Indian Journal of Science and Technology, *Vol 9(11), DOI: 10.17485/ijst/2016/v9i11/81658, March 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Comparative Analysis of Data Mining Tools and Techniques for Information Retrieval

Amit Verma[1]*, Iqbaldeep Kaur[2] and Inderjeet Singh[1]

[1]Department of Computer Science and Engineering, Chandigarh University, Mohali - 140413, Punjab, India; cu.inderjeet@gmail.com
[2]Department of Computer Science and Engineering, Chandigarh Engineering College, Mohali - 140307, Punjab, India; iqbaldeepkaur.cu@gmail.com

Abstract

Background/Objectives: There are a lot of information retrieval techniques available for getting information from different kinds of sources. Main aim of this paper is to improve information retrieval activities to a higher level. Different methods for information retrieval have been studied and discussed. It involves use of Fuzzy Ontology Generation framework (FOGA) framework along with Formal Concept Analysis (FCA) based clustering and keyword matching approach. Hidden Markov Model has been used for retrieval of data from search engines in an intelligent and efficient way for correct identification and retrieval from the database. Classification algorithm has been used for detection of community and conversion of large community graph to sub community graph for its better study and usage. **Findings:** It has been found that fuzzy ontology generation framework can automatically generate the fuzzy ontology which is very hard and time consuming task otherwise. Clustering of data can be done using the technique of formal concept analysis along with keyword matching method. There is a large amount of data available under similar words but with different meanings. So there has been a lot of problem in retrieval of exact data as required in a very short amount of time. In this case Hidden Markov Model can be used which can find the non-observable or hidden stochastic process from the observable stochastic process. **Application/ Improvements:** Generalized Expectation-Maximization algorithm used with Hidden Markov Model can find unknown parameters. By adding frequency tracking algorithm along with Hidden Markov Model, we can also track audible data from a large database. Community detection algorithm along with Informap and Bigclam algorithms used with Hidden Markov Model will increase the modularity of data. Its applications include use of information retrieval of different types of data in an extremely faster and efficient way.

Keywords: Classification, Clustering, Community Detection Algorithm, Fuzzy Ontology Generation Framework, Formal Concept Analysis, Hidden Markov Model

1. Introduction

Data denotes to the fact that there is some existing information which is represented or coded in some particular form which is suitable for its better usage or processing. This definition defines that data can be shown in written form, in the form of graphs, in the form of tables or in pictorial form. Data is mainly made or collected so that the information which is collected can be used or processed for its better usage, later. In our daily life, we store data in our mobile phones, in notepad, it can be the list of phone numbers, the list of items to buy; in case of organizations they store data of the daily work done by their organization. All this stored data is used so that it can be processed later for personal or organizational benefits. Benefits can be in the form of convenience or to make things better in the future, etc. Digital data refers to any of the quantities, characters or symbols on which particular operations are performed with the help of a computer and it can be stored, recorded or transmitted in the form of electrical signals. This definition defines that in computer data is of the digital form. Digital data

*Author for correspondence

can contain symbols, character or numeric values. Data is different pieces of information. In prehistoric times, data was stored only in human's brain, as given in the Table 1.

This definition also defines that data is some pieces of information. Data can exist in numerical form, in textual form on papers, in the form of bits in electronic memory, or it can be simple facts in the mind. Data is plural term of datum. Datum is a single piece of information. But in daily life, data can use as singular or plural form.

Digital data is usually stored on computer storage devices such as magnetic disks, optical disks or some other mechanical recording media. Digital data can be transmitted in the form of electrical signals. In today's world all the organizations are connecting to computers. They want to store their data for keeping records, for doing manipulations on stored data to make their services better. Shops store data of buying and selling to keep record of the money input, and output and the amount of benefit they are getting. Banks store data to keep track of money transfers which can be accessed by only particular persons. Keeping data in hard disks or on other means of computer storage makes it faster to upload, download. It is more secure and trustworthy. It can be the form of electrical signals. In today's world all the organizations are connecting to computers. They want to store their data for keeping records, for doing manipulations on stored data to make their services better. Shops store data of buying and selling to keep record of the money input, and output and the amount of benefit they are getting. Banks store data to keep track of money transfers which can be accessed by only particular persons. Keeping data in hard disks or on other means of computer storage makes it faster to upload, download. It is more secure and trustworthy. It can be transmitted to longer distances securely and only it can be more convenient.

Table 1. History of data storage before 1900 A.D

S.No.	Y_{ea}	I_{nv}	I_{vb}
1.	P_{hi} B.C.	C_{es}	N_{at}
2.	A_{hi} B.C.	S_{tp}	A_{eg}
3.	1725 A.D.	P_{uc}	B_{bj}, F_{ra}
4.	1890 A.D.	P_{cr}	H_{ho}, A_{me}
5.	1886 A.D.	F_{ic}	H_{eb}, A_{me}

B.C. = Before Christ, A.D. = Anno Domini, P_{hi} = Pre-historic, N_{at} = Nature

Data is mainly a collection of facts which are in the form of numbers, words, measurements, observations or descriptions of things. This definition defines that data is simply a collection of facts. It can be in numerical form, character form, measurements, observations or descriptions. It tells that data is mainly of two types, i.e., Quantitative form (it is numerical data) and Qualitative form (it is descriptive information). We can divide quantitative data into discrete and continuous form. Similarly Qualitative data is of 3 types, first is interviews, observation and written documents.

At earlier time there was no way to store the spoken data in tangible form. So data was tried to be stored by making drawings on rocks, later leaves were used to store data in textual form. But still it was not very safe for a longer time. Slowly it was felt importance to store data so that it can be used for record keeping. So that future generations can also use it. Then data was stored in written form on animal skin, because it can last longer. But it was not a permanent solution. So there was discovery of paper. In earlier times, to write there was use of flower colours or other colours from nature. Slowly pencil and pen were discovered. Data was stored by scientists, organizations, etc., to keep records. But they were destructible and it was time consuming to manipulate them. So there was discovery of electronic devices to store and share data. Data storage methods changed a lot by the time. Some of the data storage inventions after 1900 A.D. are given in Table 2.

Now we use pen-drives, memory cards and cloud-computing techniques to store and retrieve data. If we talk about past present and future then it is our ability

Table 2. Data storage inventions between 1900 A.D. and 2000 A.D

Sl. No.	Y_{ea}	I_{nv}	I_{vb}
1.	1932 A.D.	M_{gd}	G_{ut}, A_{us}
2.	1951 A.D.	M_{at}	F_{pf}, G_{er}
3.	1956 A.D.	H_{ad}	R_{ej}
4.	1969 A.D.	F_{dk}	Y_{on}, J_{ap}
5.	1978 A.D.	S_{sd}	I_{nt}, C_{al}
6.	1980 A.D.	C_{od}	J_{ar}, A_{me}
7.	1990 A.D.	O_{nb}	T_{bl}, B_{ri}
8.	1995 A.D.	D_{vd}	S_{pp}
9.	1998 A.D.	P_{fs}	A_{do}, I_{sr}

M_{gd} = Magnetic drum, G_{ut} = Gustav Tauschek, A_{us} = Austria

to receive, store and recall data, which has changed tremendously in last one century. But this evolution goes back even further. In prehistoric times (B.C.), there was usage of the original and most powerful data storage device till today, i.e., cerebral storage.

This synaptic treasure trove of data provided the best form of storage for generations of oral traditions. When the caveman memories needed to be transferred amongst them, they used walls as their medium to record data. In ancient history (B.C.), there was usage of stone tablets and papyrus/paper. Cave paintings had a point of failure that they were immovable. So for Stone Age man, carved tablets were used for data storage. But it was a problem to carry a large amount of tablets and it was also dangerous. So later, papyrus/paper was used as a means of storage. Until 19th century, papyrus was the only acceptable means of storage. Many years later in around 18th – 19th century, punch cards were used. In 1725 A.D., 'punch cards' were invented and were used for information storage in 1832 A.D. In 1890 A.D., the scientist named Herman Hollerith was the first to invent a punch card that could be read by a machine. He merged many companies to form IBM, later. Punch card was a kind of improvement in tablets. In field of papyrus/paper, there was invention of 'filing cabinet'. In 1898 A.D. the first filing cabinet was used in insurance firm. It was an earlier form of multi-file compression. In 1932 A.D., there was invention of magnetic drum, in Austria.

