International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.8, August 2012 – www.ijais.org

Semantic based Text Summarization using Universal Networking Language

S. Mangairkarasi
Assistant Professor
Department of Computer Science and Engineering
J.N.N Institute of Engineering, Chennai,
Tamil Nadu, India

S. Gunasundari
Assistant Professor II
Department of Computer Science and Engineering
Tamil Nadu, India

ABSTRACT
Text Summarization is extracting the important information from the document by leaving out the irrelevant information, and to reduce the details and collects them in a compressed way. Normally text summarization is done in single or multi documents. The advantage on processing time can be achieved in the text summarization. Converting English sentences into expressions or Interlingua is called Universal Networking Language (UNL). The given source document is preprocessed by eliminating tables and images. The preprocessed document is fed into sentence splitter and then to word separator. The given word is sent to Morphological Analyzer to find the root word. This root word is fed into UNL dictionary for finding the corresponding concepts and attributes. By using the heuristic rules, we identify the relations between concepts. UNL represents knowledge in the form of graphical format, where nodes represent concepts and links represent relations between concepts. It represents the whole document not the sentences in particular. The graph algorithm is used to find the weight age of links connected to the Universal Word. According to the highest weight age the document is summarized.

Keywords: Interlingua, Document preprocessor, UNL dictionary, Morphological Analyzer

1. INTRODUCTION
Natural Language [1] is mainly used to build a model for its analysis and generation. Natural Language Processing (NLP) has significant overlap with the field of computational linguistics, and is often considered as a subfield of artificial Intelligence. NLP may encompass both text and speech. NLP is used for automatic summarization, machine translation, information retrieval and information extraction. The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether the system answers the goals of its designers, or meets the needs of its users.

Text Summarization focuses to reduce the length or complexity of the original text without losing the main concept. Single document summarization techniques have the potential to simplify information by presenting only the relevant information contained in the document. The combination of completeness, readability and conciseness would give a good summarization.

Universal Networking Language [2] is a computer language; which is used to process information and knowledge. It is used to intercommunicate, among the people with a Linguistic Infrastructure (LI) for receiving and understanding multilingual information. The process of converting a source language expression into the UNL expression is called as "enconversion". UNL represents knowledge in the form of relations between concepts. It provides a uniform concept Vocabulary. Universal Words can also be annotated with attributes like number, tense etc, which provides further information about how the concept is being used in the specific sentence.

2. UNIVERSAL NETWORKING LANGUAGE REPRESENTATION
Universal Networking Language represents a document in the form of a graph with nodes as the universal words and relations between them as links. Information written in the Natural Language may be enconverted to UNL and UNL can be deconverted into Natural Language. UNL expresses a document using Universal Words, Relations and Attributes: Universal Words (UW) is based on English words. Every UW denotes a concept. A single UW is a character-string made up of two different parts: a headword and a constraint (list). The headword can be a word, which is used to denote concepts. The constraint (list) is used to delimit a concept within a range. UWs are restricted using constructs which describe the sense of a word in the current document. UWs are interlinked with other UWs to form a UNL expression corresponding to a natural language sentence. These links are called relations that specify the role of each word in a sentence. Speakers' views, aspect, time of the event are captured by UNL attributes. It also expresses additional information about the Universal Words that appear in the sentence.

3. UNIVERSAL NETWORKING LANGUAGE DICTIONARY
The UNL dictionary consists of root word, concept and attributes. The work of Morphological Analyzer is to find the root word. The purpose of Concept identifier is to identify the concepts of root word from UW Dictionary. In that dictionary it finds the corresponding concept and attributes for the root word.

3.1 Semi Automatic UNL Dictionary Construction
The construction of fully automatic UNL dictionary is not possible to implement and at the same time a complete manual updating of UNL dictionary is a tedious job and time taking process. Hence in this paper we try to propose an algorithm to construct a semi-automatic UNL dictionary. Fig. 1 is a diagrammatic interface for dictionary construction which can be used as Semi automatic. The set of educations documents downloaded from web are words splitted and the root word of the splitted word is identified from morphological analyzer. The identified root word is checked

in main UNL dictionary for the presence. If it not present, type the concept and select the constraints from the list which is displayed and update the UNL dictionary.

