Received: May 11, 2018

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Novel Analysis and Prediction of Students' Behaviour Using Semantic Similarity-Based Improved J48 IL Algorithm in Personalized Library Ontology

Fernandez Mary Harin Fernandez [1]* **Ramalingam Ponnusamy [2]**

[1] *Department of Computer Science and Engineering,*
Sathyabama Institute of Science and Technology, India
[2] *Department of Computer Science and Engineering, CVR College of Engineering, India*
* Corresponding author's Email: mary.fherin@gmail.com

Abstract: The Semantic web deliberates machines to develop conceptual information themselves by understanding its implications. Semantic web creates overt resources through ontology. Ontology widely offers the representation of conceptual knowledge. It has the ability to signify the domain knowledge in a distinct and unambiguous manner. We propose Ontology-based structure for personalized library usage environment which discover user's behaviour by their learning feedback. The personalized library Ontology and user's feedback is developed using Protégé editor 4.3. To identify user behaviour we determine 225 students' library learning feedback which enhance the reference information for the future users. For data classification, we propose semantic similarity-based Improved J48 induction learning algorithm which facilitates the system to identify patterns and regularities for the extracted ontology data. Impurity from the classified data is evaluated using entropy and information gain. The new decision-making for ontology subpopulation is made by estimating the highest GainRatio. Finally, the system performance is evaluated with k-fold cross-validation, accuracy, F1-score, and execution time. The proposed algorithm reduces dissimilar data and handles missing data therefore, the system can improve its efficiency and performance with enhanced accuracy up to 10% - 12% compare to existing algorithms as Random Forest, Naive Bayes, Multilayer Perceptron and J48 algorithm. It can also reduce its execution time more than 20 (ms) for the different size of datasets compare to existing C4.5 algorithm. Theoretical analysis and comparison of existing algorithms with its experimental results and system evaluation show the effectiveness and performance of the proposed system.

Keywords: Personalized library ontology, Protégé editor 4.3, Classification, Improved J48 induction learning algorithm, User behaviour, Ontology evaluation.

1. Introduction

Ontology is an increasingly large volume of conceptual data representation. Large millions of people globally use WWWs' huge volume of repository information. Current days World Wide Web becomes very significant for gathering and sharing information and its services.

Distributing and retrieving information from web repository become more complex and facing difficulties to use required information from web repository. It gradually transforms web information into semantic web, which is the way of knowledge discovery from the database repository. This uses OWL ontology language to overcome current technology and difficulties. Such as storing, transforming, distributing, presenting, retrieving and maintaining the necessary information by the user.

1.1 Web ontology language (OWL)

To develop ontology, it is the necessity to have ontology languages. There are several ontology design languages; normally acknowledged a standard language is Web Ontology Language (OWL) for representing and sharing information in the semantic web [1]. OWL would use Resource Description Framework (RDF) inferences of classes, data and object properties and include the imperative structure of primitives to develop the

articulateness [2]. Protégé editor [3] is a tool which is used to build ontology with the allocation of classes, data properties, object properties, individuals and its association. This Protégé editor consents the structure of the domain ontology and personalizes information access construction to enter data.

1.2 User behaviour analysis

The user behaviour analysis is intended using survey, recommendation and user model system which use adaptive performance regarding its communication with the user [4-5]. The purpose of doing user behaviour analysis is to classify and predict their interest in future behaviour.

1.3 Decision tree induction algorithm

Induction machine learning generally uses decision tree approach which is easy and simple to implement a large set of data. Decision tree uses the available input dataset attributes to form its top-down tree hierarchy [6-7]. It works with numerical and categorical data with high classification efficiency.

1.3.1. ID3 algorithm

Iterative Dichotomister3 (ID3) is simple decision tree developing algorithm which classifies objects, introduced by Quinlan Ross in 1986. This decision tree examines each attribute in every node in a tree and builds tree hierarchy using the top-down approach based on divide and conquer strategy [8]. Information gain estimation is used to select the splitting attributes and accept categorical attributes for developing a tree. If any noise in the dataset the result of classification efficiency is less.

1.3.2. C4.5 algorithm

C4.5 also referred as J48 algorithm implemented in Java. It is an upgraded algorithm of ID3 introduced by Ross Quinlan in 1993 [9]. It is similar to ID3 algorithm but unlike ID3, C4.5 work with distinct, incessant, and categorical attributes, handles missing information and perform tree pruning [10]. For continues attribute's categorization, a threshold value is preferred. It uses gain ratio measurements for splitting attribute in the decision tree.

The proposed methodology is used to analyze user's behaviour by their learning feedback. To identify user behaviour we determine 225 students' library learning feedback. In order to enhance data classification, we classified 225 records as 65, 90, 195, and 225 records of datasets and each record has 12 attributes. Initially, we preprocess the input data using semantic similarity measure algorithm. Each attribute in the dataset are semantically analyzed and populated into personalized library ontology. The graphical representation of ontology sub-graphs are discretized and labelled with the corresponding semantic similarity measure. Ontology subpopulation is processed with entropy, information gain and gain ratio. The highest gain ratio is chosen to take new decision making for the new ontology subpopulation. To enhance system effectiveness and performance, the classified ontology data are evaluated using 10-fold cross-validation, accuracy, F1-score and execution time which is also compared with the existing algorithms.

