www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 4 April 2017, Page No. 20984-20989
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i4.34

Survey on Various Retrieval Strategies and Utilities for effective Information Retrieval

Mr. Ramesh Babu Pittala[1], M.Nagabhushana Rao[2]
1 Research Scholar, Rayalaseema University, Kurnool
prameshbabu526@gmail.com1
2Professor, CSE Department, KL University, Vijayawada, Andhra Pradesh, India
mnraosir@gmail.com2

Abstract: Information Retrieval is mainly used for offering the fair information at the right time and also retrieving the necessary information in the desired form. Information systems which are strongly describes what the user need. The main intension of retrieval strategy states that the level of relevancy between the document and the query increases with increase in the number of items .The retrieval strategy are used for retrieving the information from highest ranked documents based on precision and recall. Different Retrieval Utilities are used to progress the performance of IR strategies. In this paper, specified the various retrieval strategies and utilities which are playing a major role in estimating the performance of IR.

Keywords: Information Retrieval, Retrieval Strategies, Retrieval Utilities, Similarity Measure.

I. Introduction

Text Mining is playing a major role to provide the require information to the user based upon their needs. It is the best methodology to provide the information to the user in simple manner. All the industries are automated their data to provide the information within the organization. To provide such accurate data, Information Retrieval system is used. It is a system used to store, manipulate and retrieve the information to fulfill the requirements from the user. Evaluation of IRS's is very important to provide the relevant information quickly. IRS may retrieve relevant or Non-Relevant, may not retrieve relevant or Non-Relevant information from the database [10].

Information retrieval uses a WAIS (Wide Area Information Servers) to access the terabytes of the information. "Text" Data Type is used to represent the user input and to process the information retrieval functionality. The terms user (end user seeking for the information), items (smallest complete unit), and document (Information storage) are most commonly user terms to represent the different retrieval methodologies. Information Retrieval is aimed to reduce the strain of the end user when finding the crucial information. The user can get the desired information through the other sources like newspapers, books, magazines etc. But the time consumption is high due to finding the items located at the different areas and finding the content manually.

1.1 Process of IR

User will try to optimize the time to get the needed information. Time retrieval time will be based on the time to search for the content from the system and the time to locating the needed information or knowledge from the system. The total time required to retrieve the relevant information is described as below

Step 1: Time to identify the Query - user have to describe what information he is looking for.

Step 2: Time to Execution of a Query – user input is submitting to the IR to search from the relevant data.

Step 3: Time to scan all the results of a submitted by the user to read the selected items.

Step 4: Time to Read the Non-Relevant documents from the result set.

Information Retrieval uses the software and some special hardware functions to locate the needed information. To satisfy the user requirements to retrieve the relevant items, to save processing time, the Information Retrieval is designed with various methodologies. Some of the important functionalities are given below.

1. Normalizing the Items: Accepting the user data in a standard format and converting the text into the searchable data structure by removing the common words and stop words.
2. Selective Dissemination of Information or Mail Process: It will compare the newly received items with the standing statements of user interest that matches the content before searching from the database. SDI will save the retrieving time by finding the required information from the mail files.
3. Indexing Process: Identifies the scope and visibility of the document and pace the terms in the index. Apply manual indexing also for the effective results.
4. Retrieve the results: After identifying the indexed terms, relevant results will be displayed.

1.2 Applications of IR:

Information Retrieval Systems are developed with an intention of retrieving the data or manage the information from the huge amount of the data bases. Many corporate companies, universities, public libraries, militaries and many more research organizations are using this IR to provide the accurate information with ease of access.

II. Performance Evolution:

Two important factor's Precision (PREC) and Recall (REC) [1][7] are used to estimates the ranking of the relevance items.

2.1 Recall

Recall is the fraction of the number of retrieved or fetched records to the aggregate number of significant records in the database.

2.2. Precision

Precision is the ratio of number_relevant_retrieved or fetched to the whole _ Retrieved (relevant and irrelevant).

Example:

Assume that the database contains 90 records in a specific topic. When the user searched for a topic, he retrieved 70 records. Out of the 70 Records, 45 records are relevant.

The precision and recall values are

Precision = 45 / (45+25) = 45/70 = 0.64 (64%)

Recall = 45 / (45+45) = 45/90 = 0.50(50%)

2.3 F-Measure

It enumerates the median of the information retrieval precision and recall values. It is computed using harmonic mean.