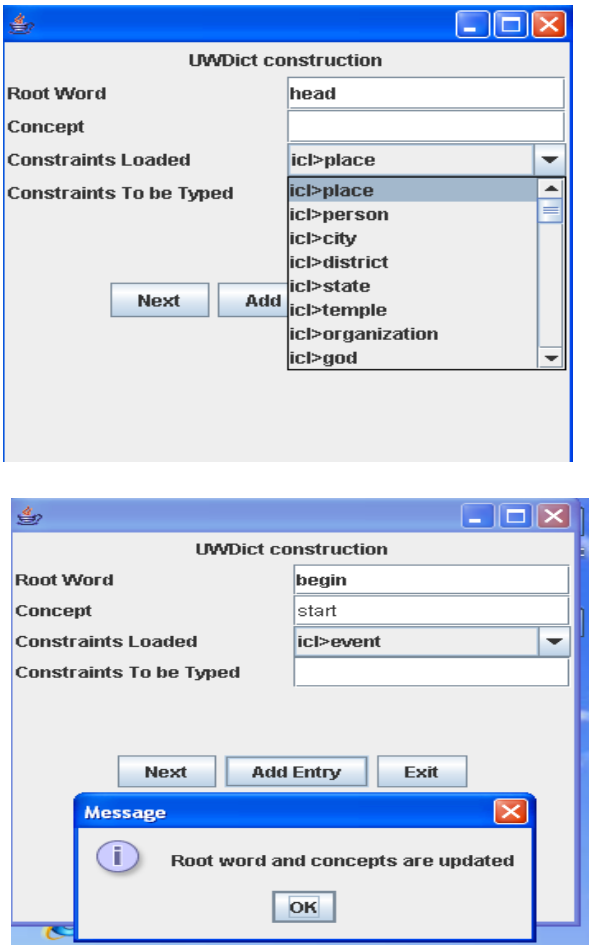

Figure 1 Semi Automatic Dictionary Construction

3.2 Representation

The various forms of representing a dictionary are arrays, Linked Lists and Trees. The Binary Search Tree is effective and efficient as it has time complexity of O(log n) and it can also be used for real time processing. BST is a node based binary tree data structures. The tree starts with the root word. It can be implemented as a linked data structure. Node consists of root word and universal word. The left sub tree of a node contains only nodes with keys (root word) less than the parent node. The right sub tree of a node contains only nodes with keys greater than the parent node. More than one root word gives as the same concept. The major advantage of binary search trees over other data structures is that the related sorting algorithms and search algorithms such as in-order traversal can be very efficient.

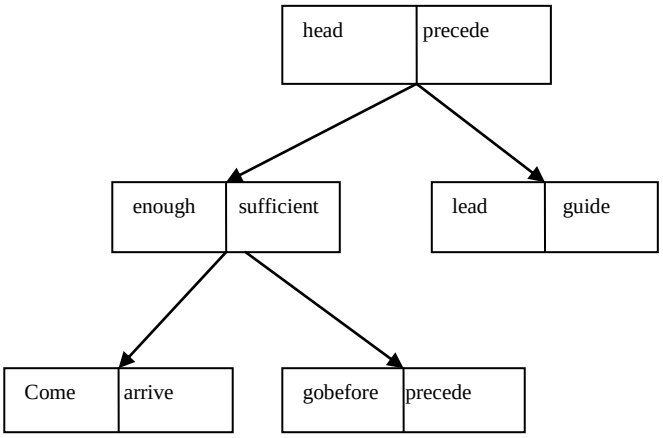

Figure 2 Binary Search tree

4. UNL RELATIONS

Relations as well as UWs are used to describe the objectivity information of sentences. There are many factors that are used to choose an inventory of relations between concepts. It has different set of relations. The UNL relations basically selects the principles such as necessary condition and sufficient condition: When a UW has relations between more than one other UWs, each relation label should be set to identify the knowledge about the concept of each UW expressed. When there are relations between UWs, each relation label should be set to understand the role of each UW only by referring to the relation label.