The rest of the paper is positioned as follows: The related work and motivation behind the work are specified in division 2. Division 3 describes the proposed method for library ontology origination, simplified semantic similarity measure algorithm, and Improved J48 induction learning algorithm. The detailed examined outcomes and its considerations are illustrated in division 4. The conclusion and future enhancement are given in division 5.

2. Literature review

Ontology data mining finds out pertinent and useful information from the database and this mining used for organization and prediction of data in the database. Datasets classification is current challenges in research. User behaviour determines, classifying the behavioural data which predict user's behaviour. Some of the data classification existing algorithms are explained below.

Vani Kapoor Nijhawan, Mamta Madan and Meenu Dave, provided the analytical comparison of two algorithms ID3 and C4.5 using weka tool [11]. They analyzed both algorithms with three sets of data as 50, 100 and 150 records. Finally, they concluded that the C4.5 algorithms' classification accuracy is better compare to the ID3 algorithm.

Boufardea Evangelia and Garofalakis John proposed an ontology-based distance learning structure which develops a recommendation system to forecast learner's evolution and their ultimate performance [12]. They used J48 algorithm to classify the input data in weka tool with percentage split of 66% and cross-validation of 10 folds. The system accuracy for percentage split is 85.2% and for cross-validation is 79.7% respectively.

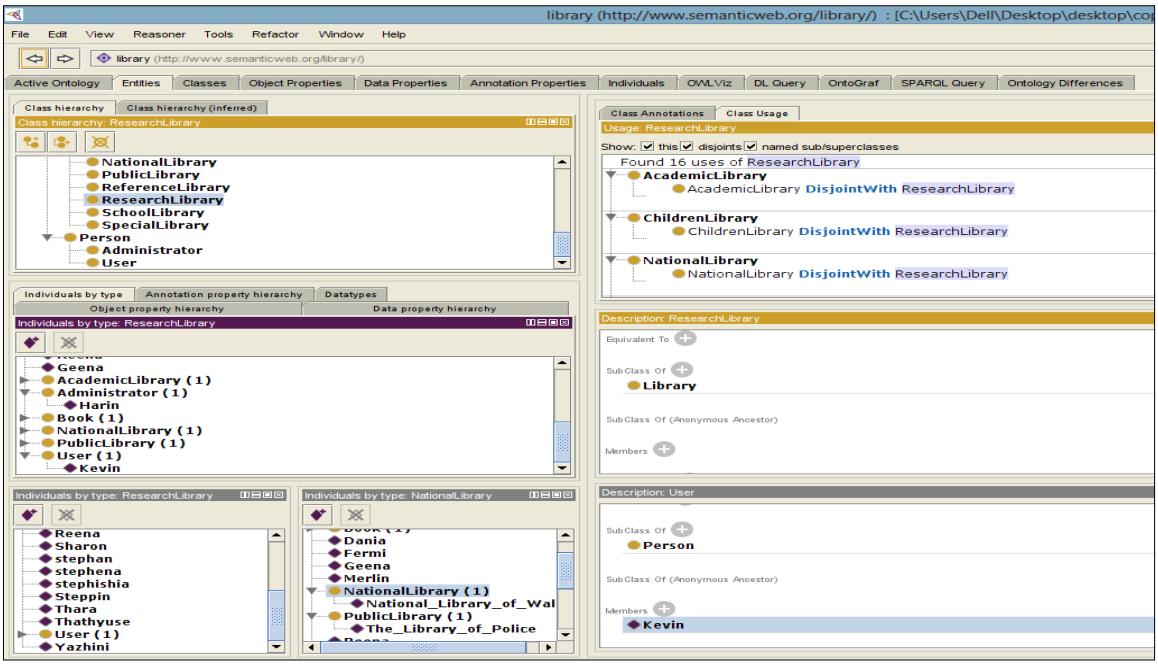

Figure. 1 Personalized library ontology using protégé editor 4.3

Jaimin Undavia, Atul Patel and Prashant Dolia, analyzed and compared students' performance using three algorithms such us Random Tree, J48 and SimpleCART in WEKA environment [13]. They worked out 128 instances in 10-folds cross-validation for forecast students' post-graduation course. Authors concluded that the J48 algorithm is better compared to Random Tree and SimpleCART, which correctly classified result as 68.75% with the execution time of 0.02 seconds.

Sukontip Wongpun and Anongnart Srivihok presented the experimental evaluation of four algorithms such as Naive Bayes classifier, C4.5 algorithm, Baysian Belief Network and RIPPER algorithm for classifying students' bad behaviors in vocational education [14]. They concluded that Correlation-based Feature Selection (CFS) evaluator with C4.5 algorithm gives the highest accuracy of 82.52% compared to genetic explores hybrid classification method. But the genetic search hybrid classification and Bayesian belief network give an improved F-measure precision value.