Given "M" points harmonic mean a1, a2, a3… am is

$$H = M * \sum_{K=0}^{M} X_k \quad (1)$$

So, the harmonic mean of Precision(PREC) and Recall(REC)

F-Measure = 2 *((PREC* REC)/(PREC+REC)) (2)

As per the above F-measure equation derived by van Rijsbgen (1979), Fβ "It will find the efficiency of system with respect to the items entered by the user, who assigns β times as much significance to recall as precision".

To fine the effectiveness measure

$$Eff = 1 - 1/(\Omega/P+(1-\Omega)/R) \quad (3)$$

Their relationship is

$$F\beta = 1-E \quad (4)$$

Where $\Omega = 1/(1+\beta 2)$ (5)

For each cluster we will calculate the F-measure to estimate the performance.

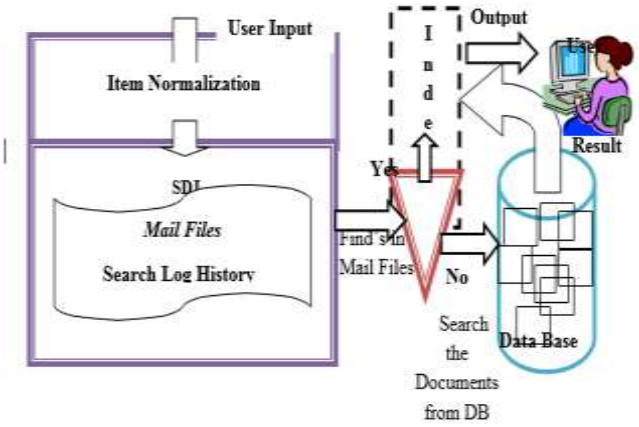

Fig 1 IR Process

III. Retrieval Strategies

Retrieval strategy is a kind of measure for computing the synonymy in a query(X) and a document(Y). The main intension of this strategy states that the level of relevancy between the document and the query increases with increase in the number of items .It is a principle which receipts a request (X) and a set of Records Y1, Y2, Y3...Yk isolates the relationship between the Query and Document SC (X, Y_k) for all the documents 1<=i<=k. In other words, similarity coefficient can be termed as "Retrieval Status Value (RSV)"

There are several types of retrieval strategies to find out SC are listed below.

3.1 Vector Space Model:

It is the most former retrieval strategy. In this model we are assuming two vectors in order to finding the similarity coefficient. This method can be applicable only when their items are expressed in the form of vectors. In this documents can be compared with respect to the similarity among their contents with the generated query items. However, a similarity coefficient can be produced by using the binary representation.

Vector space model is used for implementing vector which denotes the items in the document and the other vector denotes the term in the query. Similarity coefficient is used for defining the vectors instead of angle.

3.2 Probabilistic Retrieval:

To find the SC, probability methodology is used. This method considers each and every individual item involved in the collection (or) document collection. It is based on the fundamental approaches.

- Usage patterns to predict relevant
- Each input pattern will find whether the document is relevant or not.

This probabilistic retrieval based on the concept of estimating the relevant documents.

Types of Probabilistic Retrieval Strategies:

- Simple term weight paradigm
- Non_binary independence paradigm
- Poission paradigm
- Term components paradigm

3.3 Language Model:

Language model is constructed for each and every individual document and the possibility that the document will produce the query is calculated.

3.4 Boolean Indexing:

In this model, it assigns a score to evaluate ranks from an initial Boolean query. This can be done by combining the weight with each query term .It can be used for evaluating the similarity coefficient retrieval status value.

3.5 Latent Semantic Indexing:

In this methodology, the term document matrix is defined with the occurrences of items in a document. This matrix is used for filtering or to eliminate the noise. This Term X Document matrix is compressed via singular value decomposition (SVD). In this, the documents reflecting the same meaning will be recorded in a multi-dimensional space.

3.6 Neural Networks:

The Nodes in the network are represented as a set of neurons. These neurons get fired upon capturing by an input which generates the connections to documents. The asset of every connection in the network is inherited to the document and is placid from a document query similarity measure. These Networks are arranged by balancing the weights on link in reaction to predetermine relevancy of the documents.

