EXTENDED SEMANTIC NETWORKS

A Dissertation submitted to the University of Hyderabad in partial fulfillment of the

degree of

MASTER OF TECHNOLOGY

in

Artificial Intelligence

By

RAMESH NAGINA

(09MCMI15)

Department of Computer and Information Sciences

School of Mathematics & Computer/Information Sciences

University of Hyderabad
(P.O.) Central University, Gachibowli
Hyderabad – 500 046
Andhra Pradesh
India
June -2011

CERTIFICATE

This is to certify that the dissertation entitled **"Extended Semantic Networks"** submitted by **Ramesh Nagina** bearing Reg. No **09MCMI15** in partial fulfillment of the requirements for the award **of Master of Technology** in **Artificial Intelligence** is a bonafide work carried out by him under my supervision and guidance.

The dissertation has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Prof. P. N. Girija
Supervisor
Dept. of CIS
University of Hyderabad
Hyderabad

Head of the Department
Dept. of CIS
University of Hyderabad
Hyderabad

Dean of the School
School of MCIS
University of Hyderabad
Hyderabad

DECLARATION

I **Ramesh Nagina** hereby declare that this Dissertation entitled **"Extended Semantic Networks"** submitted by me under the guidance and supervision of Professor **P. N. Girija** is a bonafide work. I also declare that it has not been submitted previously in part or in full to this University or other University or Institution for the award of any degree or diploma.

Date: Name: Ramesh Nagina

Signature of the Student

Regd. No. 09MCMI15

ACKNOWLEDGEMENTS

There are a lot of people without whose support: physical, technical and moral, the project could possibly have neared completion .Though I would have liked very much to do so, it is unfortunately not feasible to mention all of them individually here, as that would probably occupy half of this report

I feel great to offer my sincere thanks to **Prof. P. N. Girija** madam, who gave me an opportunity to work under her guidance, I thank her for her support and encouragement to complete our project

I would like to express my gratitude to **Prof. C. R. Rao,** Head of the Department of Computer and Information Sciences and **Prof. T. Amaranath**, Dean of School of Mathematical and Information/Computer Sciences for providing the facilities required to continue my work

I wish to extend my thanks to AI Lab Staff for their co-operation .I am thankful to all my friends for their Encouragement and suggestions

I finally convey my heart full thanks to my family who has given me great encouragement and motivation

Ramesh Nagina

Abstract

Now a day's lot of information is being published in the internet. The Information in the internet is stored in various forms. It is a laborious process to search the information in the internet. The existing search engines and methods couldn't achieve the accurate results needed by the user. There is a need for better and efficient knowledge representation techniques .One of such technique is Extended Semantic Network. It is a new tool for knowledge representation and ontology construction. It looks for sets of associations between nodes semantically i.e. looks for the meaning and proximally, which is better in performance when compared with the present method of keyword association. It is a model with good accuracy and minimum human intervention, it is developed using the algorithms like k-means, PCA, word association and porter stemming. The tool thus developed is able to understand the human mind and has the ability of human reasoning.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Problem definition

In the recent years there is a wide usage of the World Wide Web for the information. The World Wide Web is playing a crucial role in one's life. Every individual is using the World Wide Web to obtain the required information .The information that is available in the internet is large. Searching the information in internet has become a laborious process for an individual because of various forms of storage of information. There are many techniques and methods that help the users to find the relevant information but there ability is limited to specific tasks. The most recently used method is document retrieval method based on the key word search.

1.2 Keyword Search:

In this method the user enters the keywords about which he needs the information and the documents consisting of such keywords are retrieved. It is the responsibility of the user to find the information needed from the set of documents retrieved.

The disadvantage of the keyword search is the documents that consists if the Key words entered in the query terms are only retrieved .It doesn't retrieve the information that contains the words that are related to the query terms semantically.eg. if one enters the query term 'automobile' only the documents containing the word 'automobile' are retrieved the documents containing the words 'Benz' 'Volvo' are not retrieved which are also the major automobile manufacturers.

This takes lot of time for the user because of large sets of data retrieved. In some cases the data obtained found to be irrelevant and sometimes the relevant dataset is omitted from the results. The main reason for this is most of the data available in the web is designed for the human comprehension .when using this data with machines the accurate results are not obtained without human intervention at regular intervals. The

major challenge faced by the present users of the web is imagining intelligent tools for knowledge representation and processing techniques that support and enable efficient analyzing of data by machines.

There are many researches that have taken place in this direction and one of the most important solutions to this is semantic web based ontologies that enable data understanding by machines. The main objective of this technique is to represent the data effectively so that the machine can better understand the data and enhance capture of the existing information. The main idea here is construction of the meaning related concept networks for knowledge representation i.e. enabling the machines to provide the output results of high quality with minimum human intervention.

1.3 Semantic Web

The Semantic Web is a "web of data" that enables machines to understand the semantics, or meaning, of information on the World Wide Web **[1]**. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling users to access the Web more intelligently.

There are many languages that are used to represent the semantic web some of the major languages are resource description frame work (RDF), web based ontologies (OWL).

1.4 Resource Description Frame Work (RDF)

The Resource Description Framework (RDF) is a family of World Wide Web Consortium(W3C) specifications, originally designed as a metadata data model, which has come to be used as a general method of modeling information through a variety of syntax formats**[2].**

In the RDF metadata data model the statements available in the web are made in the form of subject object predicate models these are called RDF triples. The subject denotes the resource, and the predicate denotes aspects of the resource and expresses a relationship between the subject and the object. E.g. consider the statement

"The sky has the color blue" in RDF is as the triple: a subject denoting "the sky", a predicate denoting "has the color", and an object denoting "blue".

1.4 Ontology

In a nutshell Ontologies can be defined as tools that allow storing domain knowledge in a much more sophisticated form than thesauri

- According to many dictionary definitions ontology can be defined as the science or study of being
- Specifically, a branch of metaphysics relating to the nature and relations of being
- A particular system according to which problems of the nature of being are investigated
- It is also stated as a theory concerning the kinds of entities and specifically the kinds of abstract entities that are to be admitted to a language system.

The major drawback of the ontologies is the following

- The high cost is incurred in the construction of the ontologies
- Lot of time is wasted in the construction of the ontologies
- Requires help from the domain experts in the construction of the ontologies.

To overcome the drawbacks of the ontologies construction a new technique of knowledge representation is called Extended Semantic Network is used.

Chapter 2

Review

2.1 Introduction to Extended Semantic Networks

It is a semi-automated ontology with efficiency equal to that of ontology. It is constructed using the mathematical modeling algorithms like k-means, PCA and word association and some heuristically developed methods to combine the human developed semantic networks with the automatically developed proximal network. This requires minimum time for construction, less cost and minimum human intervention. The efficiency is same as that of the traditional ontology. It is semi-automated ontology. It is a combination of automatically developed Proximal Network and human developed Semantic Network. It requires minimum human intervention and the cost in construction of the extended semantic networks is less when compared with that of the ontologies.

2.2 Proximal Network

In the past documents are classified by a person reading the entire documents. Sometimes it's a problem because one person can judge a document of one type and the other as another type. Moreover it is a laborious process to collect the documents of one type if the number of documents is more. So the machines should be employed to collect the documents of one type. The documents are classified based on the proximity between them. The machine generated results are effective and the time required is less

Proximity means the ability of a person to tell something when he is near to the object or when it is near to it. There are many measures to define the proximity such as the distance from one object to the other. The mostly used distance metric is the Euclidian distance.

The proximal network is built based on the proximity of the words, documents in that context. Proximal network is network of words in documents. It helps in the effective retrieval of the information. Proximal network is an automatically generated network. The proximal network consists of the word pairs and the value they share. The input to the proximal network is the set of the textual documents and the

output produced is a word pair with a proximity value. The documents are processed in several stages .The three stages of the proximal network are

1) Pretreatment process
2) Mathematical modeling process
3) Post treatment process

Pretreatment process: -In the pretreatment process the documents are processed and the word frequency matrix is produced

Mathematical modeling:-In this process several statistical algorithms are used such as K-Means, CA, and Word association to cluster the data.