agt Relation:
Rules:
An agent is defined as the relation between
a word in named entity having relation with verb as its neighboring word, and
a word in noun having relation with verb as its neighboring word
Example: John breaks the computer
agt(break(icl>do),john(iof>person))

man Relation:
Rules:
A manner relation is defined between
an event or state and a manner, and
a word that exists in a way characterized by word2
Example: I often visit him
man(visit(agt.thing,obj>thing),often)

fmt Relation:
Rules:
The range is defined as the relation between initial thing and final thing.
Where the word describes the beginning of a range and the other word describes the end.
Example: The weekdays from Monday to Friday are holidays.
fmt(Friday(icl>day),Monday(icl>day))

frm Relation:
Rules:
An origin is defined as that the relation between
A word which indicates an initial state of a thing and associated with the focused thing
Example: A visitor comes from Japan

frm(visitor(icl>person),Japan(iof>country))

and Relation:
Rules:
A conjunction is defined as the relation between
a concept that indicates a partner to have conjunctive relation to another concept.
Example: John and Mary are good friends.
and(Mary(iof>person), John(iof>person))

or Relation:
Rules:
A disjunction is defined as the relation between
a concept that indicates a partner to have disjunctive relation to another concept.
Example: John or Jack will done this work.
or(Jack(iof>person), John(iof>person))

plc Relation:
Rules:
A place is defined as the relation between
An event or a state and a place or thing understood as a place.
Example: My mother cooks in the kitchen
plc(cook(agt>thing), Kitchen(icl>place))

int Relation:
Rules:
An intersection is taken between
A class concept and another class concept
It indicates all common instances to have with a partner concept.
Example: An intersection of tableware and cookware
int(tableware(icl>tool),cookware(icl>tool)))

ins Relation:
Rules:
An instrument is defined as the relation between
An event or a state and a concrete thing
It indicates an instrument to carry out an event
Example: He cut the string with a pair of scissors
ins(cut(agt>thing,obj>thing),scissors(icl>thing))

icl Relation:
Rules:
An upper concept or a more general class concept is defined as the relation between
A class concept and a class concept
Example: A bird is a kind of animal
icl(bird(icl>animal),animal(icl>living thing))

4.1Identification of Relation

To identify the relation, first we have to start with the root word identification. With this root word, we have to identify the tenses namely verb, noun and named entity.
The combination of tenses and heuristic rules are used to identify the relations. For example, when a noun combines with the do verb this is the possibility of the agent relation. Based on cue words the probability of and, or, fmt,frm relations are to be identified. Sample rules are given in the table Tab 1.

Tab 1 Heuristic Rules

S.no	Relation	Cue word	Rules
1	and	and (conjunction)	If a concept have the conjunctive relation with another concept
2	or	or (disjunction)	If a concept have a disjunctive relation with another concept
3	agt	-	If a word is a named entity or noun and neighboring word also a verb.
4	fmt	From-to	If one word indicates as the beginning and the other word indicates as a end

5. GRAPH REPRESENTATION

UNL represents a document in the form of a graph with nodes as the universal words and relations between them as links. The main concept of graphical representation is to calculate more no of links to and from a UW. The effective graph algorithm is used to find the weight age of links connected to the Universal Word. According to the highest weight age the document is summarized. The graph is represented by using the Multi-List. Once the graph is built, it is possible to decode it to any other language

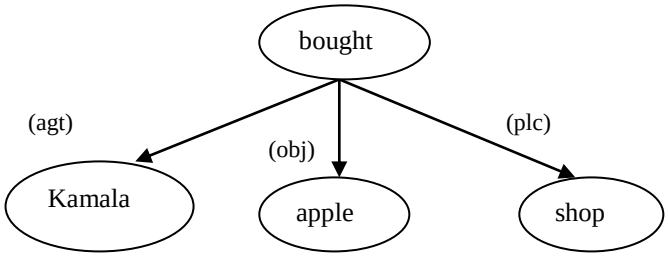

Figure 3 UNL Graph

5.1 Multi-list Representation

Multi-list representation structure is very flexible in nature. In this structure, a collection of items are represented as a list. The memory requirements of this multi-representation scheme are quite low. The whole document represents one single Multi-List representation. Each node
Consists of sentence-id, document-id, concept-id and pointer field. Each node can have any number of pointers to other nodes. The head node consists of a relation to concept from node and relation node. The relation node contains all the relations of the given the document.

