coverage of domain ontology for semantic Web services”, International conference on Software Engineering Conference (MySEC), 2014 , DOI: 10.1109/MySec.2014.6986012 Publisher: IEEE
[8] Guanghui Yang,Junkang Feng,"Database Semantic Interoperability based on Information Flow Theory and Formal Concept Analysis", IJITCS, vol.4, no.7, pp.33-42, 2012.
[9] Zahia Marouf, Sidi Mohamed Benslimane,"An Integrated Approach to Drive Ontological Structure from Folksonomie", IJITCS, vol.6, no.12, pp.35-45, 2014 DOI: 10.5815/ijitcs.2014.12.05
[10] Kavitha, A., Rajkumar, N., and Victor, S.P., An Integrated Approach for Measuring Semantic Similarity Between Words and Sentences Using Web Search Engine. The International Journal of Information Technology & Computer Science (IJITCS), 9(3), 68-78.2013
[11] “Semantic analysis of Natural Language”, by Poroshin.V.A, Saint-Petersburg State University, 2004.
[12] “Speech understanding through syntactic and semantic analysis”, Donald E. Walker Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California., DE Walker, AI center, advanced papers of the conference
[13] Hiroyuki Yamauchi, “Processing of syntax and semantics of natural language by predicate logic”, institute of space and aeronautical science, university of Tokyo, proceedings of the 8th conference on computational linguistics, pages 389-396
[14] Wood, G.C.,” Lecture on Introduction to Semantics at the University of Sheffield”.
[15] Yi Jin, The Research of Search Engine Based on Semantic Web”, International Symposium on Intelligent Information Technology Application Workshops,2008 DOI: 10.1109/IITA. Workshops.2008.193 Publisher: IEEE

Authors’ Profiles

Mrs. Rashmi S is a Research Scholar in the Department of Computer Science and Applications, Bangalore University, Bangalore, India. She also has over 5 years of teaching as well as Industry experience. She has published several papers in National and International Journals with good impact factor. She has also attended various conferences and presented papers. Her specialization in research is Data Mining. She has published several papers in various National and International conference and Journals.

Dr. Hanumanthappa M is currently working as Professor and coordinator in the Department of Computer Science and Applications, Bangalore University, Bangalore, India. He has over 17 years of teaching (Post Graduate) as well as Industry experience. He is member of Board of Studies/Board of Examiners for various Universities in Karnataka, India. He is actively involved in the funded research project and guiding research scholars in the field of Data Mining and Network Security.

How to cite this paper: Rashmi S, Hanumanthappa M,"Determining the Degree of Knowledge Processing in Semantics through Probabilistic Measures", International Journal of Information Technology and Computer Science(IJITCS), Vol.9, No.7, pp.35-41, 2017. DOI: 10.5815/ijitcs.2017.07.04