Post Treatment process: - During this stage the result obtained from the mathematical modeling process is subjected to stemming.

The output of the Post Treatment process is the Proximal Network. It is visualized with the help of the graph editor. The links used are Uml association links.

The pictorial representation of the proximal network is show below

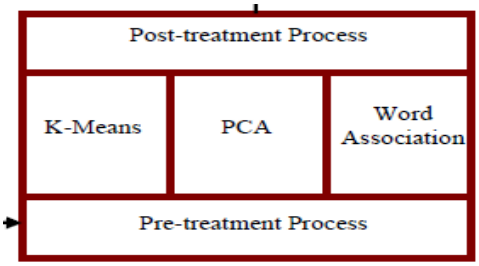


Figure 2.1: Block Diagram of Proximal Network Prototype Model

2.3 Semantic Networks

Semantic network is one of the knowledge representation techniques. It is a labeled directed graph consisting of nodes representing concepts and labeled edges representing the relation between the concepts. The semantic network consists of two parts **[3]**

1) Concept:-they are the ideas or thoughts that have meaning

2) Edges: - these are the relationships that bind the concepts

The example semantic network is shown below

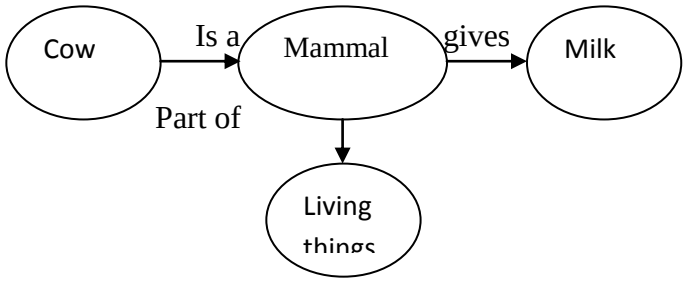


Figure 2.2: Semantic Network Example

The above graph is a semantic network with cow, mammal, milk, living things representing the concepts and the edges is a, gives, part of represents the relations they share.

The example of the semantic network is the Word Net, a Lexical database of English.

There are 6 different types of semantic networks they are

1) Definitional Networks:-this type of networks use is-a relationship between the concept and the newly defined concept
2) Assertional Networks: - In this network the information is assumed to be true unless it is marked with an external modal operator.

3) Implicational Networks: -These networks use implication relation for connecting nodes. These types of networks are used to represent patterns of beliefs, causality, or inferences.
4) Executable Networks: -These networks include some mechanism, such as marker passing or attached procedures, which can perform inferences, pass messages, or search for patterns and associations.
5) Learning Networks:- these networks are built by learning from the examples. This network may change the old network by adding or deleting the nodes or changing the values associated with the links
6) Hybrid Networks: this is a combination of the two or more networks of the above types

Thus semantic network is a system that enables capturing, storing and transferring information just as the human brain does.

2.4 Extended Semantic Network

There is a lot if information available in the World Wide Web. It is has been a great problem for the users of the World wide web to find the relevant information from the lots of information available in it.It is also a time consuming process .There is no efficient tool to retrieve the relevant information effectively. There is a need to develop techniques that enables the efficient retrieval of information as required by the user.one of such tool is the extended semantic network. The extended semantic network is a combination of the automatically generated proximal network and the human developed semantic network. The proximal network is a high recall semi-automated approach and the Semantic network is a high precision model. Thus extended semantic network is a combination of the high precision and high recall approaches.
With the help of extended semantic networks we can effectively manage and retrieve the information.

There are many methods and techniques developed by different groups and organizations but finding the right information and efficient retrieval is still a problem. Moreover

The cost of developing such tools is very high and the time required to develop such tools is more, we require the domain specialists to construct such tools. Where as in the extended semantic Networks the time required to develop this is very less since the proximal network is generated automatically using the statistical algorithms and the semantic network is developed with minimum human intervention.

The extended semantic network **[4]** is a hybrid model which just looks like ontology graph.it consists of nodes describing the concepts and edges representing the relation between them. The extended semantic networks is understandable by the machine easily it takes minimum time for the construction when compared with that if the traditional ontologies. The efficiency of this network is same as that of the ontology.

The pictorial representation of the extended semantic network is shown below

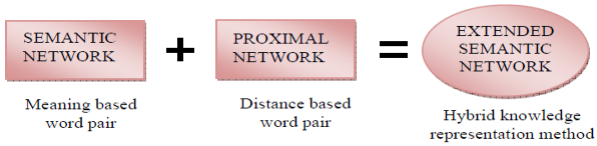


Figure 2.3: Schematic Representation of Extended Semantic Network

2.5 Human Computer Interaction

Human–computer interaction (HCI) is the study, planning and design of the interaction between users and computers. It is the combination of several fields of study such a computer science, behavioral sciences, design **[5]**. Interaction between the user and computer happens at user Interface which is both software and hardware. The output is displayed on the screen with the help of software and the user can interact with the computer with the help of hardware like keyboard and mouse. The Association for Computing Machinery defines human-computer interaction as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them." The important goal of Human computer interaction is achieving the user satisfaction.

Human computer interaction is a field which includes both the natures of human and the computer. The computer field includes topics such as computer graphics, operating system concepts, programming languages and development

environment. From the human side it includes communication theory, graphic and industrial design disciplines, linguistics, social sciences, cognitive psychology, and human factors such as computer user satisfaction are relevant. The HCI often referred as man-machine interaction.

The Major topics of HCI include

Group interfaces

These are the interfaces that allow large number of users to use a single interface. The Internet is playing a major role in this area

User Tailorability

The interfaces are as per the requirements of the end user. The end user who has great knowledge in the domain is responsible for constructing such interfaces

Embedded computation

Now a days computation is passing beyond desktop .it is transferred to application like mobiles, calculators etc. The Interfaces are designed as per the requirements of the user for these devices. It is called Embedded Computation

Augmented reality

Making a real world in computer is defined as an augmented reality .This feature helps to build the prototype of the projects.

The basic goal of HCI is to improve the interactions between users and computers by making computers more usable and receptive to the user's needs. Specifically, HCI is concerned with:

- Methodologies and processes for designing interfaces (i.e., given a task and a class of users, design the best possible interface within given constraints, optimizing for a desired property such as learnability or efficiency of use)

- Methods for implementing interfaces (e.g. software toolkits and libraries; efficient algorithms)
- Techniques for evaluating and comparing interfaces
- Developing new interfaces and interaction techniques
- Developing descriptive and predictive models and theories of interaction

Chapter 3

Design chapter

3.1 Proximal Network prototype

The proximal network prototype consists of the three stages they are

1) Pretreatment process
2) Mathematical modeling process
3) Post treatment process

3.1.1 Pretreatment process: - in the pretreatment process the textual documents are produced as input to the java program where the stop words which have less importance are removed ant the term frequency matrix is produced. The term frequency matrix is a matrix in which the rows represents the words that occur in the documents and the columns represent the documents the value of the cell is the number of times a word occurred in the particular document.

The pictorial representation of the pretreatment process is shown below

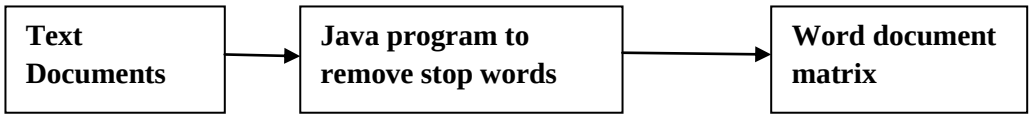


Figure 3.1: Proximal Network Pretreatment Process

The example of the word frequency matrix is shown below

Word occurrence Matrix

Doc / Word	1	2	3	4	5	6	7	8	9	10
A	1	5	6	0	0	6	5	4	10	2
B	0	4	7	14	2	3	5	7	1	2
C	16	4	6	0	0	0	0	0	0	3
D	5	5	4	4	4	6	9	4	8	6
E	1	1	1	5	5	5	5	4	19	2

Figure 3.2: Word Document Matrix

The main aim of the pretreatment process is to remove the words that have less influence while retrieving the information and make the input for the mathematical modeling process.