It was an early form of computer memory. It used the technology that used electromagnetic pulse by changing the magnetic orientation of the ferromagnetic particles which are present on the drum. In 20th century, there was invention of magnetic tapes. They were first used in 1951 A.D. Magnetic tapes replaced punch cards. It had the capability to store more data of more than 10,000 punch cards. It became most popular till 1980 A.D. In 1956 A.D., IBM introduced 'hard disk'. But it was of very big size and large till 1970 A.D., so they were not of much use. In 1990 A.D., hard disks were used for tape backups. In 1969 A.D., floppy disk was read-only 8 in disk that stored 80 kb of data. It was considered as a revolution in data storage because they were portable from one computer to other. And gradually they became cheaper and more widely used. First CD was created in 1980 A.D. by Philips and Sony. CD was a replacement to aging floppy disk. The next generation of CD was DVD; it came in 1995 A.D. in the market. It had the storage capacity of 4GB, which was of same size of CD but was around 7 times capable to store data. In 1960s A.D. there was invention of Internet. But the online backup and storage was available only after 1990 A.D. Now there was no need of device anymore, and data could be backed up from a remote location. Hard drives continued to reduce in size and they were evolved in the form of portable flash storage device in 1998 A.D. It had the capacity to store a very large amount of data. Major inventions in 'data storage' made after 2000 A.D. are given in Table 3.

Blue ray disks were used in 2000 A.D.; it had promising storage space of 400 nanometres. In 21st century, in advancement to internet technology, there was usage of 'the cloud'. It allowed data to be stored on multiple servers, hosted by third party. First successful and prolifically used SSDs were in 1978 A.D. It is a solid state hard drive. It became so popular in past few years. It used to store data with electric charge rather than magnetism. In future, we can expect 'holographic layers'. It would allow data to be encoded on tiny holograms' layers. And it would have capacity to store data for 30+ years. Another storage technique could be 'quantum storage', it will be extremely small in size, and it couldn't be read by even the smallest microscope. It would use single electron to store single bit of information and it will only be decipherable by quantum computers. There are also other future scopes in which the data storage is possible. As the newer inventions of data storage are coming, the size of hardware is decreasing and inversely proportional to it, the storage capacity is increasing.

There is a vase amount of data available in various forms; it includes paper form and electronic form. Paper data is present in the form of books, magazines, newspaper, pamphlets, etc. Electronic data has been stored in various electronic devices which are discussed above. Data available in the paper form can be copied to electronic sources for better maintenance of data. Electronic data is mainly present in the form of images, audio, video, animations in various formats or data can be present as a mixture of above forms. So as to utilize that data in a better way

Table 3. Data storage inventions after 2000 A.D

S.No.	Y_{ea}	I_{nv}	I_{vb}
1.	2000A.D.	B_{rd}	B_{da}
2.	$21^{st}C_{en}$	T_{cl}	J_{rl}, A_{me}
3.	F_{tr}	H_{qs}	U_{kn}

B_{rd} = Blue ray disk (400nm), B_{da} = Blu-ray Disc Association, Y_{ea} = Year, I_{nv} = Invention, I_{vb} = Invented by, C_{en} = Century

we have different techniques for the electronically stored form of data, it includes data mining and information retrieval.

Information retrieval refers to getting of data from a repository or database as it is. There are different types of repositories available which includes standalone repositories, it is a kind of repository which is complete in it and is not dependent on any other repository; global repository, it refers to a repository which has been shared to many other platforms and repositories for handling large amount of data to a larger number of users; local repository, it is used under a certain domain, it uses global repository; versioned repository, this kind of repository has been used for local and global objects and there is version control of it. In case of information retrieval there is no change in data. We just copy the required data from the available database.

Data mining refers to taking of required data from the database or repository in a different form than the data available in it. Change of form takes place after using certain techniques such as classification, clustering, prediction, etc. The output in data mining can be taken in the form of graphs, tables, cluster formation, etc. of the mined data. In case of data mining, information is mainly collected to do comparison on available data to get more meaningful information out of it. Different methods can be applied to perform different activities on data such as using genetic algorithm, neural network to form some other data, based on certain condition, from the available data.

Literature survey covers the various papers research related to information retrieval and data mining.

In[1] proposed about information retrieval machines. Operations of these machines are based upon code system. Retrieval theory is dependent on communication theory. Development done in practical and theoretical systems helps in the development of signalling systems. Future work would include handling and retrieval of semantic information. Retrieval system has been discussed in detail including its model, measure, coding systems and selection errors. Architecture to perform this is given.

In[2] proposed the need to have information retrieval system in organizations for handing their own publications. Various steps shown for doing it, these are observations, describing information, coordinate indexing, storing information and practical aspects. All these steps are explained in detail in this paper. It is found that using this method through its application properly will help the organizations through economically as well technologically. Comparison of various implementation methods by different means and with the help of different parameters are shown in Table 4.

In[3] proposed that decision making task for evaluating purpose. It is important to carefully evaluate the proposal. A lot of issues can be present during evaluation. Process to remove those issues is told. Actual process of evaluation is also told in step by step manner. 17 attributes and their measurement units that were used in comparing and evaluating proposals, their scaling constants are listed in the Table 5.

In[4] proposed that neural network can be used for information retrieval from a pictorial system. For auto-associative memory operations neural networks are used. Auto-associative memory operations are not much capable but at the same time the performance of auto-associative memory is very sensitive to the algebraic properties. Hierarchical approach proposed in this paper is efficient but it has some drawbacks too. Hierarchical system does

Table 4. Comparison of implementation methods

S_{ys}	P_{ar}				
	T_{im}	D_{if}	E_{xp}	C_{on}	E_{co}
E_{tn}	0	0	0	0	1
U_{tc}	0	1	0	1	1
O_{ci}	1	1	1	1	1
P_{uc}	0	1	1	1	0
C_{om}	0	1	1	1	0

S_{ys} = System, P_{ar} = Parameter, T_{im} = Time, D_{if} = Differentiation, 0 = No, 1 = Yes

Table 5. Attribute details table

A_{tt}	S_{lc}	P_{ro} 1	P_{ro} 2	P_{ro} 3	P_{ro} 4
X1	.1125	30	34	29	54
X2	.0450	2.0	3.0	2.0	2.0
X3	.0675	64	60	50	45
X4	.0675	2.8	4.0	3.0	3.0
X5	.0450	5.1	2.0	3.0	3.0
X6	.1800	91	90	80	69
	K= –.138				
Overall score	$17\pi[(1+k.ki.ui.)-1]/k$ $_{1=1}$	.805	.761	.589	.541

A_{tt} = Attribute, S_{lc} = Scaling constants, P_{ro} = Proposal

error correction at global and local levels. Performance of system investigated through storage and reconstruction of patterns. Abstraction methodology used to represent complex spatial relationships. Image processing technique is used to identify contained objects in picture. Hopfield model uses asynchronous update algorithm. Learning algorithms such as Hebbian and spectral can be used by Hopfield neural network model for the establishing stable patterns.

In[5] proposed a tool for knowledge discovery. It helps in process of information retrieval. It can acquire and hold relations in-between the words so as to describe the context of information knowledge for a subject. First all the words of context are indexed. But these can by synonym words with different meanings. So there is needed to identify context descriptors. Context can be constructed automatically (through clustering technique) or manually. We can use automatic algorithm to find relations amongst words but at the same time this process is slow. It is sometimes hard to understand the wants of a user. Context analysis used to solve information retrieval problems. With hyper dictionary, queries can be made and but in it words are consistent than other means. Probability of finding relevant information is increased. This paper shows various problems faced in information retrieval; traditional techniques used; architecture and working of 'Hyper dictionary' tool to solve information retrieval problem.

In[6] proposed various application and the related issues in case of machine learning, data mining, knowledge discovery, knowledge acquisition, information retrieval, inductive decision-making and database. There is need of a human expert's decisions; web databases, document collections and environmental datasets in case of both types of repositories i.e., unstructured. In case of collection of large documents a computerized search can be performed through information retrieval algorithms so as to retrieve multiple documents which are according to user's need. Information retrieval algorithms are used in digital libraries and online searches.

In[7] proposed the 'Web based Information Retrieval Support Systems' (WIRSS). WIRSS used to build research tools. It can access information online. It can explore the information. It can also use the information from the Web. Evaluation of this tool is shown. Work done in this field and published articles related to it are explained. Types of WIRSS are also explained.

In[8] proposed a developed model so as to mine information through graph analysis. In case of data mining, views of clustering web nodes and find item sets for association rule are presented. And in case of information retrieval, different views are there for query making, clustering of the results of query and to improve the ranking in quality is presented. A model having these attributes demonstrates flexibility, modularity, broadness in the graphical problems and applicability and applicability. In this paper the model, its applications in information retrieval and data mining and ranking information retrieval and data mining and ranking improvement is discussed.