The above stated existing data classification algorithms are analyzed and many algorithms such as Naive Bayes classifier, Bayes Network, RIPPER, Random Tree, SimpleCART, ID3 and J48 (C4.5) decision tree algorithms are executed in WEKA workbench in which the input data are not semantically identified before its processing. To enhance ontology data classification performance and its accuracy, we propose Improved J48 induction learning algorithm which reduces dissimilar data using Simplified semantic similarity measure algorithm and handles missing data in order to improve the system's effectiveness and performance.

3. Proposed work

The users' learning behavioural ontology is build using Protégé editor 4.3. We took the survey of students learning style with different dimensions as such as perspective learning, processing learning, resource feedback and mood of learning for behavioural analysis. We build ontology as classes, subclasses and identify the properties of each class and its subclasses and then we defined individuals and logical relationship between its objects. Fig. 1 shows the representation of personalized library ontology with the users' learning feedback.

3.1 Simplified semantic similarity measure algorithm

The properties of each individual are analyzed with the semantic association using semantic-based similarity measure algorithm presented in algorithm 1. It takes input as a set of *Sg* and *q* represented as subgraphs and the query in highest ontology term. In algorithm, *n* is the number of subgraphs of ontology and *m* is the number of nodes in subgraph Sg_i. It computes data content, depth, local network density, semantic distance and similarity score in input data and it returns a set of *Sg* of ontology with the assigned semantic similarity measure score for each node in *Sg* terms. This score is calculated with the generation of preliminary population and

estimation of the fitness function. This algorithm is adapted from [15]. We used simplified semantic similarity measure algorithm which is explained in [16]. This algorithm calculates the likeness strength in Ontology terms.

Algorithm 1

SimplifiedSemanticSimilarityMeasure(S_g, q);

Input: Sg={s_{g1},s_{g2},…s_{gn}} and **q**

Output: Sg={s_{g1},s_{g2},…s_{gn}} (S_{gi} with similarity score)

```
begin
  for j:= 1 to n do
   for i:= 1 to m do
    Compute the data content DC (c^j_i );//where, c^j ϵ S_gi
    Compute the depth D (c^j_i );
    Compute the local network density ND (c^j_i );
    Compute the semantic distance SD (q, c^j_i );
    Compute the term similarity score SS (q, c^j_i );
   end-for
  end-for
end
```

3.1.1. Data content approach

The data substance is calculated according to the content association. This association stores annotation which affords relations between ontology terms and resource attributes. The data content of ontology terms *DC(c)* is signified as in Eq. (1).

$$DC(c) = -\log\big(Prob(c)\big) \tag{1}$$

Where *Prob(c)* is a probability of incidence of an ontology term *c* in the relationship, which is calculated using maximum likelihood inference as given below with occurrence of the term:

$$Prob(c) = \frac{f(c)}{N} \tag{2}$$

$$f(c) = \sum_{c \in descentdants(c_i)} occurrence(c_i) \tag{3}$$

In Eq. (2) *f(c)* is the number of times that *c* happens with its entire descendants and *N* is the total number of occurrences. The frequency of the ontology term *c* is given in Eq. (3). Here the *descendants(c)* are a function which returns a set of descendants' ontology terms *c*.

3.1.2. Conceptual distance approach

In ontology term, the conceptual distance is considered by the depth *D* and the local network density *ND* factors. The ontology depth is computed as the hierarchy distance of the term in ontology graph. In Eq. (4) *d(c)* represents the ontology term level in ontology graph. Ontology graph root depth starts with the term c_0 as 1 and it increases as the height of the ontology term decreases in the graph hierarchy. Here the parameter α controls the degree of graph depth, and $\alpha \geq 0$.

$$D(c) = \left(\frac{d(c)+1}{d(c)}\right)^{\alpha} \tag{4}$$

$$E(c) = \left(1 - \beta \frac{\bar{E}}{e(c)}\right) + \beta \tag{5}$$

In Eq. (5), *E(c)* denotes the local network density of ontology. Where $\bar{E}$ is the number of edges that is separated by the number of ontology terms *e(c)* in the ontology graph and *e(c)* is the number of edges that begin from the ontology term *c*. Here the parameter β controls the degree of local network density in ontology graph and $0 \leq \beta \leq 1$.

3.1.3. The hybrid approach

This approach is derived from the perception of the conceptual distance and by integrating the data content as a decision aspect. In Eq. (6) the hybrid approach calculates the semantic distance involving ontology terms c_1 and c_n. Where $c_1, \ldots, c_n$ represents the sequential path of ontology terms with the length *n*. The function *distance* (c_1, c_n) returns the sum of edge weights with the shortest path that associates c_1 with c_n.

$$distance(c_1, c_n) = \sum_{i=0}^{n-1} D(c_i) \times E(c_i) \times (DC(c_{i+1}) - (DC(c_i)) \tag{6}$$

The semantic distance is computed between the ontology terms c_m and c_n with the below-given Eq. (7). Here ontology term c_1 is the adjacent collective ancestor of ontology terms c_m and c_n.