3.1.2 Mathematical modeling process:-

The mathematical modeling process main aim to take the input as the word frequency matrix and process it using the three clustering algorithms and produce the word pairs with the proximity value as the output .The three algorithms employed during the mathematical modeling process are.

1) K-means
2) Principal component analysis
3) Word association

3.1.2.1 K-means:-

K-means is one of the most famous clustering algorithms. It is an algorithm used to classify or to group objects into K number of group. K is a positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid **[6].**Thus the most important purpose of K-mean clustering is to classify the data. K-means utilizes the Euclidian method to calculate the number of clustering required and to decide which data falls into what cluster?

The k-means clustering is an unsupervised clustering algorithm the main idea behind the k-means clustering is to classify the data provided as the input into k clusters that are fixed priory.

Initially the centroids are chosen for k-clusters. The centroids should be chosen such that they should be far away from each other. Next the distance between the data points and the centroids are calculated and they are grouped in to k-clusters depending on the distance from the centroids. This procedure is repeated to attain the accuracy this is because some data points change from one cluster to the other cluster after each iteration .So to make the data points standard in a cluster this procedure is carried out.

The picture represents the flow chart of the k-means algorithm

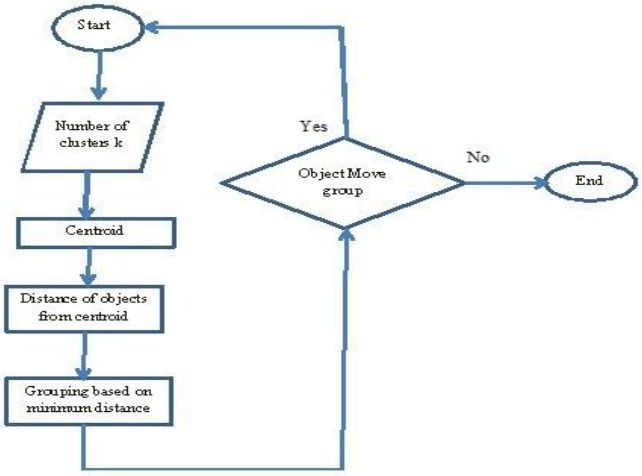


Figure 3.3: Flow chart for K-means clustering

As shown in the above diagram the number of clusters are initially defined and the centroids are selected. The distance between the data points and the centroids are calculated and the clustering is performed

The k-means algorithm uses the Euclidian distance as the distance metric. Euclidian distance between the point X(x_1, x_2, etc.) and Y (y_1, y_2, etc.) is defined as follows

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$ ----------eq(1)

The word document matrix obtained in the pre-treatment process is fed as the input to the k-means algorithm. The k-means algorithm performs the iterations until the word entities form the stable clusters .The following algorithm is employed for k-means clustering

Input:- word frequency table T

Output:-word pair table with proximity value

1)clusters = n

2)intilaise the centriods for n clusters

3) for all (words w € T) do

While (unstable(cluster))do

Distance (w,centroid)

end

end

4) for all word pairs (w1,w2)€ same cluster

Value=1

Else

Value=0;

5)Store the result in database

In the above algorithm the data stored in the database i.e. word frequency matrix is given as input to the above algorithm .and the k-means clustering is performed on the word frequency matrix and the clusters are obtained.

If the two words belong to the same cluster then the proximity value 1 is given to the word pair otherwise the value 0 is given to the word pair.ie. Two words belonging to the same cluster are considered as proximally good match and the words entities of different cluster as a bad match.

Thus the result of a k-means algorithm is a word pairs with a proximity value either 1 or 0.

3.1.2.2 Principal Component Analysis

PCA is a technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA is used for making exploratory data analysis and predictive models. PCA generally involves the calculation of the eigenvalue decomposition **[7]** of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute.

PCA is used to identify the patterns in a data and with the help of these patterns similarities and dissimilarities between data can be found. PCA is used to find the patterns in

high dimensional data which is a complex problem. It is a powerful tool for the high dimensional data.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms **[7].**

PCA is used for the dimensionality reduction of a dataset by considering the most important features that contribute to the variance and removing the high dimensional features. i.e. it considers the lower order components which contains most important aspects and removing the higher order components.

The following mathematical procedure is followed during the principal component analysis

Let X^T be a data matrix such that the mean of the dataset is subtracted from every point in data set.

The PCA transformation is given as the following

$$Y^T=X^TW \text{----------eq (2)}$$

$$=V\sum^T \text{----------eq (3)}$$

Where W= It is an m x m matrix of eigenvectors of XX^T,

$\sum$= It is an m x m rectangular diagonal matrix with nonnegative real numbers on the diagonal.

V= It is an n x n matrix.

The following algorithm is employed for calculating the PCA

Input: - word frequency table;

Output: - word pair with proximity value

1. Calculate the covariance table for matrix
2. Find out the principal components of the matrix and perform clustering

3. Calculate the distance between the word pairs
4. Filter the values
5. The result is stored in the database.

In the above algorithm the initially the word frequency matrix from the pretreatment process is given as input to it. In the next step we calculate the covariance matrix for the matrix.

In the step 2 we employ the principal component technique to find out the principal components and perform the clustering. In the step 3 we calculate the distance between the word pairs and we assign the proximity value as the inverse of the distance. In step 4 we filter the values that are less than 0.25 because they have less effect in information retrieval.

Finally the result is stored in the MySQL data base for input to the next process.

3.1.2.3 Word Association

Word association is a method in which a person says a word and the next person says the word whatever comes to his mind that is related to it.This method is used to find how the parts of a human mind works. This method can be used to find the proximity between the word entities.

Word association **[8]** is a popular game in which the word is chosen randomly. The player announces the word whatever that comes to his mind that is related to the word to the other people playing the game. The next one says the other word. This game continues till a threshold value fixed priori is reached.

According to the law of contiguity, the association strength between two words is the relative frequency of the two words occurring together in the documents. With this law it can be used to predict the proximity between the word pairs in a document.

In this word association algorithm the co-occurrence of a word pair in the set of considered documents is analyzed .the proximity value is assigned to the word pair based

on the simple mathematical calculations. The following algorithm is used for the word association.

Input: - word frequency table

Output: - word pair with proximity value

1) for all i=1 to n

For all j=1 to n

If (m_{ij}>m_{kj})

Value=$m_{jk}^2/m_{ij}*m_{jk}$

else

Value=$m_{ij}^2/m_{ij}*m_{kj}$

end

End

2) Store the word pair and the value in the database

The proximity between the word pairs is calculated by the above algorithm .m_{ij}, m_{kj} represent the position of the words. The proximity value is calculated as per the algorithm and the results are stored in the MySQL database for the next stage of the proximal network

The mean of the three statistical algorithms that are employed in the mathematical modeling stage is calculated and stored in the database for the post treatment process. This completes the mathematical modeling process.

3.1.3 Post treatment process

This stage is the final stage of the proximal network construction. The output table from the pretreatment process is taken as the input to the Post treatment process. The table is subjected to the stemming algorithm in this stage. The stemming algorithm used here is porter stemming algorithm. In stemming algorithm the words are reduced to their root words. The main aim of this process is to reduce the size of the database and for the effective retrieval of the information.

3.2 Semantic Network Prototype Model

Semantic network is a labeled directed graph with nodes representing the concepts and edges representing the relation between the concepts. Semantic Network prototype is built with human intervention. The semantic network is used to increase the efficiency of the model

The domain that is considered for the development of the semantic network is human computer interaction. The first step in the construction of semantic network consists of identifying the important concepts in the semantic network with the help of the domain experts.

Initially 144 concepts are identified and they are linked with the help of the relations. The relation links used in the construction of the semantic network are the UML links **[9]** which are explained below

3.2.1 Instantiation Link:-

This relational link is used to express the meaning "instance of it". Generally instantiation link is used to represent a concept which is an instance of the other concept. The following diagram depicts the representation of the link.