In[9] proposed about the best practice to do effort estimation using liner regression which includes stratification and local calibration. Different experiments were done on COCOMO81, COCOMO11 and NASA93 to find the best method for cost estimation. Their results are shown diagrammatically. And the reasons for large deviations in the results are also explained. Major reasons for deviations in result were because of less training, testing, making self-made assumptions and bigger size of tools were the reason. Best regression or parameter based techniques have been discussed, which includes various types of COCOMO and other techniques. Coseekmo tool is discussed, it use rejection rules to make estimation models. Factors causing deviations in COCOMO results, using Coseekmo have been explained. Applications of Coseekmo are explained in building effort models, data sources and for validating stratification. Rejection rule applied on COCOMO 81 is shown in Table 6.

In[10] proposed that knowledge retrieval model for searching, optimizing of query, resources constructing and result analysis. Use of data mining into knowledge retrieval provides different methods and retrieval strategies

Table 6. Survivors of the rejection rule from the coc81's embedded systems are shown in the table below

T_{re}			R_{es}				
N_{um}	S_{ub}	L_{er}	P_{re} (30)	M_{ea}	Sd	(sd/M_{ea} %)	C_{or}
P_{re}	17	E	46	40	34	85	.93
P_{re}	17	LC	48	38	34	88	.86
R_{ou}	17	LC	50	39	34	87	.86
P_{re}	16	LC	50	39	34	87	.85
R_{ou}	16	LC	47	43	38	89	.81
P_{re}	15	LC	47	43	38	88	.85
R_{ou}	15	LC	41	45	42	93	.88

T_{re} = Treatment, R_{es} = Results, $_{um}$ = Numbers, P_{re} = Precise, R_{ou} = Rounded, S_{ub} = Subsets, L_{er} = Learn, P_{re} = PRED

for mining, dynamic learning and adaptability. Various knowledge retrieval algorithms are explained. A training instance set for the classification algorithm has been shown in Table 7.

In[11] proposed the way to find experts for right job. It can be done in the best way if there is some proper way to do it. Most of the experts are available on famous social networks, such as Facebook, MySpace, Twitter, LinkedIn, Xing. These sources can be used to connect individuals to make a project team. The graph can be formed to see how closely one individual is related to the other and how closely it can communicate and collaborate. Lappas et al., proved team problem is NP-Hard, so they propose two approximation methods i.e., Rarest first and Enhanced Steiner algorithm by defining the communication cost, without considering skill grading. Lappas et al. gave 'Enhanced Steiner algorithm'. It can associate a specific no. of experts with the specific skill. This paper adds to Rarest First algorithm with new definition of communication cost to both distance and skill level of individuals. Each individual has a set of skills. There must be a minimum skill level of expert. Rarest First algorithm was proposed by Lappas et al. for the tasks which are basic so as to resolve the problem of forming a team. It's given that first of all those individuals must be found who have rarest skill. After that the individuals with other skills are explored. In[11] made generalized diameter algorithm. It uses the method that find the skill of the individuals and the papers published by individual related to skill are counted to measure size of skill set of individual. Generalized Diameter algorithm decreases the cost of communication and cardinality, as related to Generalized Steiner algorithm proposed by C. Li et al. Generalized diameter algorithm has the capacity to form expert team in proper manner, from a social network with diverse attributes.

In[12] shown a survey on challenges faced in design of intelligent information retrieval systems. Web based information retrieval system is discussed. Various problems in it and how to deal with it are explained. Major research challenges in information retrieval and how to solve them effectively is explained. Search quality and soft computing in information retrieval system is explained. Soft computing which includes web mining, soft web mining and in it fuzzy sets, Probabilistic information retrieval, artificial neural network and genetic algorithms are discussed.

In[13] proposed that there is importance to get results according to user needs and not only related to query words. It introduces a new method of integration the sensitivity of data in the search engines naturally, which can be done using sensecalc and sensiaugment approaches. Conceptual framework describes different components of sensitivity model. Overview of architecture of sensitivity-based search engine is given. First of conceptual framework shows various notations and various definitions have been given. Later the related work is given.

In[14] proposed a protocol to privately query about the patients' data which is stored on cloud from different geographic locations. Cloud storage is not a safe method to store data without proper protocols which can efficiently retrieve the private data. An older protocol PIR which is weak is discussed. Solution to this problem is discussed, which includes, one query solution, anonymizer and

Table 7. Training instance set

D_{cn}	S_{er}	A_{gt}	W_{eb}	I_{mr}	C_{on}	I_{nt}	I_{me}	S_{em}	M_{ed}	G_{ri}	S_{em}	I_{nt}	T_{ef}	O_{nc}	D_{at}	O_{bf}	A_{sr}	D_{ot}
D01	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	P_{os}
D02	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	P_{os}
D03	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	P_{os}
D04	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	P_{os}
D05	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	P_{os}
D06	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	N_{eg}
D07	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	N_{eg}
D08	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	N_{eg}
D09	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	N_{eg}
D10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	N_{eg}

D_{cn} = Document number, S_{er} = Semantic retrieval, A_{gt} = agent, W_{eb} = web, I_{mr} = Image retrieval, C_{on} = Content, I_{nt} = Intelligent, I_{me} = Image emotion, S_{em} = Semantic identification, P = P_{os} instance, N_{eg} = Negative instance

then overall solution is explained. The whole background process for getting data in this case is explained. Limitations of method are also told in the end.

In[15] proposed that there are faster hardware/software solutions for performing clustering. Hierarchical merging is proposed as the best technique for it. The method to perform hierarchical merging is explained in a step by step manner. Two major activities of sorting and merging used in it are explained in detail. Architecture for sorting and merging for hardware and software has been shown diagrammatically. The difference in acceleration achieved by software only and hardware/software sorters is by a factor ranging between 1.13 and 4.08 only. Acceleration in different types of sorters is shown in the Table 9.

In[16] proposed a hybrid approach to image retrieval. They also proposed rough set based approach to audio/video retrieval. Both these are explained in detail. Features of images are told. Summary correspondence between attributes and objects like audio files, video files, text etc. has been shown in Table 9.

In[17] proposed the new way to use fuzzy rule-based language and clustering to retrieve ontology based information. FOGA is used in combination with FCA and with keyword matching approach. It includes various properties of FCA based clustering. It also includes the algorithm for mapping of hybrid ontology with the keywords. Experiments on text files are done. And clustering results are evaluated by Precision, recall and F-measures. Where precision is the correctly retrieved document from all the retrieved documents. Recall refers to those documents which are correctly retrieved with respect to the total number of retrieved documents. Calculation of precision and recall for FOGA and keyword matching has been given in Table 10.

Calculation of F – measures for FOGA and keyword matching has been given in Table 11.

Table 8. Acceleration achieved in sorting by software only and hardware/software data sorters has been shown

S_{or}	N_{lt}	S_{pu}
1	64	4.08
100	6400	1.56
1000	64000	1.34
5000	320000	1.24
10000	640000	1.22
20000	1280000	1.20
50000	3200000	1.18
100000	6400000	1.16
300000	19200000	1.14
600000	38400000	1.13

S_{or} = Sorting blocks (L/N), N_{lt} = Number of L items, S_{pu} = Speed up

Table 9. Correspondence between features and objects

O_{bj}	F-1	F-2	F-3	F-4	F-5
O-1	3	4	5	2	2
O-2	3	4	5	2	2
O-3	4	5	5	2	5
O-4	5	5	3	4	3
O-5	4	3	5	4	3
O-6	5	5	3	4	4
O-7	4	5	5	3	5
O-8	5	3	4	4	3
O-9	4	3	5	4	4

O_{bj} = Object

Algorithm for Clustering Fuzzy Ontology has been given below.

```
Algorithm Clustering_Fuzzy_Ontology
    Begin:
    Involve C1, C2, C3;
    Where C1∈ P1, C2∈ P2, C3∈ P3;
    //P1 and P2 are properties; C1 and C2 are the clusters
    for (Cn=1; Cn<Cl; Cn++)
    //where n=1 to 3
    {
    Apply mapping through fuzzy ontology;
    }
    for (Kn=1; Kn<Knl; Kn++)
    {
          for (Dn=1; Dn<Dnl; Dn++)
    {
                    Generate database;
    {
          for (Dn=1; Dn<Dnl; Dn++)
    {
    generate database;
    }
    }
    if Dn = Dnl;//where Dn∈ Kn
    {
    (Exit)
```

Table 10. Precision and recall

A_{pp}	P_{re}	R_{ec}
S_{tf}	0.79	0.87
K_{mg}	0.85	0.72
H_{fk}	0.89	0.65

A_{pp} = Approaches, P_{re} = Precision, R_{ec} = Recall, S_{tf} = Standard FOGA, K_{mg} = Keyword matching, H_{fk} = Hybrid FOGA with keyword

Table 11. Show F - measures

A_{pp}	F_{ms}
S_{tf}	0.52
K_{mg}	0.76
H_{fk}	0.85

F_{ms} = F-measures

```
Else
perform clustering;
}
    }
    }
End.
```

In[18] proposed the method for supporting fuzzy semantic information retrieval in the book domain. This new method helps in high recall and high precision of search results. First of all, all the details of a particular book are collected. Then classes and sub-classes are found. After it the properties and rules and found and saved in an OWL or RDF data format.