$$distance(c_m, c_n) = distance(c_1, c_m) + distance(c_1, c_n) \tag{7}$$

$$distance_{norm}(c_m, c_n) = min\left\{1, \frac{distance(c_m, c_n)}{max\{DC(c)\}}\right\} \tag{8}$$

$$similarity(c_m, c_n) = 1 - distance_{norm}(c_m, c_n) \tag{9}$$

To find semantic distance from the adjacent collective ancestor of ontology terms c_m and c_n, the sum of minimum semantic distance score of *distance*(c_1, c_m) and *distance*(c_1, c_n) are measured. Eq. (8) calculates the semantic distance between differences of data content with the normalization of

the semantic distance. Similarity score for each ontology terms is measured between c_m and c_n by adapting the semantic distance as given in Eq. (9).

3.2 Improved J48 induction learning algorithm

The Improved J48 induction learning algorithm utilizes the simplified semantic similarity measure algorithm to compute the ontology term similarity score. This score link each ontology attribute terms and the query ontology terms.

3.2.1. Preprocessing

The proposed systems' flow diagram is exposed in Fig. 2. The preprocessing of ontology term is done by the similarity measure algorithm to develop the feature of the attributes.
Here each attribute's values are semantically analyzed and populated into personalized library ontology. The ontology conceptual representation of the graph is separated into numerous subgraphs in order to compute the semantic similarity score for each ontology terms and to create the primary population simple and quicker. The experiment model of Improved J48 induction learning classification decision tree algorithm is given in algorithm 2.

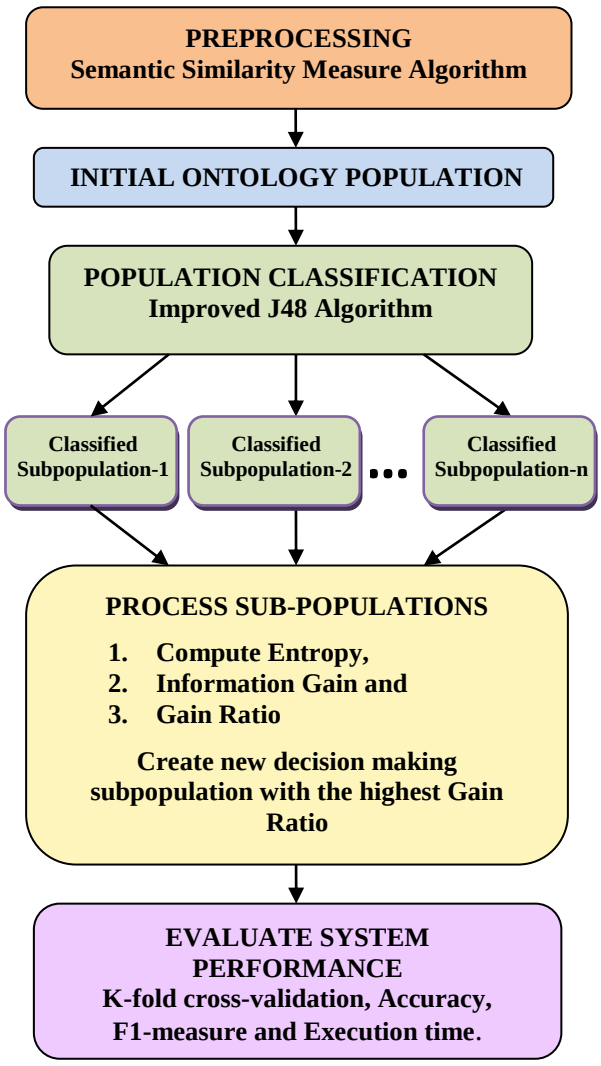

Figure. 2 Proposed system's flow diagram

This takes input as ontology subgraph *Sg* with semantic score. Before processing the attribute classification, the continuous attributes in subgraphs are discretized and labelled with the corresponding semantic linguistic labels and the missing attribute instances are handled by placing the average value of its semantic similarity score. Finally, the classified data are evaluated using 10-fold cross-validation, accuracy, F1-score and execution time.

Algorithm 2

Input: Ontology graph with similarity score
Output: Classified attribute data.

1. Input ontology subgraphs with semantic similarity score $S_g = \{s_{g1}, s_{g2}, \ldots s_{gn}\}$.
2. If continuous attribute in S_g then,
   Discretize that attributes $a=1,2,3,\ldots$
   Using semantic linguistic labels with the highest compatibility of input values.
3. If any missing value v in attribute a_m then,
   Set v value in a_m using the average value of semantic similarity measure score of a_m with semantic linguistic labels.
4. Evaluate the entropy outcome and information gain of each attribute to split training set.
5. Define the test node of the tree until all training sets are classified.
6. If the attribute instances are incorrectly classified then, Repeat step 2 to 5 until it classifies all values correctly.
7. Apply 25% of confidence limits for a post-pruning process as default.
8. Calculate the accuracy and execution time.