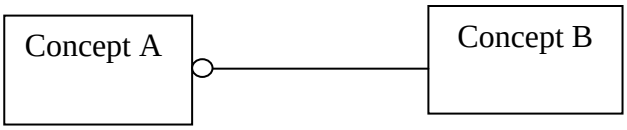


Figure 3.4: General Representation of Instantiation Link

In the above picture concept B is an instance of Concept A. The instantiation link is drawn as line with a circle at the end towards the super class.to understand the instantiation link consider the following example.

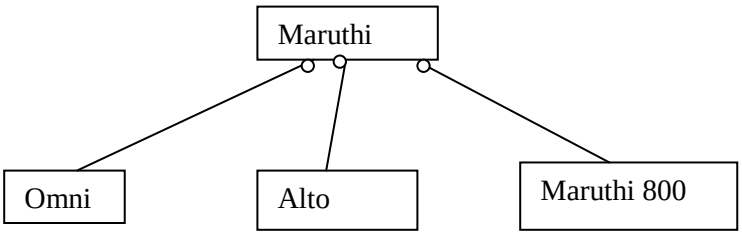


Figure 3.5: Example of Instantiation Relationship

In the above example the Maruthi is the superclass and the Omni, Alto, Maruthi 800 are the instances of the Maruthi.

3.2.2 Composition Link:-

Composition link is a form of aggregation and it is described as a "whole part" relationship. It consists of the Owner and the parts. The composition link is used to describe the relation between all the objects that constitute to form a single object. Even if one part is deleted the whole concept gets affected. The composition link in the semantic network relates the group or single concept to the other concepts .It defines the "part of" relationship. The following diagram represents the representation of composition link

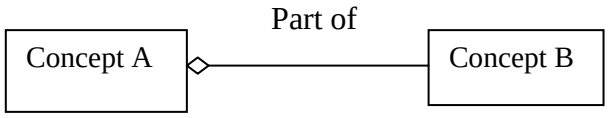


Figure 3.6: General Representation of Composition Link

The composition link is drawn as the solid filled diamond and on end of the line in the above figure the concept B is Part of Concept A.Consider the following example

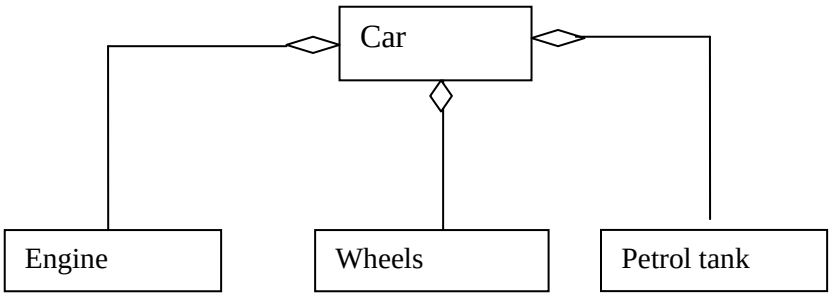


Figure 3.7: Example of Composition Link

In the above example the car is the main object the parts of the car are Engine, Wheels and Petrol tank, without these the concept of the car is affected.

3.2.3 Inheritance Link:-

This link is used to inherit the properties from the parent to offspring. Inheritance is defined as "Is a" relationship. For example consider the concept animal and dog.in this case the dog is an animal.it inherits all the properties of the animal such as it has fourlegs,it eats shouts etc.in the inheritance relation the properties owned by the superclass are transferred to their children. In the semantic network the multiple-inheritance is also used in which a class can inherit properties from more than one class. The following is the general representation of the inheritance.

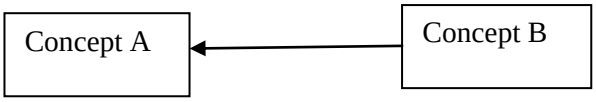


Figure 3.8: General Representation of Inheritance Link

In the above diagram the concept B inherits all the properties of Concept A. The inheritance link is a solid line with arrow head pointing towards the Parent.

Consider the following example.

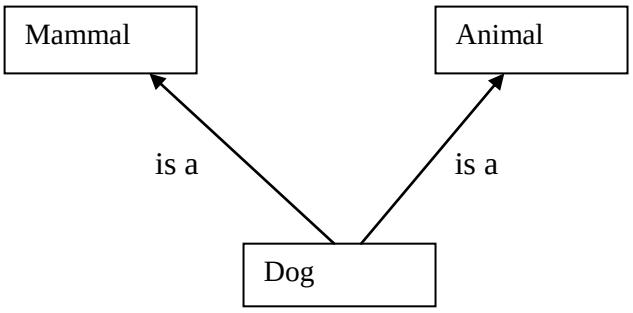


Figure 3.9: Example of Inheritance Link

In the above dog inherits the properties from the two super classes Mammal and Animal. i.e. Dog inherits all the properties of an animal such as four legs,

shouts, eats etc. as well as the properties of the mammal which feeds their Childs with the milk. In the construction of the semantic network this link is used to inherit the properties of a concept to its children. This is how the inheritance link is used in construction of the Semantic network prototype Model.

3.2.4 Association link:-

The association link is used to represent a simple relationship between two concepts. If the two concepts are related with an association link refers to the concepts communicate by sending the messages. The following diagram represents the general representation of the Association link

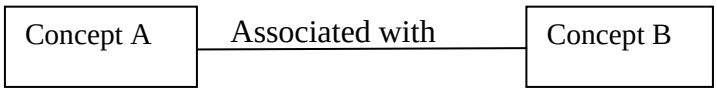


Figure 3.10: General representation of Association Link

In the above diagram concept A is associated with the concept B. The association link is a solid straight line between two nodes .the following example depicts the use of association link.

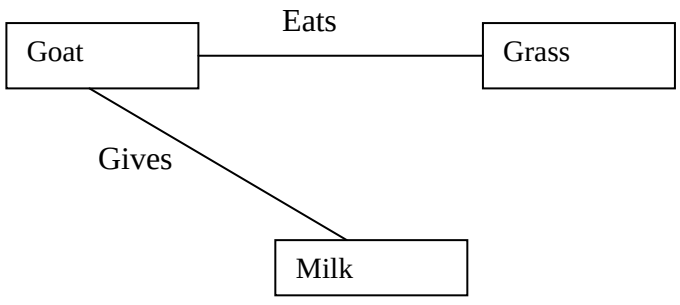


Figure 3.11: Example of Association Link

In the above example the concepts Goat is related to the Milk with association relationship gives and it is related to the concept grass with the association link eats. Thus the above diagram illustrates the use of the association link.

3.2.5 Determining the values

The links are assigned values on the random basis the Composition link is assigned a value of 0.85 followed by the instantiation link with the value 0.80 and the inheritance link with the value 0.75.

3.2.6 Semantic Network design

Semantic network is designed by considering the important concepts in the Human computer interaction domain. The 144 concepts are identified as the first step in the construction of the Semantic network. The human computer interaction is considered as the core concept of the Network. It is then divided in to the sub concepts. This stage completely involves the human. It is the responsibility of the network designer to gather all the related concepts of the domain and establish the relations between them.

After identifying the important concepts of the domain, the relations between the concepts are established carefully by using the links mentioned above. Initially the aggregation link is used to derive the sub concepts from the human computer interaction domain because these constitute the most important topics of the human computer interaction domain. The main concepts of semantic network domain include foundations, design process, models and extra. The topics are subdivided to form the semantic network by connecting them with the appropriate relational links.

The semantic network is the central core of the extended semantic network, so it is constructed by carefully identifying the important concepts in the human computer interaction domain and relating them with the help of the relational links. The semantic network is stored in the database for the further use. It is stored as the word pairs and a value is assigned to the word pairs which is the value of the relational link that is used to join the word pairs.

The concepts of the semantic network are classified in to following four categories they are

1) Foundations: - All the basics concepts and the introduction concepts of human computer interaction are grouped in to this category.