In[19] proposed a method to make intelligent music retrieval system. All the music is present online, but polyphonic music can be hard to find. Hidden Markov Models can be used in music search engine. Whole of the process in which HMM can be used in music retrieval is explained and results are shown in tabular form. Fundamental algorithm for tracking the frequency has also been used. Frequencies are quantized and mapped. Other algorithms are also used along with HMM such as Generalized Expectation-Maximization algorithm and frequency tracking algorithm. And there is 90% accuracy in this process. Algorithm for methodology used is given below.

```
AlgorithmHidden_Markov_Model
Begin:
    Initialize S_d, Y_t(i), E_t(i,j), t
    //S_d= Songs database
    for(S_dn=1; S_dn<S_dnl; S_dn++) //n= 1 to 20
    {
    generate Hmmn
    if (Hmmn ≤ 12) //Hmm = Hidden Markov Models
    {
    create database;
    }
    else
    exit (1)
    }
    for (Dhmm = 1; Dhmm<Dhmm_l; Dhmm++)
    //Database hidden markov models
    {
    trace through Sh;
    //Sh = Schroeders histogram
    }
    for (Udhmm=1; Udhmm<1;Udhmm++)
    {
    apply Baum-Welch algorithm;
    }
    If (Udhmm ≤ T)
    {
    Database added (Dhmm);
    Else
    Exit
    }
    End.
```

In[20] proposed a technique and algorithm on for using large community graph to retrieve sub-community graph. First of all a large community graph is converted to compressed community graph and then to sub-community graph. And it is observed that retrieved sub-community graph is simpler, efficient and easier in terms of complexities. Earlier algorithms for solving this problem are also discussed. It included community detection algorithm which increased the modularity. Modularity had the problem that it couldn't detect the small sized communities from a big sized community. So there was a multilevel approach but it had quality issues. Infomap and Bigclam were other algorithms proposed. First of all members of community are calculated and shown in the matrix form. Total edges of matrix are calculated numbers are assigned for first row and first column in it. Then edges have been created for similar and dissimilar members of the community and the

expanded matrix is displayed. Procedure for edge creation and for extraction of sub community graph is given. As an example a compressed graph of community has been taken and analyzed. All the steps are shown diagrammatically for this example. Classification algorithm for solving this problem is given below.

Algorithm Classification
Begin:
Int C_{no}, T_{cm}=0, C_n, C_{cam}, E_{am}
for (i=1; i<n; i++);
{
print rows;
}
for (j=1; j<n; j++);
{
print columns;
}
create C_n[i] [j];
//C_n = Matrix to hold community numbers
if |C_n|≠0;
{
T_{cm} = T_{cm}+C_n [i] [j]
//T_{cm} = To store total number of community members
}
else
exit (0);
for(i=1; i≤2; i++)
{
for (j=1; j≤2; j++)
{
E_{am}[i];
//E_{am} = Expanded adjacency matrix of order (TCM+1) X(TCM+1)
}
E_{am}[j];
}
exit (0)
for (i=1; j=index+1; i<n; i++; j++)
{
if E_{am} [i] [j] = 0; Calculate C_{cam} [] [];
else
//C_{cam} = Compressed Community Adjacency Matrix of order nXn.
calculate C_m;
//C_m = Community menu
}
if E_{am} [i] [j] = 0;
calculate C_{cam} [] [];
else
calculate C_m;
Calculate Ecreation;
if Ecreation = 1
similar item
else
dissimilar items
End
Function C_{cam}
C_{cam} [i] [j];
where i = 1 to n;
for (i=j; i<n; i++)
{
C_n [i] [j] = C_{cam} [i] [j];
}
Function Ecreation
Ecreation [i] [j]
for(i=2; j=2)
if (E_{am}[i-1][1] = E_{am}[1][j-1]
and E_{am} [i][j]=0),
then
any Edge from E_{am} [i–1]
if Yes, then
Row: = Row+1.
End

In the table below major details of all the research papers is given. First column includes name of the research paper. Second column has author names. Third column includes the year of publication. All the papers above and in the table below are arranged in increasing year wise manner. It will help to know that various inventions done in the field of information retrieval with the passage of time can be understood in a better way. Next column includes the tools used in each paper. Column after that enlists all the techniques that have been used in each paper. There are a lot of data mining techniques also which could be used in information retrieval process. Next column includes the names of the algorithms used. Next column includes important works done in the paper. Last column includes the list of work which has not been done or the future work of the paper which is possible. There are all details of literature survey in Table 12.

2. Data Flow Diagrams

Data flow diagrams related to information retrieval system, up to level 2, are give below: -

Table 12. Details of literature survey

T_{ie}	A_{us}	Y_{ea}	T_{oo}	T_{ec}	A_{lo}	F_{ee}	F_{ws}
C_{ci}	C_{nm}	1954	N_{av}	N_{av}	N_{av}	R_{sc}, R_{ds}	T_{si}
D_{ir}	M_{dr}, W_{fm}	1963	C_{ig}	I_{ts}, T_{fr}, S_{ct}	N_{av}	O_{rp}	N_{av}
E_{pd}	R_{ks}, A_{si}, K_{na}	1978	N_{av}	D_{at}, L_{it}, T_{if}, W_{rt}	N_{av}	E_{re}, E_{dh}, M_{et}	A_{dm}, N_{po}
P_{ir}	A_{sp}, A_{rl}	1992	N_{av}	N_{av}	N_{nt}, I_{pt}, P_{ht}	A_{mo}, l_{rp}	H_{ap}, S_{oa}
H_{ti}	L_{kw} , S_{lh}	1998	A_{va}	T_{ir}	F_{ba}, A_{ma}	A_{fw}	A_{ic}
D_{mi}	Spu, H_{mc}	2002	D_{el}	D_{mt}	I_{ra}	D_{ir}	N_{av}
W_{in}	J_{ty}, Y_{yy}	2003	N_{av}	N_{av}	N_{av}	S_{fr}	C_{ef}
A_{wd}	A_{rp}, R_{by}	2005	N_{av}	N_{av}	A_{fc}, A_{pa}	N_{mg}	N_{av}
B_{ps}	T_{mn}, Z_{ch}, J_{hn}, K_{lm}	2006	C_{ot}, L_{rt}	T_{bt}, M_{gt}	R_{ba}, W_{sa}, Q_{ma}	B_{ee}	N_{id}, N_{bm}
R_{kd}	Y_{ha} , Y_{fh}	2010	N_{av}	N_{av}	K_{ra}, A_{rn}, C_{ra}, C_{rl}, I_{la}, R_{el}, A_{re}, C_{ls}	R_{dr}	N_{av}
E_{es}	F_{ar}, M_{ys}, S_{th}, A_{lh}	2011	B_{tt}	D_{sp}, F_{mp}	R_{fa}, E_{sa}, G_{da}, S_{ga}	U_{pe}, R_{fa}, M_{ce}	H_{se}
S_{ii}	M_{wa}, D_{ma}	2012	P_{mt}, M_{td}	S_{ct}, R_{ag}, S_{qm}, T_{if}, P_{et}, W_{st}, A_{it}, I_{da}, D_{mt}, C_{mt},W_{um}, I_{rt}, P_{bt}	G_{em}, F_{ma}, P_{ai}, I_{cn}, M_{sa}, I_{ra}, C_{ma}, F_{sa}, B_{sa}, B_{qp}	D_{si}	M_{pd}, N_{sh}
A_{tp}	K_{aa}, J_{ay}	2013	L_{in}, F_{bs}, G_{oa}, P_{md}	S_{at}, P_{ft}, P_{at}, C_{at}	W_{as}, S_{am}, G_{ra}, G_{ka}	P_{mr}	N_{av}
F_{iu}	M_{ia}	2013	N_{av}	N_{av}	N_{av}	I_{bu}	S_{br}, C_{sd}
P_{ia}	F_{kd}, K_{ee}, S_{mn}	2014	N_{av}	N_{av}	C_{la}, G_{aa}	E_{ps}	N_{av}
H_{ai}	V_{sy}, I_{os}, J_{sa}, A_{sn}, A_{rr}	2015	S_{dk}, V_{ds}	H_{lm}, H_{lc}, D_{ts}	D_{ma}	M_{cp}, V_{ld}	N_{av}
H_{fi}	B_{al}	2015	T_{mt}, P_{et} , F_{nt}	C_{lt}, F_{og}, Irt, F_{bc}, K_{ot}	K_{pm}, C_{lu}, C_{la}, E_{xa}, P_{tk}, S_{aa}	G_{ii}	D_{aa}
D_{df}	R_{es}, V_{sc}	2015	S_{wt}	E_{st}, W_{st}, S_{st}, I_{rt}, T_{dr}	I_{ea}, S_{aa}, S_{er}	I_{rp}	N_{av}
M_{ir}	S_{ca}, M_{ss}, A_{gi}	2015	N_{av}	N_{av}	F_{ta}, , B_{wa}, G_{ma}	P_{pr}	N_{av}
A_{rc}	B_{ro}, A_{nm}, J_{am}	2015	N_{av}	R_{aw}, G_{mt}	C_{da}	S_{eg}	N_{av}