3.2.2. Information gain and gain ratio

The semantically linguistic labelled data are alienated to calculate impurity degrees between the attributes. The information theory computes to choose the accurate attribute to split the input data. The evaluated dissimilarity of impurity degrees is described as information gain.

The J48 algorithm exploits entropy to determine the degree of impurity in the input subgraph. If an attribute has a distinct value then the preferred attribute is homogenous else it is heterogeneous.

Table 1. Different aspects of student's learning behavioural patterns

Processing Dimension	Perception Dimension	Input Dimension	Resource Cost	Resource Quality	Specification	Mood	Resource Feedback
Medium Reflective	Neutral	NIL	Very High	Poor	Not upto the Expectation	Very Sad	Completely Irrelevant
Medium Active	Medium Sensitive	Poor	High	Can Improve	Not Good	Sad	Not Interested
Neutral	Medium Intuitive	Neutral	Unsatisfactory	Better	Fine	Not so Happy	Average
Active	Extremely Sensitive	Medium Visual	Satisfactory	Good	Good	Happy	Relevant
Extremely Active	Extremely Intuitive	Extremely Visual	Good	Excellent	Excellent	Very Happy	Good

Entropy distinguishes the impurity of the subjective set of examples. It estimates attribute values between dissimilar classes. The entropy calculated in Eq. (10), where $Prob_i$ denotes the likelihood of randomly acquiring one value as an i^{th} value from the set, R is all occurrences in the dataset and N is a number of different values in a set. The result of entropy is zero when the probability is 1 and $log_2(1) = 0$. The entropy attains the highest value when all data classes have the same probability.

$$Entropy(R) = -\sum_{i=1}^{N} Prob_i log_2 Prob_i \quad (10)$$

The information gain computes the predictable reduction in entropy reasoned by separating the instances according to the specified aspect. The computation of information gain is given in Eq. (11) where R denotes overall dataset, $|R_k|$ is a number of occurrences with k value of an attribute *Attr*, $|R|$ is a sum of occurrences in dataset R, n is a set of distinct values of an attribute *Attr* and $Entropy(R_k)$ denotes the subset of occurrences for attribute *Attr*. Information gain determines the reduction in entropy attained because of the correct split. We choose the accurate attribute split by accomplishing the most reduction with maximum information gain.

$$GAIN(Attr, R) = Entropy(R) - \left(\sum_{k=1}^{n} \frac{|R_k|}{|R|} Entropy(R_k)\right) \quad (11)$$

$$GainRatio = \frac{GAIN(Attr,R)}{SplitInfo} \quad (12)$$

$$SplitInfo = -\sum_{i=1}^{k} \frac{n_i}{n} log \frac{n_i}{n} \quad (13)$$

Gain ratio deals with the alteration of information gain. The information gain gets biased when the number of value is higher in the count and results in overfitting. This overfitting should be reduced by adjusting the value of information gain which happens by considering intrinsic information of a subset of class. The gain ratio is calculated using Eqs. (12) and (13), here *SplitInfo* value is efficient for normalization of information gain. The utmost value of *GainRatio* is chosen as a decision node.

4. Experiment result and discussion

An ontology-based structure for the personalized library is proposed to identify user's behaviour by their learning feedback. Fig. 1 shows personalized library ontology which is developed using Protégé editor 4.3. The proposed system is developed using the emotional recommendation dataset obtained from the students while exploiting the library resources. The student's behaviours are predicted and databases are constructed on these estimated behaviours in order to create ontology structure. The recommended student's behaviour dataset consists of 225 records and each record includes 12 attributes. The student's behavioural activities are analyzed with the different aspect of students' accessing library learning behaviour patterns shown in Table 1.

The classified subpopulation ontology data is evaluated using entropy measurement. The entropy differentiates the impurity of the subjective set of data and estimates the attribute values between dissimilar classes.

The information gain calculates the conventional diminution in entropy logically by separating the instances according to the specific feature. The information gain and gain ratio are estimated using Eqs. (11), (12) and (13) for the different size and sets of data shown in Table 2.

Here we use 2 different sizes of datasets as 65 records and 90 records and each record contains 12 attributes. To determine the diminution in entropy, we select the exact attribute split by accomplishing the most reduction with highest *GainRatio* with

Table 2. Information gain and gain ratio comparison analysis with different size of datasets

Attributes	Attributes with different number of records					
	65			90		
	Gain	SplitInfo	GainRatio	Gain	SplitInfo	GainRatio
Resources	1.219	2.78	0.438	1.089	2.754	0.395
Frequency of Visit	1.219	2.78	0.438	0.013	0.852	0.015
Processing Dimension	1.219	2.78	0.438	0.804	2.034	0.395
Perception Dimension	0.658	2.251	0.292	0.453	2.246	0.202
Input Dimension	0.449	1.632	0.275	0.551	1.696	0.325
Resource Cost	0.556	1.685	0.33	0.377	1.765	0.214
Resource Quality	0.161	1.714	0.094	0.256	1.725	0.148
Specification	0.354	1.564	0.226	0.413	1.618	0.255
Mood	2.303	1.83	1.258	2.043	1.783	1.146