2) Design process:-the interaction designs, rules, software process, evaluation techniques, universal design are grouped in to this category

3) Models:-the concepts such as cognitive models, socio-organizational issues and stakeholder requirements.etc are grouped in to this category

4) Extra: - the extra concepts such as World Wide Web, ubiquitous computing, hypertext are grouped in to this category

Thus all the concepts are grouped and the relations are established between them to complete the construction of the semantic network

The data from the MySQL is used to view the semantic network using the graph editor.

3.3 Extended semantic Network prototype

The main idea in developing the extended semantic network is to overcome the problems faced by the present world in information retrieval and information classification. It is a hybrid which is the combination of the two different models i.e. proximal network which is developed using the mathematical models and semantic network which is developed by the humans. Thus extended semantic networks have advantages of both the semantic network and the proximal network.

The main advantages of the extended semantic network are the time required for the construction of the network is less when compared with other techniques. It also requires minimum cost for the construction of the network and also requires minimum human intervention. The efficiency is same as that of the ontologies.

The Extended semantic network is a three phase approach in the first phase the documents relating to the field of the human computer interaction are collected and processed using the mathematical algorithms to form the proximal network. In the second phase the semantic network is constructed with the help of domain experts and the final is carefully integrating the semantic network with the proximal network.

The construction of the extended semantic network consists of the semantic network as the central core of the network the reason for semantic network being the central core is the semantic network is constructed with the help of the domain experts it is very reliable. It is very efficient while searching with the help of search tool. It is combined with the proximal network by identifying the common nodes between the both networks and extending it with the nodes of the proximal network.

The technical design process is started by considering the output tables of the semantic network and the proximal network. The semantic network table is copied in to the extended semantic network table. After the completion of this process we start the process of combination of the semantic network with the proximal network by identifying the common words between them.

We start with the first word in the first row of the extended semantic network and find the matching word in the proximal network output then carry out the breadth first search and add the nodes to the extended semantic network table i.e. they are stored as a word pair entity with the proximity value they share.

Certain constraints are given while constructing the extended semantic network such as each node is assigned a level so that no relation can be drawn from the lower level node to the upper level node but vice versa is possible. If the word pair in the extended semantic network has two values then the average of the values is considered.

The following algorithm is used to combine the Semantic network with the proximal network to form the extended semantic network

Input:-output tables of semantic network and proximal network

Output - table of extended semantic network

Step1:-Copy the entire semantic table in to the extended semantic network table

Step2:-Identify the common word entities between the proximal network and Semantic network

Step3:-Add such nodes to the output table and carry out the breadth first search and add nodes of the proximal network to extended semantic network

Step4:- Store the output in the MySQL database.

The Extended semantic network is visualized using the Graph Editor.

3.5 The HCI platform

The HCI platform consists of the user interface and search tool that is developed for the end users to facilitate them to retrieve the information needed by them .It also enables the users to upload for the documents that can be used by the other users of the platform .The user interface is developed using the php which is a scripting language for the web and mysql.

The users are provided with the username and password to access this platform so that they can search for the related information

The different modules involved in the user interface are

1) Login screen: - This enables only the authorized users to use the platform restricting the unauthorized users. It consists of the username and password fields
2) Search tool: - This enables the users to search for the information. Here the user enters the key word regarding which he needs the information. The tool searches for the information by using the extended semantic networks and retrieves the related information to the user.
3) Uploading: - With the help of this feature the users can upload the documents relating to the human computer interaction design. Only the textual documents are allowed to upload for the platform
4) Registration: - this enables the new users to register for the platform .it includes the fields such as first name, date of birth, desired username, password etc.
5) Change password: - this feature enables the already registered users to change their passwords.

With the help of this platform the users can effectively share the information related to the human computer interaction domain .This is a rich interaction platform for the users .This enables the user to retrieve the information they need and also enables them to download the information.

The search tool uses the Extended semantic Network to find the related information so that the information required by the user is retrieved .the search is carried semantically i.e it is carried as per the user requirements

3.6 Graph editor

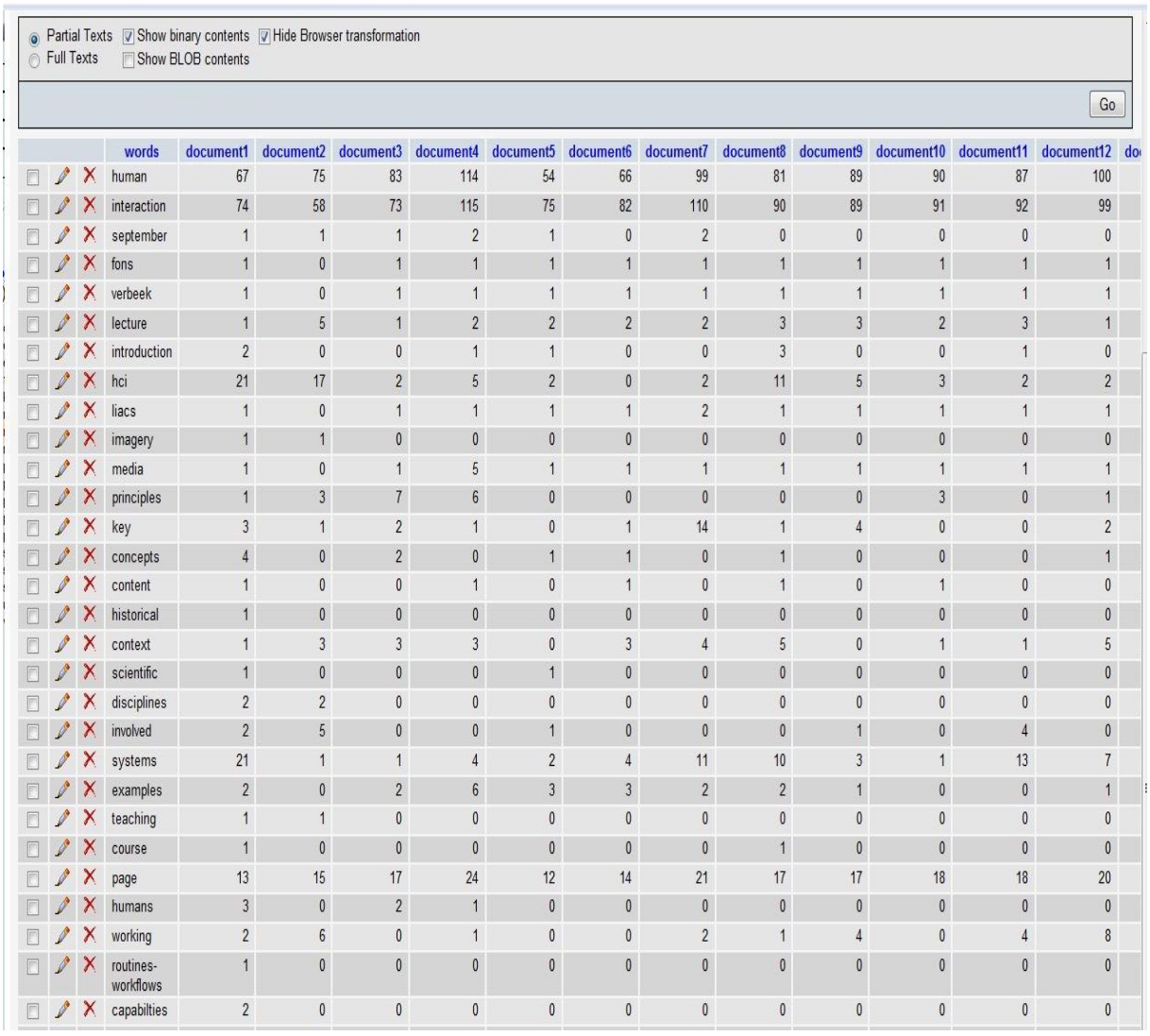
Graph editor is designed with the java language it is used to view the semantic network. The nodes in the graph are represented with the rectangle and the links are used to connect the nodes, this graph editor is used for the easy understanding of the networks and the relations between the nodes in the network

Chapter 4

RESULTS

4.1 word occurrence matrix

The output of the word occurrence matrix is shown below it is obtained by removing the stop words from the documents the output is shown below. The rows represent the words and the column represents the documents. The value in the cell is the number of time the word occurred in a document