E_{es} = An Effective Expert team Formation in Social Networks Based on skill grading, F_{ar} = Farnoush Farhadi , M_{ys} = Maryam Sorkhi, S_{th} = Sattar Hashemi, A_{lh} = Ali Hamzeh, M_{wa} = Mohd Wazih Ahmad, D_{ma} = Dr. M.A. Ansari

In this DFD (Figure 1.) basic functions of information retrieval are given. First there is need to retrieve information. Then retrieval functions are used. In the end, information is retrieved.

In this level 1 DFD (Figure 2.), there are two inputs shown. First there is need of information. Then information formulation is done and query is sent. Database has indexed data items. They are represented in needed form. Retrieval functions use the user query to retrieve information.

In this level 2 DFD (Figure 3.), functions of information retrieval are further shown in detail. First there is information need. User goes to a retrieval interface. It parses the query. One of the many tools can be used to work on the query. Name of some useful tools for information retrieval are given. Those tools contact the database

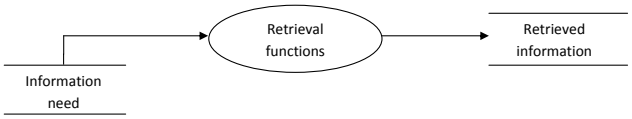

Figure 1. Level 0 DFD.

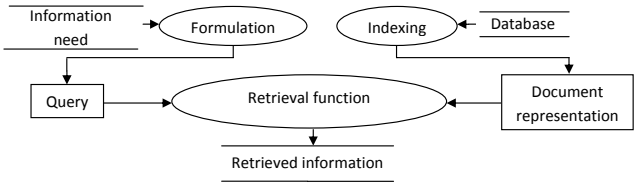

Figure 2. Level 1 DFD.

and then retrieval techniques applied to data. Names of some useful techniques for retrieval function are given.

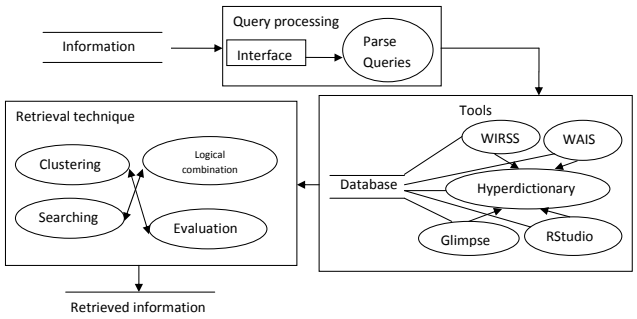

Figure 3. Level 2 DFD.

And after this the retrieved information is received in documented form.

3. Tools and Techniques

There are multiple tools available online for data mining and information retrieval. Some of the data mining tools can be used for information retrieval. Following is the list of some important tools with their release date, language in which they are written; either they are graphical user interface or command line interface, main purpose of the tool, various features of the tools and pros and cons of the tool. This helps in better understanding of the tools usage and in a particular environment for a particular purpose. Here is the list of some the most important data mining tools which are used mostly these days. Most of these tools are free to download and can be used for different purposes as given in the Table 13.

Data mining techniques are used to mine the data in a particular representation (such as in the form of clusters, graphs, tables, bar charts, box plots, etc.) and to mine some completely new information based on previous observations. There are a lot of data mining techniques available. Classification technique can be used to classify data based on some particular attributes. Clustering technique can be used to make clusters of data based on their properties. Neural networks are used to simulate intelligent activities of brain and to mine data based on certain rules and to mine data after performing certain kind of learning based on repetitions using this technique. Genetic algorithm can be used to mine data to get best results (with best attributes) out of the available dataset. These techniques are mainly implemented using data mining tools. Below is the list of important techniques that can be used. Their algorithms, subtypes, specific app, pros and cons are given in Table 14.

Table 13. Tools for data mining

N_{am}	R_{el}	L_{ae}	G_{ci}	M_{pe}	F_{es}	P_{cs}
R_{pr}	1993	C, F_{or}, R	Both	S_{cs}	S_{gt}, L_{nm}, C_{st}, T_{sa}, C_{af}, C_{lt}	F_{tu}, GNU GPL 2+
O_{ra}	1997	C++, P_{yt}, Qt	Both	G_{en}	M_{cl}, $B_{im}T_{tm}$, D_{ta}	O_{ps} for N_{ae}, GNU GPL 3
R_{ap}	2001	Java	GUI	G_{en}	A_{at}, M_{cl}, D_{tm}, T_{tm}, P_{da} and B_{sa}	O_{ps}, F_{tu}
R_{att}	2009	R_{pr}	GUI	E_{tt}	D_{tt}	F_{tu}, O_{ps}, GNU GPL v2
W_{ek}	2015	Java	GUI	G_{en}, M_{cl}	D_{pp}, C_{af}, C_{lt}, R_{gs}, A_{sr} and V_{il}	O_{ps}, GNU GPL

R_{ap} = Rapid Miner, W_{ek} = Weka, R_{pr} = R programming, F_{or} = FORTRAN, O_{ra} = Orange, R_{att} = Rattle GUI, G_{en} = General data mining, A_{at} = Offers advanced analytics through template-based framework, N_{am} = Name, R_{el} = Release, L_{ae} = Language, G_{ci} = GUI/CLI, M_{pe} = Main purpose, F_{es} = Features, P_{cs} = Pros and cons

4. Proposed Work

To make the system more robust and choice based, an improved version of algorithm in term of, freedom of selective database (through the process of normalization) has been achieved and proposed below.

Proposed algorithms
Begin:
Involve C1, C2, C3
Where C1∈ P1,C2∈ P2,C3∈ P3
//P1 and P2 are properties; C1 and C2 are the clusters.
for (Cn=1; Cn<Cl; Cn++)
// Where n=1 to 3
{
Apply mapping through fuzzy ontology
}
for (Kn=1; Kn<Knl; Kn++)
{
for (Dn=1; Dn<Dnl; Dn++)
{
Function Generation::Function Normalization
Initialize R_{SET}
// RSET is raw data set
}
Select $Data_{(Proc)}$
data processing
}

Table 14. Techniques for data mining[21]