3.7 *Genetic Algorithm:*

It is an algorithm which is used for generating an optimal query in order to determining the relevant documents by estimation. For this, in which we are random (or) estimated weights for the new queries. The existence is due to the closeness of the relevant documents. In which, all the less fit queries are eliminated.

3.7.1 Fuzzy Set Retrieval:

Document are graphed to a fuzzy set Boolean query operations are matched in to the operations union,

intersections and complement, which results to the effectiveness of membership grouped with the query. This result of membership associated with i^{th} query. The strength of membership associated with the query will represents the Similarity between the query and document.

3.7.2 Inference Networks:

This model decides how a document relates to a query using Bayesian network. In this the relevancy of documents can be achieved by using the "Evidence" exists in the document(Y). The outcome is reference as the similarity coefficient (SC).

Among these various retrieval strategies, finding the SC, Vector space model is described with an example. On the same way all the strategies will apply the SC and retrieves the ranked documents for desirable information.

Example:

The following example will describes to find the similarity between the query and document. To find the SC, query terms and document terms are described to understand the problem easily.

- t= Number of individual terms in the document
- tf_{ij}= Number of happening of term t_j in document y_i this is referred to as the term frequency.
- yf_{ij}= Number of documents which contains t_j this is called as document frequency.
- iyf_j=log [y/yf_j]

Where y is the total number of documents. t is the total number of terms, and the proportion between the aggregated amount of documents (y) to the document frequency(yf) is called as inverse document frequency (iyf). Which is used to computes the weights of the documents corresponding to a given term. The product of term frequency and inverse document frequency will generate the term weights in a document.

The following equation is used to find the j^{th} record in the vector equivalent to document i.

$$yf_{ij}=tf_{ij}(iyf_j) \qquad (6)$$

The product of two factors are used to find the similarity coefficient (SC)

$$SC(X, Yi) = \sum wx_j \times Y_{ij} \qquad (7)$$

Vector Y (y_{i1}, yi2, yi3…y_{if}), Vector X (w_{x1}, wx2, wx3….w_{xt})

Let us consider the following example to find out SC

X: Design computer processor

Y1: A computer contains a processor

Y2: Design of processor is good

Y3: Good processor design improves computer performance

Total number of distinct terms t=11

Term Frequency (TF)	Y 1	Y2	Y3	Document Frequency (DF)	Inverse Document Frequency (IDF)
A	1	0	0	1	0.477
Computer	1	0	1	2	0.176
Contains	1	0	0	1	0.477
Design	0	1	1	2	0.176
Good	0	1	1	2	0.176
Improves	0	0	1	1	0.477
Is	0	1	0	1	0.477
Of	0	1	0	1	0.477
Performance	0	0	1	1	0.176
Processor	1	1	1	3	0

Terms T= A, Computer, Contains, Design, Good, Improves, Is, Of, Performance, Processor

Total number of documents (Y) =3

Calculation of TF (Term Frequency), DF(Document Frequency), IDF (Inverse Document Frequency) are shown in below table.

Table 1: Calculation of TF, DF, IDF

The Weights' of the Terms in a Document are listed In Fig 1.

The Similarity Coefficient (SC) of Query and Document for the given documents are

SC(X,Y1)= (0)(0.477)+(0.176)(0.176)+(0.477)(0)+(0)(0)= 0.03

SC(X,Y2)= (0)(0)+(0)(0.176)+(0.176)(0.176)+(0)(0) = 0.031

SC(X,Y3)= (0)(0)+(0.176)(0.176)+(0.477)(0)+(0.176)(0.176) = 0.062

The SC of the values 0.031, 0.031 and 0.062 respectively. The order of retrieving the documents is Y1, Y2 and Y3 or Y2, Y1 and Y3 as finding the similarity measure in Fig 2.

IV. Retrieval Utilities:-

To enhance the performance of the Retrieval Strategy different Retrieval Utility are used, it is a technique. A utility may plug into any strategy. Among these utilities, most are used to add or remove the terms and others just purify the query by using document instead of using the total document. There are different retrieval utilities to improve the performance of the retrieval strategies. Following are some of the retrieval utilities already existed.