Partial Texts Show binary contents Hide Browser transformation
Full Texts Show BLOB contents

Go

words	document1	document2	document3	document4	document5	document6	document7	document8	document9	document10	document11	document12
human	67	75	83	114	54	66	99	81	89	90	87	100
interaction	74	58	73	115	75	82	110	90	89	91	92	99
september	1	1	1	2	1	0	2	0	0	0	0	0
fons	1	0	1	1	1	1	1	1	1	1	1	1
verbeek	1	0	1	1	1	1	1	1	1	1	1	1
lecture	1	5	1	2	2	2	2	3	3	2	3	1
introduction	2	0	0	1	1	0	0	3	0	0	1	0
hci	21	17	2	5	2	0	2	11	5	3	2	2
liacs	1	0	1	1	1	1	2	1	1	1	1	1
imagery	1	1	0	0	0	0	0	0	0	0	0	0
media	1	0	1	5	1	1	1	1	1	1	1	1
principles	1	3	7	6	0	0	0	0	0	3	0	1
key	3	1	2	1	0	1	14	1	4	0	0	2
concepts	4	0	2	0	1	1	0	1	0	0	0	1
content	1	0	0	1	0	1	0	1	0	1	0	0
historical	1	0	0	0	0	0	0	0	0	0	0	0
context	1	3	3	3	0	3	4	5	0	1	1	5
scientific	1	0	0	0	1	0	0	0	0	0	0	0
disciplines	2	2	0	0	0	0	0	0	0	0	0	0
involved	2	5	0	0	1	0	0	0	1	0	4	0
systems	21	1	1	4	2	4	11	10	3	1	13	7
examples	2	0	2	6	3	3	2	2	1	0	0	1
teaching	1	1	0	0	0	0	0	0	0	0	0	0
course	1	0	0	0	0	0	0	1	0	0	0	0
page	13	15	17	24	12	14	21	17	17	18	18	20
humans	3	0	2	1	0	0	0	0	0	0	0	0
working	2	6	0	1	0	0	2	1	4	0	4	8
routines-workflows	1	0	0	0	0	0	0	0	0	0	0	0
capabilties	2	0	0	0	0	0	0	0	0	0	0	0

Figure 4.1: Snapshot of Word Document Matrix

4.2 k-means

The output of the k-means clustering that produces the word pair entity with the proximity value is shown below

word1	word2	value
human-computer	quick	0
human-computer	understanding	1
human-computer	jeff	0
human-computer	raskin	0
human-computer	interface	0
human-computer	relaxed	0
human-computer	forgiving	0
human-computer	rewired	0
human-computer	shortcomings	0
human-computer	able	0
human-computer	cope	0
human-computer	useful	1
human-computer	prototype	1
human-computer	visibility	1
human-computer	controls	1
human-computer	effects	0
human-computer	sensible	0
human-computer	affordance	1
human-computer	sort	0
human-computer	operations	1
human-computer	manipulations	0
human-computer	object	1
human-computer	orientation	0
air	handling	1
human-computer	reason	0
human-computer	internalize	0
human-computer	phones	0
human-computer	sorts	0
human-computer	particular	1
human-computer	constraining	0
human-computer	psychology	0
human-computer	things	0
human-computer	poet	0
human-computer	restricting	0
human-computer	way	1
human-computer	design	0
human-computer	de-activation	0
human-computer	greying-out	0
human-computer	originally	0
air	house	0
human-computer	gibson	0

Figure 4.2: Snapshot of Output of K-means Algorithm

4.3 Principal Component Analysis

The result of the principal component analysis is shown below

word1	word2	value
verbeek	flavors	0.301511344577764
verbeek	reward	0.301511344577764
verbeek	targeted	0.301511344577764
verbeek	highlight	0.301511344577764
verbeek	motivate	0.301511344577764
imagery	explicit	0.5
verbeek	bots	0.301511344577764
verbeek	avatar	0.301511344577764
verbeek	user-facing	0.301511344577764
verbeek	inferring	0.301511344577764
verbeek	vcs	0.301511344577764
verbeek	coach	0.301511344577764
verbeek	autonomy	0.301511344577764
verbeek	agent	0.301511344577764
verbeek	accord	0.301511344577764
verbeek	reactivity	0.301511344577764
verbeek	proactivity	0.301511344577764
verbeek	collaboration	0.301511344577764
verbeek	collaborate	0.301511344577764
imagery	danger	0.577350269189626
verbeek	mediatechnology	0.288675134594813
verbeek	connecting	0.288675134594813
verbeek	registering	0.301511344577764
verbeek	reacts	0.301511344577764
verbeek	wired	0.301511344577764
verbeek	december	0.301511344577764
verbeek	hammond	0.301511344577764
verbeek	workout	0.301511344577764
verbeek	heart	0.301511344577764
verbeek	blood	0.301511344577764
verbeek	transpiration	0.301511344577764
verbeek	inter-face	0.301511344577764
verbeek	kismet	0.288675134594813
verbeek	robot-baby	0.301511344577764
verbeek	humanoid	0.301511344577764
verbeek	learns	0.301511344577764
verbeek	human-robot	0.301511344577764
verbeek	obtaining	0.301511344577764
verbeek	ai	0.301511344577764
verbeek	bruzard	0.301511344577764