N_{am}	A_{lg}	S_{ue}	S_{pa}	P_{ro}	C_{on}
A_{rt}	K_{ca}, L_{ea}	B_{ac}	S_{mp}, T_{sp}, etc.	S_{ti}, N_{nh}	N_{nr}, H_{nc}, etc.
D_{ec}	ID3	A_{ec}	N_{av}	F_{to}	O_{de}
C_{la}	G_{em},F_{oc}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	D_{ct} and F_{is}	C_{di}	T_{cc}
C_{la}	O_{li}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	I_{ft}	D_{yu}	L_{sm}, T_{cc}
C_{la}	V_{fd},C_{vf}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	D_{ct}	H_{sl}	N_{cd}, T_{cc}
C_{la}	L_{wc}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	C_{cw}	H_{sl}	N_{cd}, T_{cc}
C_{la}	CDM	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	D_{ct} and B_{yn}	S_{de}	U_{ic}
C_{la}	O_{ds}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	U_{mc}	H_{sl}	T_{cc}
C_{la}	E_{bc}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	U_{cd}	S_{pd}, C_{da}, D_{yu}, H_{ia}	L_{sm}, T_{cc}
C_{la}	A_{nn}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	I_{cc}	D_{yu}	L_{sm}, T_{cc}
C_{la}	S_{ca}	$C_{dt,}$ $B_{ay,}$ $N_{nt,}$ $S_{vm,}$ C_{lb}	S_{cd}	D_{yu}	L_{qs}
C_{lu}	S_{tr}, L_{oc}	$P_{am,}$ $H_{am,}$ $D_{bm,}$ $G_{bm,}$ M_{bm}	K-Medians	I_{cl}	L_{qs}
C_{lu}	C_{lu}	$P_{am,}$ $H_{am,}$ $D_{bm,}$ $G_{bm,}$ M_{bm}	C_{pm}	T_{sc}, D_{rd}, H_{ia}	O_{fc}
C_{lu}	D_{st}	$P_{am,}$ $H_{am,}$ $D_{bm,}$ $G_{bm,}$ M_{bm}	$D_{bm}C_{lu}$	H_{qc}, D_{rd} in R_{ds}	H_{ic}
C_{lu}	A_{ws}	$P_{am,}$ $H_{am,}$ $D_{bm,}$ $G_{bm,}$ M_{bm}	P_{re}	E_{pd}, H_{sl}, D_{yu}, S_{pd}	H_{ic}
C_{lu}	H_{ps}	$P_{am,}$ $H_{am,}$ $D_{bm,}$ $G_{bm,}$ M_{bm}	P_{bc}	E_{dd}, I_{nu}, H_{is}	H_{ic}

A_{rt} = Artificial neural network, D_{ec} = Decision trees, C_{lu} = Clustering, C_{la} = Classification

```
If Data(Proc)= Data(Pre-Proc)
// Data(Pre-Proc)is data pre processing
        {
Create DDT
    }
        for (DDT(n)=1, DDT<DDT(n) , DDT(n)++)
Create DDT
    //DDT(n)Discrete Decision Table for n attributes
}
    for (DDT(n)=1, DDT<DDT(n) , DDT(n)++)
        {
        RDT : Reduce Decision Table
        }
for (RDT=1, RDT< RDT(n) , RDT(n)++)
    //Dn= Data set
{
    Create Dn
    }
        Perform Normalization for Dn
Select database of choice Dc;
}
    //Where Dc∈ Kn
    // where Dc∈ Dn
}
If Dc = Dcl
(Exit)
Else
Perform clustering
}
End
```

5. Conclusion

In present era, there exist many applications which require management of large amount of data. Information retrieval from the huge database is the most relevant activity using different techniques proposed till date which provides accurate results. The proposed technique puts forward a scope to retrieve data in much efficient way as compared to others. This proves to be beneficial in the cases where proficient information retrieval is required. Development of technique which provides higher accuracy in the existing ones is sought afterwards.

6. References

1. Mooers C. Choice and coding in information retrieval systems. Transactions of the IRE Professional Group on Information Theory. 1954; 4(4):112–8.

2. Delminger M, Marsden WF. Development of an information retrieval system for an electronics R&D laboratory. IEEE Transaction on Engineering Writing and Speech. 1963; 6(1):10–9.
3. Sarin RK, Sicherman A, Nair K. Evaluating proposals using decision analysis. IEEE Transaction on Systems, Man, and Cybernatics. 1978; 8(2):128–31.
4. Stafylopatis A, Likas A. Pictorial information retrieval using the random neural network. IEEE Transactions on Software Engineering. 1992; 18(7):590–600.
5. Wives LK, Loh S. Hyperdictionary: A knowledge discovery tool to help information retrieval. 1998 Proceedings String Processing and Information Retrieval: A South American Symposium; Santa Cruz de La Sierra. 1998. p. 103–9.
6. Michael CH, Gey F, Piramuthu S. Data mining and information retrieval. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, HICSS'02; 2002. p. 841–2.
7. Yao JT, Yao YY. Web-based information retrieval support systems: Building research tools for scientists in the new information age. 2003 Proceedings IEEE/WIC International Conference onWeb Intelligence,WI' 03; 2003. p. 570–3.
8. Pereira AR, Baeza-Yates R. Applications of an web information mining model to data mining and information retrieval tasks. 2005 Proceedings 16 International Workshop on Database and Expert Systems Applications; 2005. p. 1031–5.
9. Menzies T, Chen Z, Hihn J, Lum K. Selecting best practices for effort estimation. IEEE Transactions on Software Engineering. 2006; 32(11):883–95.
10. Hao Y, Zhang Y-F. Research on knowledge retrieval by leveraging data mining techniques. In 2010 International Conference on Future Information Technology and Management Engineering (FITME); Changzhou. 2010. p. 479–84.
11. Farhadi F, Sorkhi M, Hashemi S, Hamzeh A. An effective expert team formation in social networks based on skill grading. 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW); Vancouver, BC. 2011. p. 366–72.
12. Ahmad MW, Ansari MA. A survey: Soft computing in intelligent information retrieval systems. 2012 12th International Conference on Computational Science and Its Applications (ICCSA); Salvador. 2012. p. 26–34.
13. Adda M. A formal model of information retrieval based on user sensitivities. Procedia Computer Science. 2013; 19:428–36.
14. Dankar FK, El Emam K, Matwin S. Efficient private information retrieval for geographical aggregation. Procedia Computer Science. 2014; 37:497–502.
15. Sklyarov V, Skliarova I, Silva J, Sudnitson A, Rjabov A. Hardware accelerators for information retrieval and data mining. 2015 International Conference on Information and Communication Technology Research (ICTRC); Abu Dhabi. 2015. p. 202–5.
16. Garimella RM, Gabbouj M, Ahmad I. Image retrieval: Information and rough set theories. Procedia Computer Science. 2015; 54:631–7.
17. Balasubramaniam K. Hybrid fuzzy-ontology design using FCA based clustering for information retrieval in semantic web. Procedia Computer Science. 2015; 50:135–42.
18. Remi S, Varghese SC. Domain ontology driven fuzzy semantic information retrieval. Procedia Computer Science. 2015; 46:676–81.
19. Chithra S, Sinith MS, Gayathri A. Music information retrieval for polyphonic signals using hidden markov model. Procedia Computer Science. 2015; 46:381–7.
20. Rao B, Mitra A, Mondal J. Algorithm for retrieval of sub-community graph from a compressed community graph using graph mining techniques. Procedia Computer Science. 2015; 57:678–85.
21. Kholghi M, Hassanzadeh H, Keyvanpour MR. Classification and evaluation of data mining techniques for data stream requirements. 2010 International Symposium on Computer Communication Control and Automation (3CA); Tainan. 2010. p. 474–8.
22. Shankr R, Sundarajan M. Manufacturing quality improvement with data mining outlier approach against conventional quality measurements. Indian Journal of Science and Technology. 2015; 8(15).
23. Nagarajan S, Chandrasekaran RM. Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques. Indian Journal of Science and Technology. 2015; 8(8).
24. Purusothaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology. 2015; 8(12).
25. Suganya P, Sumathi CP. A novel metaheuristic data mining algorithm for the detection and classification of parkinson disease. Indian Journal of Science and Technology. 2015; 8(14).
26. Kalaiselvi C, Nasira GM. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology. 2015; 8(14).
27. Shankar R, Sundararajan M. Manufacturing quality improvement with data mining outlier approach against conventional quality measurements. Indian Journal of Science and Technology. 2015; 8(15).
28. Lohita K, Sree AA, Poojitha D, Devi TR, Umamakeswari A. Performance analysis of various data mining techniques in the prediction of heart disease. Indian Journal of Science and Technology. 2015; 8(35).
29. Azad C, Jha VK. Data mining based hybrid intrusion detection system. Indian Journal of Science and Technology. 2014; 7(6).