DOI: 10.18535/ijecs/v6i4.34

4.1 R

Table 2: Calculation of Similarity Measure

Doc Id	A	Computer	Contains	Design	Good	Improves	Is	Of	Processor	Performance
Y1	0.477	0.176	0.477	0	0	0	0	0	0	0
Y2	0	0	0	0.176	0.176	0	0.477	0.477	0	0
Y3	0	0.176	0.477	0.176	0.176	0.477	0	0	0	0.477
X	0	0.176	0	0.176	0	0	0	0	0	0

elevance Feedback:-

This is one of the most popular retrieval utility, the basic assumption is to find the relevancy in different passes. In each pass, the end user purifies or filters the input depends on generated results of the existed query. The implemented items are appended to the query based on the selection of relevant query so that the existing items can be reweighted based on the user feedback.

There is two model used for relevance feedback, one is the combining of new-fangled items to the original request and removal of items from query these comes under vector space model and the other is the probabilistic model based on reweighting existing items. The main concept of relevance feedback is that it runs a query, collect the feedback from the various user to improve the query, and repeat the outcome process.

4.1.1 Vector space model using Relevance Feedback

This method is used to finding the rankings of the various documents in the database. In this, query is denoted by vector X and document by vector Yi will measure the query and document similarity coefficient SC (X, Yi).

The equation given below forms the new query X' from the old query X.

$$X'=X + 1/d_1 \sum_{i=1}^{d1} A_i - 1/d_2 \sum_{i=1}^{d2} Bi \quad (8)$$

Ai and Bi are components of A and B.

Instead of using values d1 and d2 the arbitrary weights are used:

$$X' = \alpha X + \beta \sum_{i=1}^{d1} Ai/n1 - \gamma \sum_{i=1}^{d2} Bi/n2 \quad (9)$$

It is not important to use all the relevant or non-relevant document so in that case threshold na and nb are used therefore the equation now becomes:

$$X'= \alpha X + \beta \sum_{i=1}^{\min(da,d1)} Ai/n1 - \gamma \sum_{i=1}^{\min(db,d2)} Bi/n2 \quad (10)$$

Instead of using all the non-relevant document from the database only the top ranked non-relevant document are used.

$$X'= \alpha X + \beta \sum_{i=1}^{d1} Ai\text{-}B1 \quad (11)$$

$$X'=X + \sum_{i=1}^{d1} Ai - \sum_{i=1}^{d2} Bi \quad (12)$$

4.2 Clustering

To decrease the search space provided to react to a query, document clustering were used to group the document. Similar types of document are clustered into one. Different clustering algorithms or principle are used and clusters are designed with either a top-down or bottom-up approaches.

In top-down approach, the total collection of documents is taken as single cluster and fragmented into smaller clusters.

The reversal of top-down approach referencing as hierarchical agglomerative is termed as bottom-up approach.

4.2.1 Hierarchical Agglomerative Clustering:

The following steps are repeated until they become one cluster.

1. Excessive Synonymy of the two clusters are found.
2. Similarity is calculated and grouped as a single cluster. Identify the Similarity for the other clusters.

Assume document P, Q, R, S, T exist and document(Y)-document(Y) similarity exists. Then it becomes:

{{P}, {Q}, {R}, {S}, {T}}

We assume the superlative similarity is between P and Q then it becomes:

{{P,Q},{R},{S},{T}}

This algorithms is based on how {P} is clustered with {Q} in the initial stage. Once it is clustered, a new synonymy degree is calculated, which denotes the similarity of documents to newly created cluster {P, Q}.

4.2.2 Ward's Method:

El-Hamdouchi and Willett proposed that the clusters are generated by combing more than one document will increase the sum of the distance for every document (Y) in the database to the centroid(C) combined of the cluster containing the same document. The centroid(C) is calculated as the average vector in the vector space.

$$C_j= \sum_{i=1}^{n} t_{ij}/n \quad (13)$$

The j^{th} element of the centroid vector is enumerated as the average of all of the j^{th} elements of document vectors.

4.3 Passage based retrieval:

In this method, queries are matched to portion of documents and the result of every portion are brought into a solitary similarity coefficient (SC). The scope of each portion is either fix or diverse depends on principle.

4.4 N-grams:

To provide adaptability to noise, that means to remove the noise from the query given by the user so that it can be easy to find the relevant document, n-grams were used. The assumption is to split the terms into word decompose into magnitude n, strategic equivalent algorithms are used to checks the text will match or not? This method also used for detecting and checking of spelling errors, text compression.