Figure 4.3: Snapshot of output of Principal Component Analysis

4.4 Proximal Network

The output of the proximal network using the graph editor is shown below

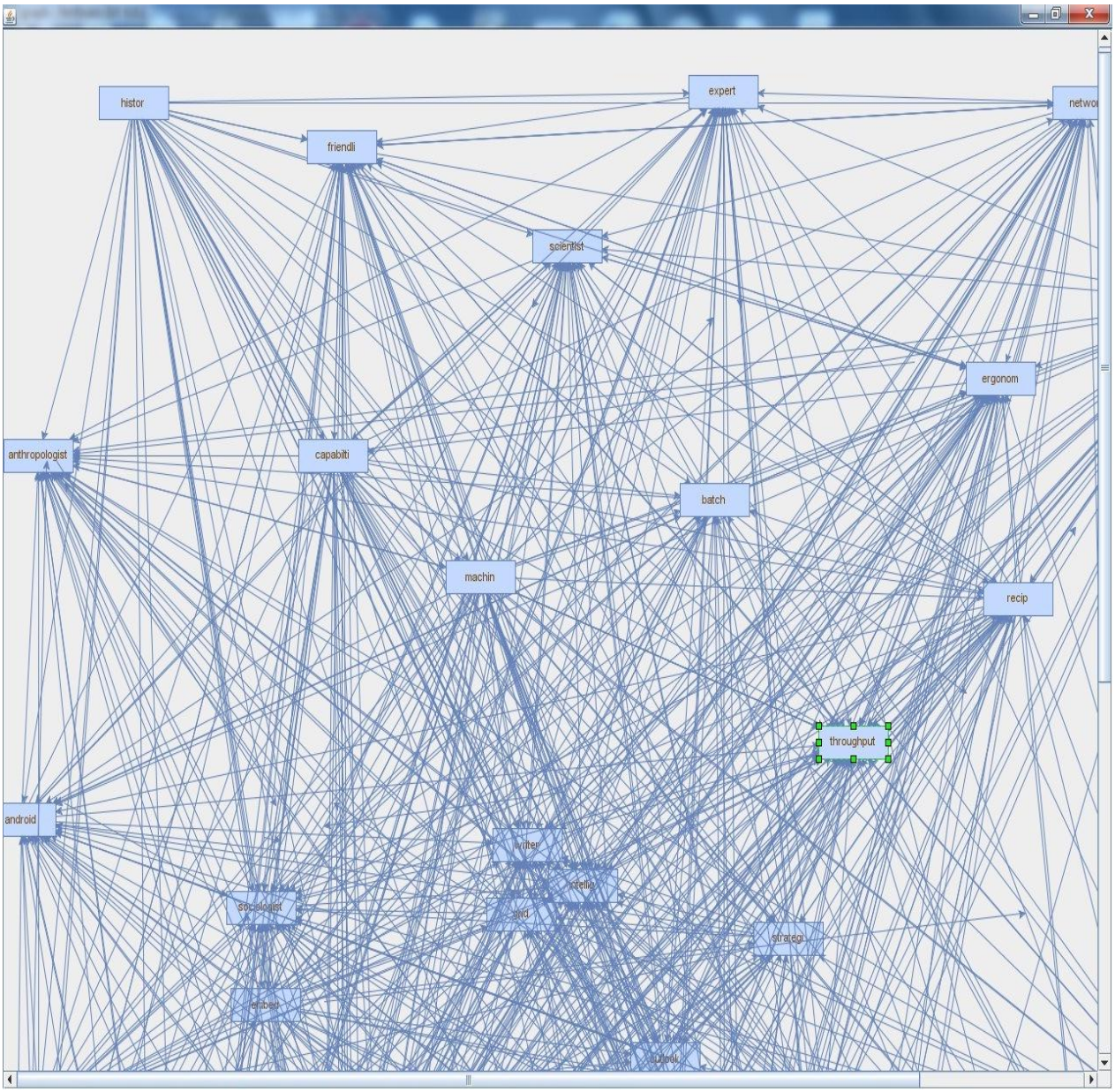


Figure 4.4: Graphical Visualization of Proximal Network

4.5 Semantic Network

The output of the semantic network is shown below.

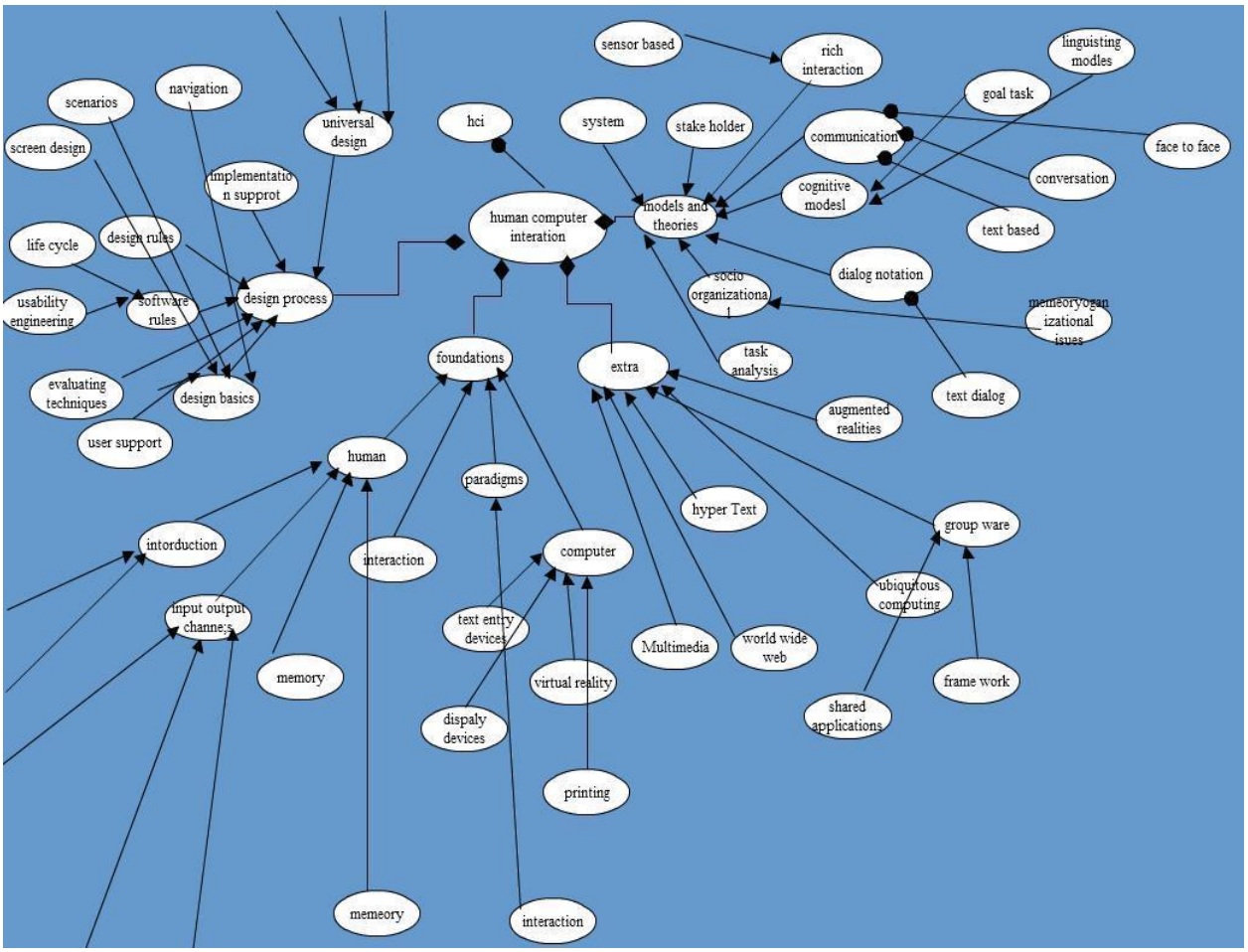


Figure 4.5: Graphical Visualization of Semantic Network

4.6 Extended Semantic Network

The output of the extended semantic network is shown below

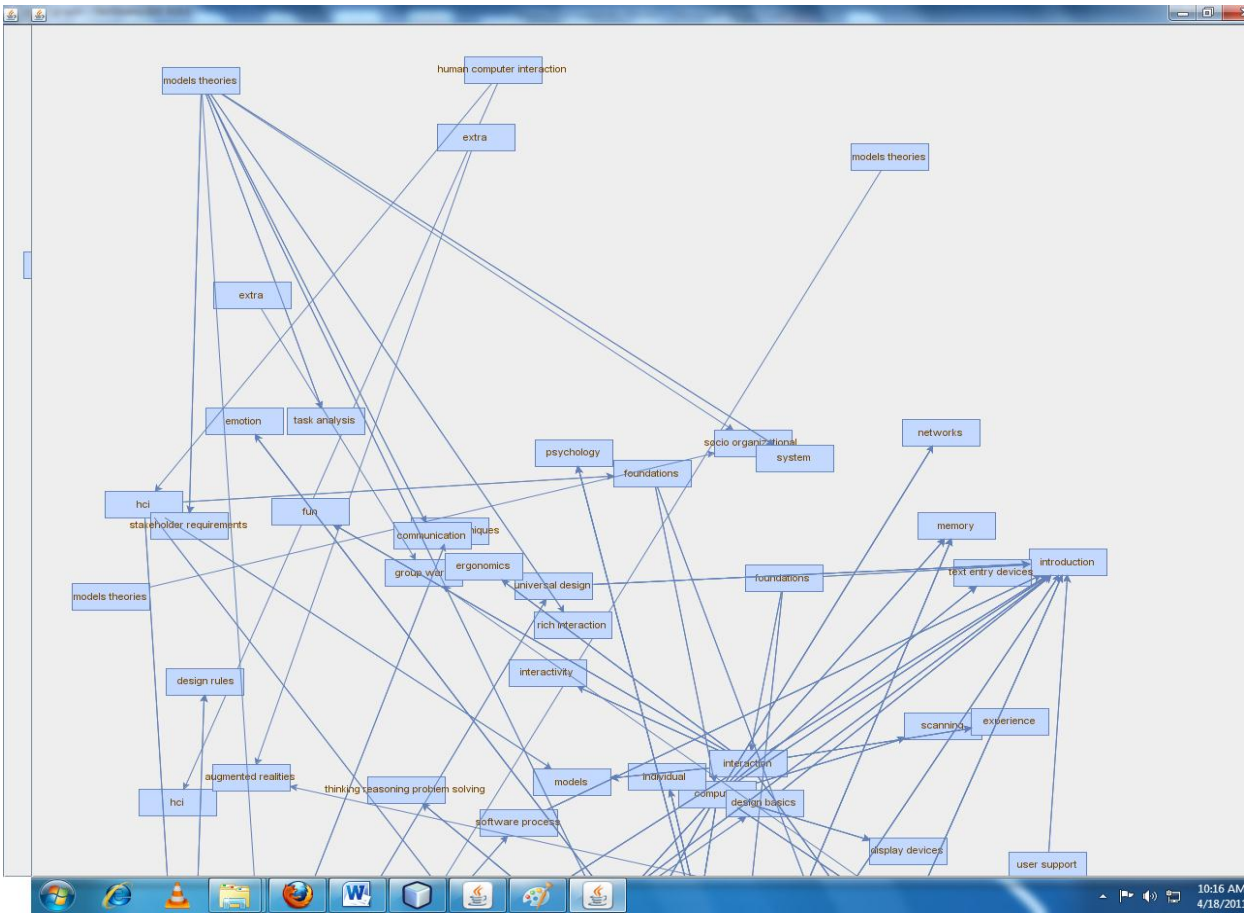


Figure 4.6: Graphical Visualization of Extended Semantic Network

4.7 HCI Platform

The output screens of the graphical interface are shown below.