30. Rajalakshmi V, Mala GSA. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014; 7(7).
31. Ananthapadmanabhan IR, Parthiban G. Prediction of chances - diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology. 2014; 7(10).
32. Manikandan G, Sairam N, Sharmili S, Venkatakrishnan S. Achieving privacy in data mining using normalization. Indian Journal of Science and Technology. 2013; 6(4).
33. Murugananthan V, Kumar BLS. An adaptive educational data mining technique for mining educational data models in elearning systems. Indian Journal of Science and Technology. 2016; 9(3).
34. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journal of Science and Technology. 2016; 9(3).
35. Chakradeo SN, Abraham RM, Rani BA, Manjula R. Data mining: Building social network. Indian Journal of Science and Technology. 2015; 8(2).
36. Noersasongko E, Julfia FT, Syukur A, Purwanto, Pramunendar RA, Supriyanto C. A tourism arrival forecasting using genetic algorithm based neural network. Indian Journal of Science and Technology. 2016; 9(4).

Appendix

N_{at} = Nature, S_{tp} = Stone tablets and papyrus/paper, A_{eg} = Ancient Egyptians, Greeks and Romans, P_{uc} = Punch cards, B_{bj} = Basile Bouchon and Jean- Baptiste Falcon, F_{ra} = France, P_{cr} = Punch cards read by machine, H_{ho} = Herman Hollerith, A_{me} = America, F_{ic} = Filing cabinet, H_{eb} = Henry Brown, G_{er} = Germany, H_{ad} = Hard disk, R_{ej} = Reynold Johnson IBM, F_{dk} = Floppy disk (8in), Y_{on} = Yoshiro Nakamatsu, J_{ap} = Japan, S_{sd} = Solid state hard drive, I_{nt} = Intel, C_{al} = California, C_{od} = Compact Disk, J_{ar} = James Russell, O_{nb} = Online backup, T_{bl} = Tim Berners-Lee, B_{ri} = Britain, S_{pp} = Philips, Sony, Toshiba, and Panasonic, P_{fs} = Portable flash storage, A_{do} = Amir Ban, Dov Moran and Oron Ogdan, I_{sr} = Israel, J_{rl} = Joseph Carl Robnett Licklider, H_{qs} = Holographic layers and quantum storage, U_{kn} = Unknown, D_{vd} = D.V.D. (4GB), F_{tr} = Future,D_{sp} = Dijkstra's shortest path, F_{mp} = Fuzzy mathematical programming model, R_{fa} = RarestFirst algorithm, E_{sa} = Enhanced Steiner algorithm, G_{da} = Generalized Diameter algorithm, S_{ga} = Skill grading algorithm, B_{tt} = Bibsonomy tag tool, M_{ce} = Minimizing the communication cost among experts, P_{ir} = Pictorial information retrieval using random neural network, A_{sp} = Andreas Stafylopatis, A_{rl} = Aristidis Likas, N_{nt} = Neural network techniques, I_{pt} = Image processing techniques, P_{ht} = Proposed Hierarchical techniques, A_{mo} = Autoassociative memory operation during pictorial information retrieval, l_{rp} = several learning and recall phases can be deviced, H_{ap} = Hierarchical approach used need further experimentation, S_{oa} = Other algorithms can be considered for increasing capacity, R_{kd} = Research and knowledge retrieval by leveraging data mining techniques, J_{ay} = Jame You, T_{dm} = Tharam Dillon, J_{il} = James Liu, Spu= Selwyn Piramuthu, H_{mc} = H. Michael Chung, J_{ty} = J. T. Yao, Y_{yy} = Y. Y . Yao, C_{nm} = Calvin N. Mooers, A_{rp} = Alvaro R. Pereira Jr, R_{by} = Ricardo Baeza-Yates, Y_{ha} = Yan Hao, Y_{fh} = Yu-feng Zhang, D_{el} = DELI, D_{wh} = Data warehousing techniques, D_{at} = Data aggregation techniques, D_{mt} = Data mining techniques, H_{si} = Technique for handing semantic information, P_{ag} = Proposed algorithm, I_{ra} = Information retrieval algorithms, A_{pa} = Apriori algorithm, K_{ra} = Knowledge retrieval algorithm, A_{rn} = Association rule mining algorithm, C_{ra} = Concept retrieval algorithms, C_{rl} = classification retrieval algorithm, I_{la} = Inductive learning algorithm, R_{el} = retrieval algorithm, A_{re} = Association rule extraction algorithm, C_{ls} = Classification algorithm, R_{dr} = Rapid development of research and applications of data mining, M_{mg} = New model to mine graph applications, R_{ss} = Retrieval systems' applicability for development of signalling systems, S_{ih} = Find new techniques for handing semantic information in retrieval, S_{fr} = More structures and functionalities for research work, C_{ef} = There is need to continually extend functionality and integrate different systems, D_{ir} = Various data mining and information retrieval papers discussed, M_{dw} = New approach of multimedia data warehousing to increase flexibility, efficiency, integration, indexing, I_{ce} = Identify context more easily, H_{kw} = Sometimes Hard to know what user really wants, M_{cl} = Machine learning, D_{tm} = Data mining, T_{tm} = Text mining, P_{da} = Predictive analytics, B_{sa} = Business analytics, O_{ps} = Open source, P_{yt} = Python, F_{tu} = Free to use, D_{pp} = Data preprocessing, C_{af} = Classification, C_{lt} = Clustering, R_{gs} = Regression, A_{sr} = Association rules, V_{il} = Visualization, S_{gt} = Statistical and graphical techniques, L_{nm} = Linear and nonlinear modeling, T_{sa} = Time-series analysis, C_{st} = Classical statistical tests, S_{cs} = Scientific computation and statistics, N_{ae} = Novice and experts, B_{im} = Bioinformatics, D_{ta} = Data analytics, E_{tt} = To edit and test data and teach R language, D_{tt} = Dataset can be partitioned for training, validation and testing, , N_{nt} = Neural Networks, S_{vm} =