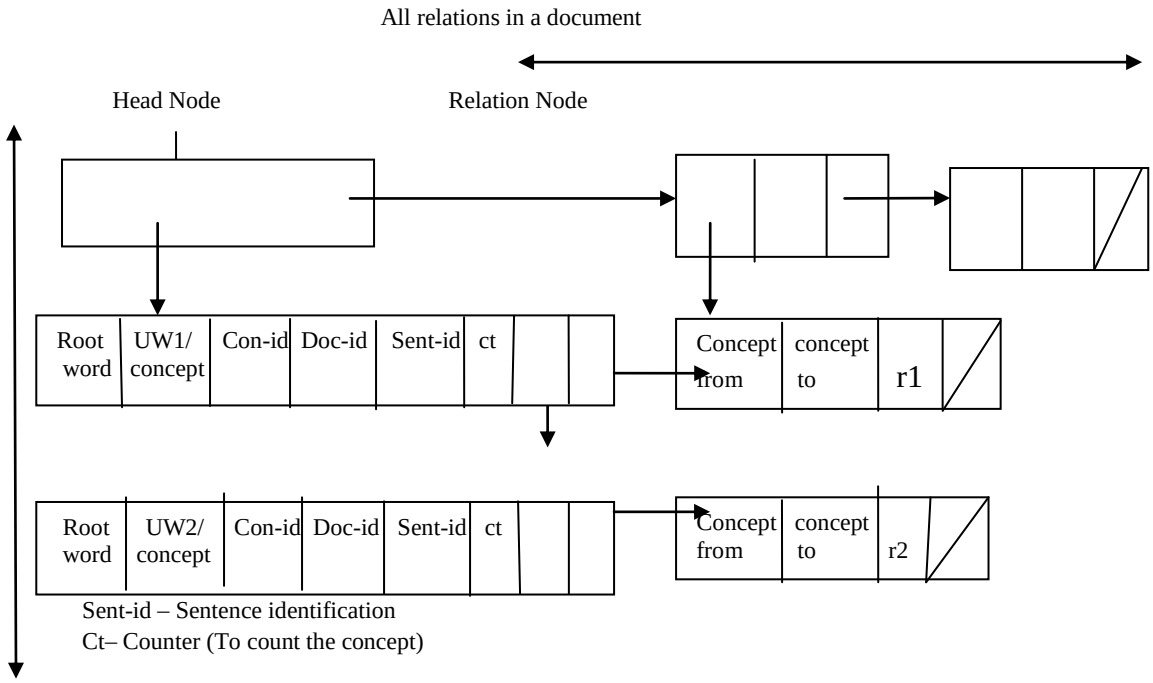

Figure 4 Structure of Multi-List Representation

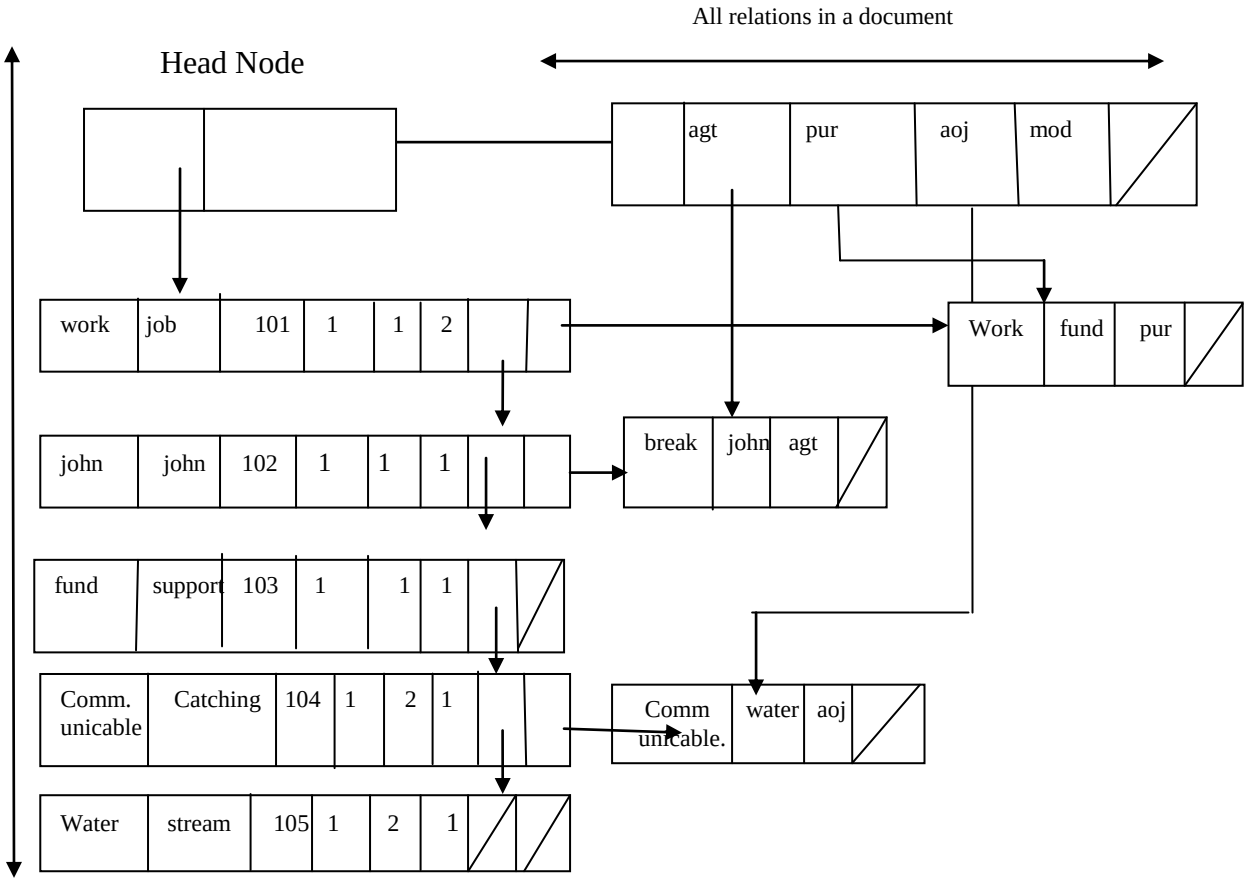

Figure 5 Example for Multi-list representation

6. IMPLEMENTATION

Documents are collected from websites based on the topic education. The collected information contains images, unwanted things so we removed that from the document. We eliminated the stop words, because this is most important process. Then the document is splitted into sentences using sentence splitter. The delimiter is blank space. Sentences are again splitted into words. The word is sent to the Morphological analyzer to identify the root word. This root word is sent to the UNL dictionary and it identifies the concept. Tenses of root words are also identified. With the help of cue words, tenses of root word and heuristic rules relations between concepts are found. The graph is constructed for the identified concepts and relations. The graph is represented using MultiList.

While constructing graph, counter field in the concept node is periodically updated. Counter values are used to know the important concepts, based on the threshold. Sentences which are having important concepts are finally picked for the summarized output. We have tested with the documents from web related to education domain. Manually we verified the output with senior people. We found the result is good.

7. EXPERIMENTAL SETUP

7.1 Ideal Summary Preparation:

Our documents are collected from commercial available news service providers like Google. Each document includes the title, image, table, so we have first created an ideal summary for evaluation our results.

Evaluation is a essential step for both single and multi-document summarization. We evaluate the automated summary with the human generated reference summary based on ranking or sentences by judges. Ranking means, assigning weights based on the level of importance sentences, and also ordering the sentences in the descending order of weights.

7.2 Experimental Result and Analysis:

To obtain ideal results, we distributed our documents to three judges and asked them to rank the sentences according to their importance. Their age group varies from 25 to 40 and all of them post graduates. Table 1 shows the agreement among judges for 3,2 and 1 sentences agreement respectively for 30% and 100% agreement.

Table 2 Agreement among Evaluators

Number of judges agree on a sentence	Agreement at 30%	Agreement at 100%
3	24.73	9.35
2	40.53	45.26
1	30.85	48.05

Table 3 Comparison between Human generated summary and our judges

DOC ID	Compression ratio	
	20%	30%
Doc 1	100	75.5
Doc 2	100	80.2
Doc 3	100	100
Doc 4	100	100
Doc 5	100	100
Doc 6	80	75.3
Doc 7	100	100
Doc 8	100	80
Doc 9	90	100
Doc 10	100	65.3
Doc 11	100	100
Doc 12	100	85.6
Doc 13	80	75
Doc 14	100	70.8
Doc 15	100	100
Doc 16	100	100
Doc 17	100	80.2
Doc 18	95	100
Doc 19	100	70
Doc 20	100	100
Doc 21	100	100
Doc 22	100	100
Doc 23	100	80
Doc 24	89	100
Doc 25	100	100
Accuracy	97	89