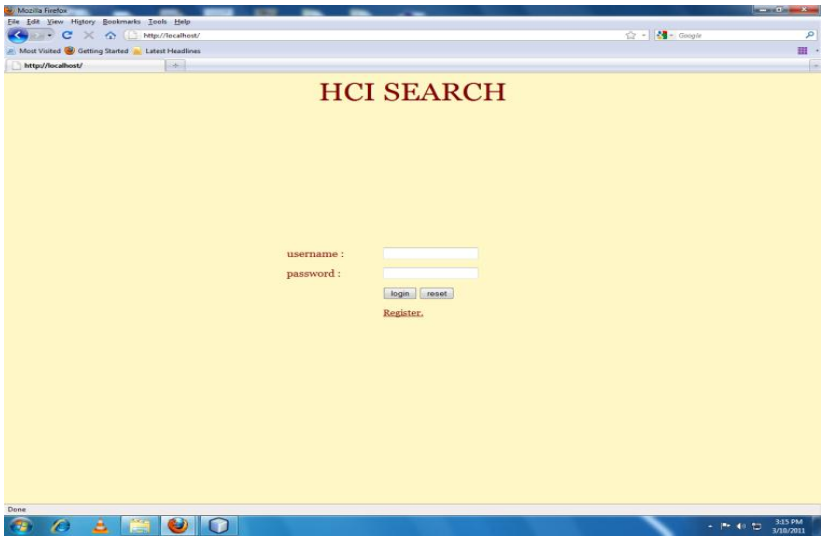


Figure 4.7: Login Form

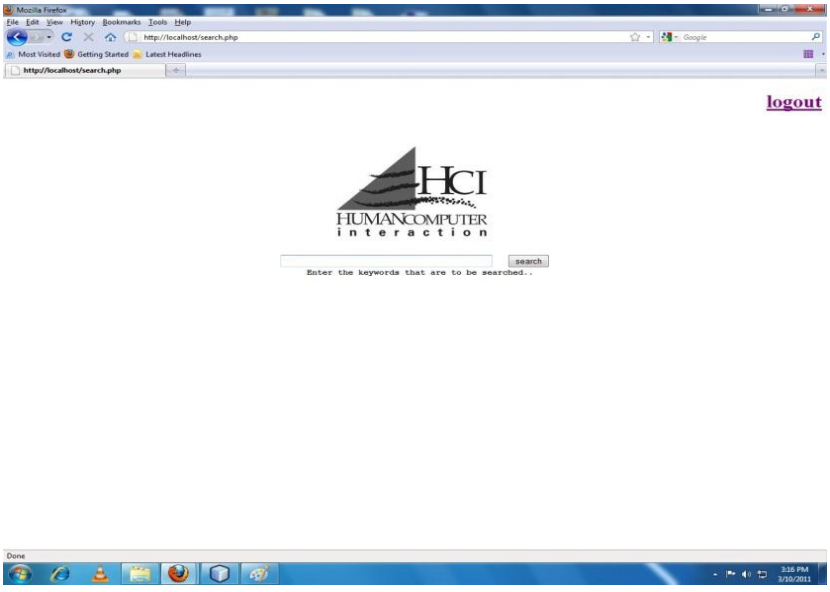


Figure 4.8: Search Screen

4.8 comparison of keyword Search and Extended semantic search

The performance of a search engine is measured in terms of the precision and recall. These are defined in terms of the retrieved documents i.e. list of documents produced by a search engine for a query and set of relevant documents i.e. list of documents that are relevant to the topic

Precision

Precision is defined as the fraction of retrieved documents that are relevant to the search. The formula for the precision is shown below.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

In other terms precision can be defined as number of correct results divided by number of all returned results.

Recall

Recall is defined as the fraction of documents that are relevant to the query that are successfully retrieved

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In other words the recall is the number of correct results divided by the number of results that should have been returned.

The Performance of the Extended Semantic Network compared with the Key word Search is tabulated below.

	Key Word Search		**Extended Semantic Search**	
No. of Documents	**Precision**	**Recall**	**Precision**	**Recall**
10	0.4	0.66	0.6	0.8
15	0.375	0.6	0.55	0.75
20	0.3	0.5	0.45	0.66
25	0.216	0.33	0.35	0.45

From the above table the precision and the recall values of Extended semantic search are greater when compared with the ordinary key word search.so we can conclude that the Extended semantic search results in the most relevant data needed by the user it is more effective tool for the information retrieval

Chapter 5

Conclusion and Future Work

Thus extended semantic network is a semi-automated ontology construction. It is the combination of two networks semantic network and the proximal network. The semantic network is constructed with the help of domain experts and the proximal network is constructed through mathematical modelling algorithms. The extended semantic network has more advantages when compared with the traditional ontology. It is easy to construct, it requires minimum time for the construction when compared with the traditional ontology .The cost incurred in the development of the extended semantic network is also less.

The extended semantic network construction is a three phase approach in which the proximal network is constructed using the algorithms such as PCA, K-means, word association and the semantic network is constructed using the links like association,inheritance,instantiation and composition and the final stage is combining both networks to form the Extended Semantic Network and a Platform is also developed for the sharing of the information among the users using php language and graph editor is developed to view the network structure.

The Extended semantic network is used for the effective information retrieval and for the classification of data it is a new technique of the knowledge representation. The future work of the Extended semantic network includes using more algorithms related to neural networks and genetic algorithms in the mathematical modelling process of the proximal network design to improve the accuracy of the clustering and developing a user aid modelling as per the requirements of the user.Finding out the new methods for combining the proximal network with the semantic network to form the Extended semantic network. Assigning the appropriate values for the relational links that are to be used in designing semantic network.

Thus Extended semantic network is a new tool for retrieving the information effectively as required by the user which has the ability of the human reasoning

References

[1] Berners-Lee, T., Hendler J. and Lassila O., "The Semantic Web", Published in Scientific American Magazine, USA, 2001.

[2] Brickley, D. and Guha, R.V., "Resource Description Framework (RDF) Schema Specification", published in Proposed Recommendation: World Wide Web Consortium, 1999.

[3] Jonh F. Sowa, "Semantic Networks", Published in Encyclopedia of Artificial Intelligence, Ed. Stuart C Shapiro

[4] Shetty R. T. N., Riccio P. M. and Quinqueton J., "Hybrid method for knowledge processing, integration and representation", IEEE-IRI '06 proceedings, Hawaii, USA, 2006.

[5] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale (2003), "Human–Computer Interaction". 3rd Edition. Prentice Hall, 2003

[6] MacQueen J.B., "Some Methods for classification and Analysis of Multivariate Observations", Published in Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:pp 281-297, USA, 1967.

[7] Jolliffe I.T., "Principal Component Analysis" , Published in Springer Series in Statistics, 2nd ed., Springer, XXIX, 487 p.28 illus. ISBN 978-0-387-95442-4, New York, USA, 2002.

[8] Packard V., "The Hidden Persuaders", Published in Penguin, paperback edition, p.129, 1961.

[9] Rational Corporation: UML Notation Guide 2, 2000.