Support vector machines, C_{lb} = Classification based on association, P_{am} = Partitioning methods, H_{am} = Hierarchical agglomerative (divisive) methods, D_{bm} = Density based method, G_{bm} = Grid-based methods, M_{bm} = Model-based methods, B_{ac} = Back Propagation, K_{ca} = Kohonen clustering Algorithm, L_{ea} = learning algorithms, S_{mp} = Stock market prediction, T_{sp} = Travelling salesman problem, S_{ti} = Simple to implement, N_{nh} = Neural networks often exhibit patterns similar to those exhibited by humans, N_{nr} = Neural networks cannot be retrained, H_{nc} = Handling of time series data in neural networks is a very complicated, A_{ec} = Apply to ethical considerations, F_{to} = they force you to think as many outcomes, as you can think of, O_{de} = Outcomes of decisions may be based primarily on expectations and not related to actual decisions, D_{ct} = Decision tree, F_{is} = Frequent item sets, C_{di} = Concept drift detection incremental mining models, T_{cc} = Time consuming and costly learning, I_{ft} = Uses info-fuzzy techniques for building a tree-like classification model, D_{yu} = Dynamic Update, L_{sm} = Low speed Storage memory problem, H_{sl} = High speed Need less memory space, N_{cd} = Non-adaption to concept drift, N_{cd} = Non-adaption to concept drift, C_{cw} = Classification based on classes weights, B_{yn} = Bayes network, S_{de} = Suitable factor for distance measurement between events, U_{ic} = User defined information complexity, U_{mc} = Using micro-clusters ideas that each micro-cluster is associated with a specific class label which defines the class label of the points in it, S_{pd} = Single pass , D_{yu} = Dynamic update, C_{da} = Concept drift adoption, H_{ia} = High accuracy, U_{cd} = Using combination of different classifiers, I_{cc} = Incremental classification, I_{cl} = Incremental learning, L_{qs} = Low clustering quality in high speed and low accuracy, S_{cd} = Scalable classification for numerical data streams, C_{pm} = Concepts of a pyramidal time frame in conjunction with a micro-clustering approach, T_{sc} = Time and space efficiency concept, D_{rd} = Drift detection, O_{fc} = Offline clustering, R_{ds} = Real time data stream, H_{qc} = High quality and efficiency Concept, H_{ic} = High complexity, E_{pd} = Efficient pattern detection, E_{dd} = Efficient for high dimensional data stream, I_{nu} = Incremental update, H_{is} = High scalability, O_{ds} = On-demand stream classification, E_{bc} = Ensemble-based Classification, A_{nn} = ANNCAD, S_{ca} = SCALLOP, S_{tr} = STREAM, L_{oc} = LOCALSEARCH, C_{lu} = CluStream, A_{ws} = AWSOM, H_{ps} = HPStream, P_{re} = Prediction, P_{bc} = Projection based clustering, D_{st} = D-Stream, M_{ed} = Medical, G_{ri} = Grid, S_{em} = Semantic, I_{nt} = Information integration, T_{ef} = Teaching files, O_{nc} = Ontology construction, D_{at} = Database, O_{bf} = OBFM, A_{sr} = Association rule, D_{ot} = Document type, U_{pe} = Upgrade, B_{ps} = Best practices in software effort estimation, T_{mn} = Tim Menzies, Z_{ch} = Zhihao Chen, J_{hn} = Jairus Hihn, K_{lm} = Karen Lum, C_{ot} = Coseekmo tool, L_{rt} = Linear regression tools, T_{bt} = Regression based technique, M_{gt} = Model generation technique, R_{ba} = Regression based algorithms, W_{sa} = Wrapper attribute selection algorithm, Q_{ma} = Quinlan's M5P algorithm, B_{ee} = Best effort estimation method of 1901, N_{id} = Need to include difinitions, N_{bm} = need for better modelling methods, E_{pd} = Evaluating proposals using decision analysis, R_{ks} = R.K. Sarin, A_{si} = A. Sicherman, K_{na} = K. Nair, D_{at} = Decision analysis technique, L_{it} = lottery indeferences, T_{if} = tradeoff indifferences, W_{rt} = weighting and rating technique, E_{re} = Evaluation and ranking become easy, E_{dh} = evaluators can make decisions and show them to higher authorities, M_{et} = multiple evaluators can work together, A_{dm} = Only appropriate for decision making, N_{po} = not appropriate for political and organizational criteria, S_{ii} = A survey: Soft computing in Intelligent Information Retrieval Systems, P_{mt} = Performance measure tools, M_{td} = mathematical tools to discover redundancy and dependency, S_{ct} = Soft computing technique, R_{ag} = rank aggregation technique, S_{qm} = Search quality measuring techniques, T_{if} = technique for implicit feedback, P_{et} = performance evaluation technique, W_{st} = web search technique, A_{it} = AI techniques, I_{da} = intelligent data analysis technique,D_{mt} = data mining techniques, C_{mt} = content mining techniques,W_{um} = web usage mining techniques, I_{rt} = information retrieval techniques, P_{bt} = probabilistic based techniques, G_{em} = Genetic algorithm, F_{ma} = fuzzy matching algorithm, P_{ai} = probabilistic algorithm, I_{cn} = intelligent classification algorithm, M_{sa} = meta-search algorithms, I_{ra} = IR algorithm, C_{ma} = current mining algorithm, F_{sa} = fuzzy search algorithm, B_{sa} = Boolean search algorithm, B_{qp} = Boolean query processing algorithm, Dsi = In depth survey on intelligent information retrieval, M_{pd} = Mind programmed, N_{sh} = neuro-sasisfaction and hybrid solutions for soft computing based information retrieval not discussed, A_{tp} = A Survey: Soft computing in Intelligent Information Retrieval Systems, K_{aa} = Kulkarni Ashish A, J_{ay} = Jagannath Aghav, L_{in} = Lint, F_{bs} = Findbugs, G_{oa} = Goanna, Pmd, S_{at} = Static analysis techniques, P_{ft} = profile feedback technique, P_{at} = program analysis technique, C_{at} = code analysis technique, W_{as} = Weiser's algorithm, S_{am} = slicing algorithm, G_{ra} = graph reachability algorithm, G_{ka} = Gen/Kill algorithms, P_{mr} = Program analysis help in software analysis, manintenance and reengineering, H_{ai} = Hardware

accelerators for Information retrieval and data mining, V_{sy} = Valery Sklyarov,I_{os} = I ouliia Skliarova, J_{sa} = Joao Silva, A_{sn} = Alexander Sudintson, A_{rr} = Artjom Rjabov, S_{dk} = Software development kit, V_{ds} = Vivado Design suite, D_{ma} = Data mining algorithms, M_{cp} = Method for clustering objects described in paper, V_{ld} = very large data sets can be processed, E_{xp} = Expandability, C_{on} = Convenience, E_{co} = Economics, E_{tn} = Etched noted cards, U_{tc} = Uniterm cards, O_{ci} = Optical coincidence, P_{uc} = Punched cards, C_{om} = Computers, A_{hi} = Ancient history, C_{es} = Cerebral storage, M_{at} = Magnetic tape, F_{pf} = Fritz Pfleumer, M_{ea} = Mean, C_{or} = Corelation, T_{cl} = The cloud, C_{ci} = Choice and coding in information retrieval system, C_{nm} = Calvin N. Mooers, R_{sc}= Retrieval function susceptible for communication theory, R_{ds} = Retrieval function good for development of signalling system, T_{si} = Technique for handling semantic information not discussed, D_{ir} = Development of information retrieval system for an electronics R&D technology, M_{dr} = Merlin delminger, W_{fm} = William F. Marsden, C_{ig} = Coordinate indexing, I_{ts}= Indexing techniques, T_{fr} = techniques for reconstructing, S_{ct} = superimposed coding techniques, O_{rp} = Organization and retrieval of organization's internal publication, H_{ti} = Hyperdictionary: a knowledge discovery tool for information retrieval, L_{kw} = Leandro Krug Wives, S_{lh} = Stanley Loh, A_{va} = Altavista, T_{ir} = Traditional technique of information retrieval, A_{fw} = Algorithm find relationship among words automatically, A_{ic} = Automatic identification of context of user's query, A_{wd} = Applications of an web information mining model to data mining and information retrieval tasks, A_{rp} = Alvaro R. Pereira Jr, R_{by} = Ricardo Baeza-Yates, N_{mg} = New model to mine information in graph applications, F_{iu} = Formal model of information retrieval based on user sensitivities, M_{ia} = Mehdi Adda, I_{bu} = Information retrieval based on user sensitivities, S_{br} = Sensitivity based recommender system to explore, C_{sd} = combine search with sensitivity based search, P_{ia} = Efficient private information retrieval for geograpical aggregation, $_{kd}$ = Fida K. Dankar, K_{ee} = Khaled El Emam, S_{mn} = Stan Matwin, G_{aa} = Geographical aggregation algorithm, E_{ps} = Efficient protocol to privately query server's database, H_{fi} = Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web, B_{al} = Balasubramaniam K, T_{mt} = Text matching tool, P_{et} = Protege tool, F_{nt} = Fuzzy ontology tool, F_{og} = FOGA technique, F_{bc} = FCA based clustering technique, K_{ot} = Keyword optimization techniques, K_{pm} = Keyword pattern matching algorithm, E_{xa} = Expansion algorithm, P_{tk} = Proposed term keyword algorithm, G_{ii} = Good interpretation of information, D_{aa} = Need of document annotation algorithm, M_{ir} = Music information retrieval for polyphonic signals using Hidden Markov Model, S_{ca} = S Chithra, M_{ss} = M S Sinith, A_{gi} = A Gayathri, F_{ta} = Fundamental frequency tracking algirthms, F_{da} = Frequency domain algorithm, B_{wa} = Baum-Welch algorithm, G_{ma} = GEM algorithm, Ppr = Promising polyphonic music retrieval, D_{df} = Domain Ontology Driven Fuzzy Semantic Information Retrieaval, R_{es} = Remi S, V_{sc} = Varghese SC, S_{wt} = Semantic web tool, T_{ie} = Title, A_{us} = Authors, Y_{ea} = Year, T_{oo} = Tools, T_{ec} = Techniques, A_{lo} = Algorithm, F_{ee} = Features, F_{ws} = Future works = Semantic web tool, I_{ea} = Information extraction algorithm, S_{aa} = Semantic annotation algorithm, S_{er} = Semantic expansion reasoning algorithm, I_{rp} = Improve recall and precision in serach engines, I_{rp} = Improve recall and precision in serach engines, A_{rc} = Algorithm for Retrieval of Sub-Community Graph from a Compressed Community Graph using Graph Mining Techniques, B_{ro} = Bapuji Rao, A_{nm} = Anirban Mitra, J_{am} = Jayanta Mondal, R_{aw} = Random walk, G_{mt} = Graph mining technique, C_{da} = Community detection algorithms, S_{eg} = Simpler and efficient sub-community graph, C_{dt} = Classification by decision tree induction, B_{ay} = Bayesian classification, A_{lg} = Algorithm, S_{ue} = Subtype, S_{pa} = Specific app, P_{ro} = Pros, C_{on} = Cons, G_{em} = GEMM and F_{oc} = FOCUS, O_{li} = OLIN, V_{fd} = VFDT and C_{vf} = CVFDT, L_{wc} = LWClass