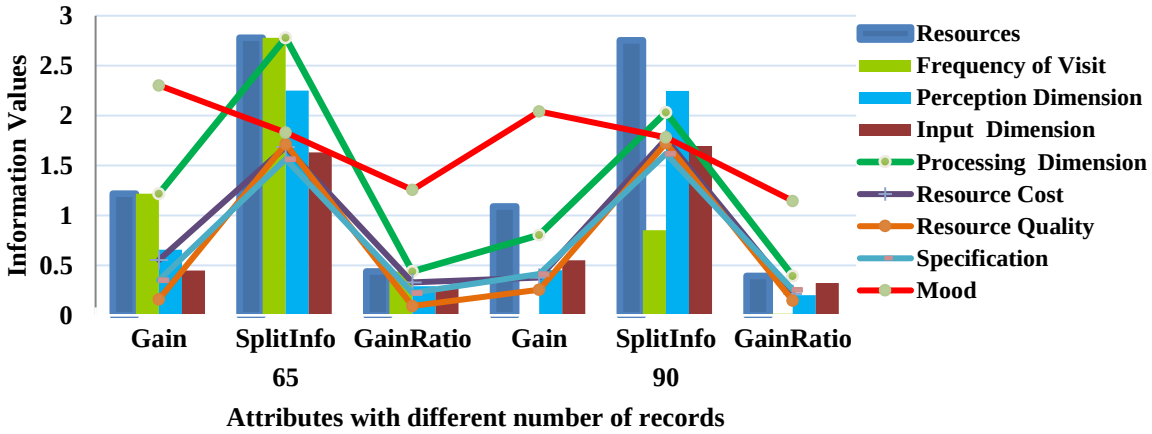

Figure. 3 GainRatio comparison analyses with different attributes and size of data sets

respective to resource feedback. Fig. 3 shows the comparison chart with respect to resource types, user's input, perception, and processing dimensions, learning mood and resource feedback attributes. The highest value of *GainRatio* is preferred as a decision node.

Performance measures:

A confusion matrix shown in Table 3 which is frequently used to portray the execution of a characterization demonstrates on an arrangement of analysis data in which the accurate data are recognized. The correctly predicted user behavioural ontology data observation is referred as True Positives *(TP)* and True Negatives *(TN)* and incorrectly predicted observation is denoted as False Positives *(FP)* and False Negatives *(FN)*. Utilizing this consideration, we estimate Precision, Recall, Accuracy, F1-score, Sensitivity, and Specificity. The Specificity identifies with the classifier's capacity to recognize negative outcomes.

We use 10-fold cross-validation for estimating classification efficiency. We divide our input dataset into both training set and testing set. This implies

Table 3. Confusion matrix

Actual Condition		Predicted Outcome	
		Positive	**Negative**
	Positive	**True Positive (TP)**	**False Negative (FN)**
	Negative	**False Positive (FP)**	**True Negative (TN)**
		$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FN})}$	$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{(\boldsymbol{FP}+\boldsymbol{TN})}$

the dataset to isolate into 10 sections. 1-fold used as a testing set and remaining 9- folds are utilized as the training set. The proposed system's performance is measured with the precision, recall, accuracy, sensitivity, specificity, F1-score and execution time.

These metrics are estimated and compared with the existing algorithm as J48 [11-12], Random Forest [13], Naive Bayes [14], and Multilayer Perceptron [17] algorithms. The experiment output of sensitivity, specificity and F1-score performance comparison is shown in Table 4. Improved J48 induction learning algorithm provides the higher performance values compared with the existing algorithms.

Precision- Precision is the proportion of accurately identified true positive values to the aggregate true positive with the false positives. Precision is the level of responsive records in search results rather than non-responsive reports.

$$Precision = \frac{TP}{(TP+FP)} \quad (14)$$

Recall - Recall also called Sensitivity and it is the proportion of accurately identified true positive values to the aggregate true positives with the false negatives. A Recall is the level of aggregate responsive reports that appear in the search result.

$$Recall = \frac{TP}{(TP+FN)} \quad (15)$$

Table 4. Performance comparison of existing and proposed algorithms

Performance	Datasets	Algorithms				
		Random Forest	Naive Bayes	Multilayer Perceptron	J48	Improved J48
Sensitivity	65	0.277	0.277	0.277	0.385	0.578
	90	0.322	0.333	0.378	0.467	0.667
	195	0.385	0.436	0.441	0.379	0.723
	225	0.422	0.44	0.462	0.422	0.649
Specificity	65	0.819	0.819	0.819	0.846	0.895
	90	0.831	0.833	0.844	0.867	0.917
	195	0.846	0.859	0.86	0.845	0.931
	225	0.856	0.86	0.866	0.856	0.912
F1-Score	65	0.28	0.28	0.308	0.397	0.571
	90	0.227	0.282	0.368	0.398	0.59
	195	0.262	0.362	0.348	0.289	0.669
	225	0.306	0.369	0.372	0.368	0.54

Table 5. Proposed system's precision and recall values.