8. RELATED WORK

Natural Language [1] is used to build intelligent computer systems such as machine translation systems, natural language interfaces to databases, and also to gain better result on human's communication using natural languages. Universal Networking Language [2] is a computer language, which is used to process information and knowledge. UNL replicates the functions of natural languages in human communication. Automatic Summarization [3] is the creation of a shortened version of a text by a computer program. This program analyzes a text statically and linguistically determine important sentence and generate a summary text from these important sentence. Shanmugasundaram Hariharan [4] proposed MEAD is a tool kit for multilingual summarization and evaluation. It implements multiple summarization algorithms such as Centroid based and position based algorithms. Bracewell,D.B et all [5] says Machine translation is required for translating keywords, answers and document. Morphological Analysis is the identification of word stems is a fundamental part of Natural Language Processing. Amitabhu Mukerjee et all [6] proposed UNL delivers a single homogeneous Language-independent encoding; it is possible to achieve the question answering; it was developed as a universal knowledge-encoding mechanism. [8] Says generating an effective summary requires the summarizer to select, evaluate, order and aggregate items of information according of their relevance to a particular subject. [9] Proposed Evaluation of summary is based on intrinsic and extrinsic evaluation methods. In [10] multi document summary, they use two dimensions of cognitive styles to assess the summary from a set of documents. [11] used novel method structure cosine similarity to cluster a document in a new way, and consider the factors such as quality and efficiency to improve the performance of document. [12] Proposed extractive approach to create a gene cluster and expectation maximation based algorithm is used to identify sub topics and extract the relevant topic automatically. In [13] text analyzer is used to analyze the structure of text based on automatic text categorization. [14] Shows a sentence scoring approach to select a suitable sentence based on their rank.

9. CONCLUSION & FUTURE WORK

The most important work is that we have implemented an improved methodology, which analyzes the document, and translate into UNL graph. We have introduced UNL as a language for expressing knowledge and information that can be described in natural language text.

We also proposed method to find summarized document from UNL graph. The advantage of our method when we compared to others is UNL overcomes language barriers. UNL is language dependent. So the given document which is expressed in UNL can be transformed to other language providing UNL knowledge base for that language.

The future work of this project is to develop a tool for evaluation and update the UNL Dictionary with the use of Root words found by Morphological Analyzer. And also identify more UNL relations by adding more heuristic rules.

10. REFERENCES

[1] Book-Natural Language processing by Akshar Bharati and Rajeev Sangal

2]http://en.wikipedia.org/wiki/Universal_Networking_Language

[3]http://en.wikipedia.org/wiki/Automatic_summarization

[4] Shanmugasundaram Hariharan , "Extraction based multi document summarization Using single document summary Cluster" , Advance Soft computing Application of vol.2/vol.2.1.1.March 2010. ISSN 2074-8523, ICSRS publication.

[5] Bracewell,D.B., Kuriowa,S., Ren,F.," Multilingual single document keyword extraction for information retrieval "IEEE proceeding of International conference on Natural Language Processing and Knowledge Engineering, 2005. 517 – 522

[6] Amitabhu Mukerjee, Achla M Raina, Kumar Kapil, Pankaj Goyal, Pushpraj Shukla, Universal Networking Language- A tool for language independent semantics?" International conference on the convergence of Knowledge, 2003.

[7]Martins, C.B., and Rino, L.H.M. (2002),"Revisiting UNLSumm Improvement through a case study", Workshop on Multilingual Information Access and Natural Language Processing, IBERAMIA'2002. (ISBN 84-607-6057-X)

[8] Jade Goldstein, Vibhu Mittal, Jaime carbonell, Mark Kantrowitz,"Multi-Document Summarization by sentence Extraction",Workshop on Automatic Summarization,ANLP/NAACL'2000.

[9] Gragomir R.Radev, HongYan Jing, Malgorzata Budzikowska,"Centroid-based Summarization of multiple documents: Sentence extraction, Utility-Based evaluation, and User-studies", Citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.4383.

[10] Hien Nguyen;Santos,E; Russell,J. "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans,2011. Volume: 41, Issue: 6, 1038-1051

[11] Soe-Tsyr Yuan;Jerry Sun " Ontology-based Structured cosine similarity in document Summarization: with applications to mobile audio-based knowledge management", IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 2005. Volume:35, Issue:5, 1028-1040

[12] Xiaohua Hu; Park, E.K.; Xiaodan Zhang, "Microarray Gene Cluster Identification and annotation Through Cluster Ensemble and EM-based Informative", IEEE Transactions on Information Technology in Biomedicine,2009. Volume: 13, Issue: 5, 832-840

[13]Devasena.C.L.;Hemalatha,M. "Automatic Text categorization and summarization using rule reduction", International Conference on Advances in Engineering, Science and Management (ICAESM) 2012, 594-598.

[14]Suanmali,L;Salim,N;Binwahlan,M.S.,"Fuzzy Genetic Semantic Based Text Summarization", IEEE Ninth International Conference on Dependable, Automatic and Secure Computing (DASC),2011. 1184-1191.