4.4.1 D'Amore and Mah:

With this method, only a defined amount of N-grams can occur for value n, therefore, a mathematical model was developed to evaluate the noise and demonstrate similarity measures.

The weight for N-grams is calculated as specified below (i in document j):

$$w_{ij} = (f_{ij}\text{-}p_{ij})/\sigma_{ij}$$

f_{ij}-> frequency of N-grams in document j
p_{ij}->predicted no. of happening of N-grams
σ_{ij} ->standard deviation

4.4.2 Damashek:

This algorithm is depend on vector space model. Instead it uses a centroid(C) vector to estimate noise. The similarity between a query(X) and a document(Y) is computed as

$$SC(X,Y)=\sum_{j=1}^{t}(wxj-\mu X)\,(wyj-\mu Y)/\sqrt{\sum_{j=1}^{t}(wxj-\mu x)^2\ \sum_{j=1}^{t}(wyj-\mu Y)^2} \quad (14)$$

Where X and Y signifies centroid vectors defined to categorize the input query and document language. The weight, wx_j and wy_j indicate the term weight for each component in the query and the document vectors. The centroid value is measured as the fraction of a over-all of happening of n-grams to whole.

4.4.4 Pearce and Nicholas:

The hypertext links were generated by the expansion of Damashek's by Pearce and Nicholas. The relations are acquired by calculating similarity among a chosen form of text and remaining from the document.

4.4.5 Teufel:

To reduce the noise of N-grams, stop list algorithms and stemming algorithms (removing suffixes/prefixes) are used. The assumption is that, if the document P is similar to Q, and Q is similar to R . Then P would be approximately similar to R. It uses new coefficient H,

H=M+N-(MN) and M is a direct similarity coefficient and N is indirect measure.

4.5 Regression Analysis:

Regression analysis is used to recognize the precise constraints that tie up with the data. It is mainly used for prediction purpose based upon the parameters we can predict, it may meet to the exact result or may not. In these we get only two results that is it can be yes or no. The following example determines the one of the equation that predicts increment in employee pay based on their experience.

Experience	Incrementing Salary
6	5000
2	1000

A simple polynomial regression could be implementing by using the below equation

$$IS= \alpha E + \beta$$

α, β to predict the salary i.e., incrementing salary(IS) based on Experience(E).

The first usage of logistic regression is given by cooper. Therefore the logistic regressions are given in two stages:

Stage1:

Log O(R/C1) =c0+c1x1+c2x2+c3x3
Log O(R/C2) =d0+d1x4+d2x5+d3x6

Stage2:

Log O(R/c1, c2...cN) =e0+e1z+e2N
z is used to compute the sum of composite result
$z=\sum_{i=1}^{N} log\ O(R/Ci)$ (15)

The result of first stage is applied to the second stage.

4.6 Thesauri:

One of the most important retrieval utility is thesauri. These are used to improve the effectiveness. It includes all the synonyms of the particular term. For example, if we want to know about party it must also contain the word gathering and another word. Therefore the synonyms words are there in the thesauri. If the user enters a query Q the system goes to thesauri and checks for synonym word and enhances the query and gives the relevant document.

4.7 Semantic Network:

A semantic network, is a network that consists of collection of nodes and arcs. Arcs are represented with the type of correlation they represents. The data structure represented with the arcs and nodes is called as a frame. The distinct items in the frame are termed as slots. A semantic network is used when there is a relation between concepts that is nothing but the nodes and arcs.

(
parrot
(is-a bird)
(has-color green)
(size small)
)

In the above example, parrot is node and is-a , has-color are the link to node parrot. Semantic network endeavor to determine the dissimilarity difficulty in which the items in a query do not peer those acquired in the document. The semantic network uses a module WordNet to search for a item from the database. It specifically contains frames sketched for words .It includes synonym, antonyms.

4.8 Distance Measure:

To measure the distance between single node and another node in semantic network the distance measure is used. let us consider an example for computing the distance between node u and v then the below formula is used

DIST (u, v) =Min Number of edges separating u and v.