Different Size of Datasets	Precision	Recall
65	0.59	0.577
90	0.591	0.598
195	0.716	0.646
225	0.598	0.527

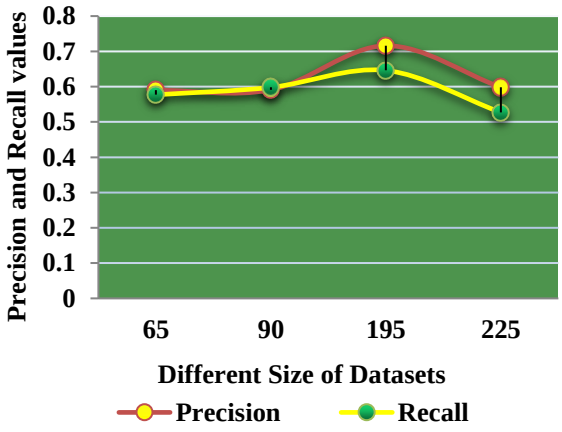

Figure. 4 Proposed system's precision and recall curve

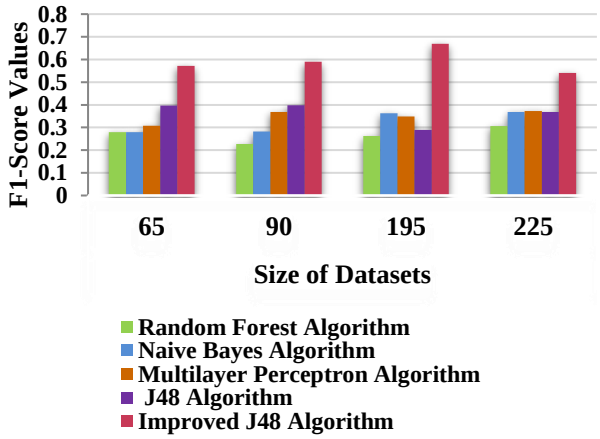

Figure. 5 F1-score comparison chart for different size of datasets

The experiment result of Precision and Recall are computed using Eqs. (14) & (15). Table 5 shows the analyzed result of precision and recall for the proposed system and its graphical representation is shown in Fig. 4.

F1-Score: F1-score used for imbalanced data analysis in the dataset. It is measured using the Eq. (16) which calculates the accuracy of average weighted *precision* and *recall* result. The estimation of F1-score ranges from 0 and 1. The highest value of F1-score is an excellent classification measure. Our proposed system's F1-score experiment output is graphically represented in Fig. 5.

$$F1 - score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (16)$$

Accuracy-Accuracy is the proportion of accurately anticipated perception to the aggregate perceptions. The experiment result is computed using Eq. (17) and Table 6 shows the accuracy comparison with existing [18] and proposed

Table 6. Accuracy comparison

Dataset Size	Random Forest Algorithm	Naive Bayes Algorithm	Multilayer Perceptron Algorithm	J48 Algorithm	Improved J48 Algorithm
65	71.1	71.1	71.1	75.4	83.1
90	72.9	73.3	75.1	78.7	86.7
195	75.4	77.4	77.6	75.2	88.9
225	76.9	77.6	78.5	76.9	86

Table 7. Execution time comparison between C4.5 and Improved J48 algorithms

Size of dataset	C4.5 Algorithm(ms)	Improved J48 Algorithm(ms)
65	40.8	28
90	62.3	42.4
195	137	89.6
225	156.4	106.7

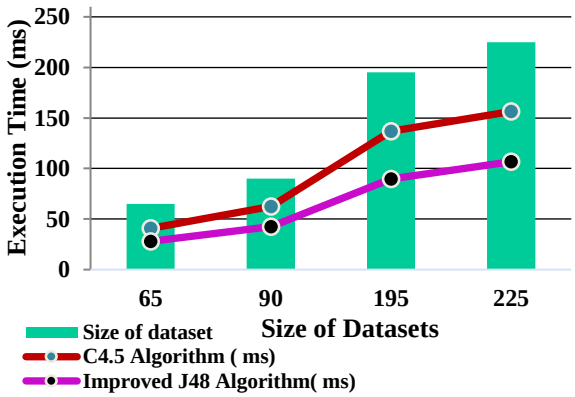

Figure. 6 Execution time comparison chart

algorithm. Concretely, the proposed system can improve its accuracy more than 10 % with the existing Random Forest, Naive Bayes, Multilayer Perceptron and J48 algorithms.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100 \qquad (17)$$

Table 7 shows the execution time represented in milliseconds for our proposed Improved J48 algorithm, which is evaluated with different size of data sets and compared with existing C4.5 algorithm [19]. The proposed algorithm can reduce its execution time more than 20 (ms) for the different size of datasets. The graphical representation of execution time is shown in Fig. 6.

The ontology created using this proposed method aids an enhanced ontology with the evaluation metrics. The parameters such as accuracy, F1-score, and execution time are the focal assessment metrics to evaluate ontology data. The comparisons of existing Random Forest, Naive Bayes, Multilayer Perceptron and J48 algorithms, our proposed Improved J48 algorithm provides higher accuracy for the data classification with less execution time. These predictive techniques give approaches to anticipate a new student related to the same domain will perform the same search with lesser time.