4.9 R-Distance:

In this method, all the distinct approach in each set is averaged to the distance between all the available aggregation of two sets. For example, a query X for items ((p AND q AND r) OR (s AND t)) and document Y with items (i1 AND i2) then the similarity is calculated as given below

C1=d(p,i1)+d(p,i2)+d(q,i1)+d(q,i2)+d(r,i1)+d(r,i2)/6
C2=(d(s,i1)+d(s,i2)+d(t,i1)+d(t,i2))/4

To find the R-distance of partitioned normal form query X, and a document Y with items (i1, i2, i3 ...in) and C_{ij}, indicates j^{th} term in concept i

SC(X,Y)=min(SC1(C1,Y),SC1(C2,Y)....SC1(Cm,Y))

$$SC1\ (C_i, Y) =1/\min \sum_{i=1}^{n}\sum_{j=1}^{m} y\ (t_i, C_{ij})$$

SC (X, Y) =0, if X=Y

4.10 K-Distance:

The distance among two nodes is attained by acquiring the smallest route amongst the two nodes and then linking the edges along the path. The space between items i_i and i_j is acquired by:

$$d_{ij}=wii,x1+wx1,x2+.....+wxn,ij$$

c1=min(d(p,i1),d(p,i2))+min(d(q,i1),d(q,i2))+min(d(r,i1),d(r,i2))/3
c2=min(d(s,i1),d(s,i2))+min(d(t,i1),d(t,i2))/2

The k-distance of a division normal form query X and document Y with terms (i1, i2...in) is derived as

$$SC(X, Y) = SC1(X, Y) + SC1(Y, X)/2$$

$$SC1(X,Y) = \min(SC2(c1,Y), SC2(c2,Y), SC2(cm,Y))$$

$$SC2(ci,Y) = 1/n \left(\sum_{j=1}^{n} min(y(c_{ij}, i_j))\right)$$

$$SC(X,Y) = 0 \text{ , if } X = Y$$

4.11 Parsing:

To determine a set of tokens which appears to be as the body of text is a main aspect of every information retrieval system.

4.11.1 Single terms:

Stemming and stop words are used. Stemming is used to standardize the terms by removing suffixes and prefixes. The main intension of the stemming is that due to the mismatch in the suffixes and prefixes no term in the query and the relevant document should be missed. For example, the user who includes the term compressed also match on "compress" and "compressing". Therefore porter stemming algorithms are most commonly used. These algorithms remove common suffixes and prefixes. Stop words are the words like the connecting words i.e. a, and, these, that.

4.11.2 Simple phrases:

The terms that are not detached by a stop term, punctuation mark (Ex: Comma, Semicolon etc.), or special character (Ex: +, “”, - etc) are found first.

4.11.3 Complex phrases:

The Natural Language Processing (NLP) is used to recognize a document by shaping a canonical structure that signifies the document and it consist of the part of speech, named entity tagging and syntactic parser.

V. Conclusion

Describes about the functionalities of the Information Retrieval and varies retrieval strategies, retrieval utilities to improve the performance of IR. Retrieval Strategies used to find the similarity measure to retrieve the relevant results. Various strategies described the parameters to improve the performance of IR and explained with the example.

VI. Acknowledgement

I would like to take this opportunity to thank Sri. D.Manohar Reddy, MLA,Founder of Trinity Educational Group, Sri.DasariPrashanth Reddy, Chairman, TrinityCollege of Engineering and Technology, Karimnagar, Telangana, India, Dr.SG Sangashetty, Principal, Sri. N.Radha Krishna, AO, Trinity College of Engineering and Technology, Karimnagar, Telangana, India.

VII References

[1] A Hybrid Semantic Similarity Measure for Sptial Information Retrieval By Angela Schwering
[2] The SIGSPATIAL, ACM Volume 3 Number 2, 2011.
[3] Normalizing Spatial Information to improve Geographical Information Indexing Andretrieval In Digital Libraries Damien Palacio and Christian Sallaberry and Mauro Gaio.
[4] Damien Palacio and Christian Sallaberry and Mauro Gaio by Damien Palacio and Christian Sallaberry and Mauro Gaio
[5] Spatial Information Retrieval by Wenwen L, Phil Yang Bin Zhou

VIII. About the Authors

1. **Mr.P.RameshBabu**, is a research scholar in Rayalaseema University, Kurnool, Pursuing Ph.D in Computer Science & Engineering and Working as a Assoc.Professor, HOD CSE Department in Trinity College of Engineering and Technology, Karimnagar Telangana

2. **Dr.M.Nagabhushana Rao** Completed Ph.D in Computer Science and working as a Professor, CSE Department,KL University, Vijayawada, Andhra Pradesh, India