5. Conclusion and future enhancement

To optimize students' learning library resources, we proposed personalized library ontology which discovers vital resources accessible for students to gain knowledge on time. Here we used protégé editor 4.3 to develop library ontology for students' learning feedback which recognizes students' behavior. To analyze student's behaviour we used 225 students library usage feedback with their learning style as processing dimension, perception dimension, input dimension and learners' interest as resource feedback, resource cost satisfaction, resource quality, resource specifications and finally this feedbacks are analyzed with respect to the learner's learning mood.

We proposed Semantic similarity-based Improved J48 induction learning algorithm to classify student's feedback which aids the system to discover pattern and regularities from the mined ontology data. The impurity from the classified data is evaluated using entropy measurement, information gain and GainRatio. The proposed method's performance is assessed with 10-fold cross-validation, F1-score for imbalanced data, accuracy for balanced data, and execution time, which is also compared with existing Random Forest, Naive Bayes, Multilayer Perceptron and J48 algorithms. Concretely, the proposed system can improve its accuracy more than 10 % and reduces its execution time more than 20(ms) for the different size of datasets. The proposed system showed the improved system's efficiency and its performance which reduces dissimilar data, evaluates imbalanced and balanced data and also handles the missing data. The proposed method is used to increase the accessible resources and afford enhanced learning services for students' requisites and expectations. In future, we can enhance our proposed work with a new algorithm to accomplish further reliable outcomes.

References

[1] N.F. Noy and M. Klein, "Ontology Evolution: Not the Same as Schema Evolution", *Knowledge*

and Information Systems, Vol. 6, No. 4, pp.428-440, 2004.

[2] S. Decker, S. Melnik, F.V. Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Hoorocks, "The Semantic Web: The Roles of XML and RDF", *IEEE Internet Computing*, Vol. 4, No.5, pp.63-73, 2000.

[3] B. Kapoor and S. Sharma, "A Comparative Study Ontology Building Tools for Semantic Web Applications", *International Journal of Web & Semantic Technology*, Vol.1, No.3, pp.1-13, 2010.

[4] B.K. Baradwaj and S. Pal, "Mining Educational data to Analyze student's performance", *International Journal of Computer Science and Applications*, Vol.2, No.6, pp.63-69, 2011.

[5] F. Ahmad, N.H. Ismail, and A.A. Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques", *Applied Mathematical Sciences*, Vol.9, No.129, pp.6415 -6426, 2015.

[6] T.M. Mitchell, "Machine Learning", *McGraw-Hill Publisher*, ISBN: 0070428077, 1997.

[7] M.R. Tolun, H. Sever, M. Uludag, and S.M. Abu-Soud, "ILA-2: An Inductive Learning Algorithm for Knowledge Discovery", *Cybernetics and Systems*, Vol.30, No.7, pp.609-628, 1999.

[8] A.H. Mohamed and M.H.S.B. Jahabar, "Implementation and Comparison of Inductive Learning Algorithms on Timetabling", *International Journal of Information Technology*, Vol.12, No.7, pp.97-113, 2006.

[9] J. Ross Quinlan, "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers*, ISBN 978-0080500584, 1993.

[10] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14, No.1, pp.1-37, 2007.

[11] V.K. Nijhawan, M. Madan, and M. Dave, "The Analytical Comparison of ID3 and C4.5 using WEKA", *International Journal of Computer Applications*, Vol.167, No.11, pp.1-4, 2017.

[12] B. Evangelia and G. John, "A Predictive System for Distance Learning Based on Ontologies and Data Mining", In: *Proc. of 4th International Conf. on Advanced Cognitive Technologies and Applications*, pp.152-158, July 2012.

[13] J.N. Undavia, A. Patel, and P. Dolia, "Comparison of Classification Algorithms to Predict Student's Post Graduation Course in Weka Environment", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.3, No.9, pp. 1250-1253, 2013.

[14] S. Wongpun and A. Srivihok, "Comparison of attribute selection techniques and algorithms in classifying bad behaviors of vocational education students", In: *Proc. of 2nd IEEE Conf. on Digital Ecosystems and Technologies*, pp. 526-531, Feb 2008.

[15] J.J. Jiang and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", In: *Proc. of International Conf. Research on Computational Linguistics*, pp.19-33, 1997.

[16] R.M. Othman, S. Deris, and R.M. Illias, "A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences", *Journal of Biomedical Informatics*, Vol. 41, No.1, pp. 65-81, 2008.

[17] S.K. Depren, Ö.E. Askin and E. Öz, "Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS", *Educational Sciences: Theory & Practice*, Vol.17, No.5, pp.1605–1623, 2017.

[18] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", *International Journal of Applied Engineering Research*, Vol.7, No.11, pp.1-5, 2012.

[19] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, "Predicting Students' Performance using ID3 and C4.5 Classification Algorithms", *International Journal of Data Mining & Knowledge Management Process*, Vol.3, No.5, pp.39-52, 2013.