Neural Network and Optimization Based K-Means Clustering for Mining Search Results from Web Database

V. Sabitha
Research Scholar
Sathyabama University
sabitha0583@gmail.com

Dr.S.K.Srivatsa
Senior Professor (ICE), St.Joseph's college of Engg, Chennai

Abstract: Recently, the usage of the web to carry services has developed more and more striking as a novel paradigm for building software. Semantically interpreting web services is ahead more consideration as a significant factor to support the computerized matchmaking and composition of web services. In our proposed method, the web annotation is created from the web result page for an input query. The method extracts the search result folders and calculates feature resemblance with the input query. At this time, the feature weighting is done, that allocates a weight dynamically by means of PSO. Besides, the similar information was grouped with adapted k-means and the relevant group is separated for further process. Then, certain relevant scores are also calculated for the relevant group of search records on the basis of the identical information units, and are also ranked using neural network. Lastly, the best result is enfolded as an annotator for the search query.

I. Introduction

In the database community, the word "Data Mining" seemed around 1990. Data mining is the abstraction of information from the large quantity of observational data sets, to determine unsuspected relationship and design unseen in data, précis the information in new customs to type it comprehensible and valuable to the data consumers. Web usage mining is the implementation of data mining method to mechanically discover and abstract valuable data from a specific web site [9], [10].

In Egypt e-government presently offers 85 services to citizens main reasons are behind hand governments choice to transfer online by data mining method [1]. Adaptive learning schemes depended on data mining method current student modeling methods for computer-assisted language learning [2]. A hybrid model depended on neuro – fuzzy clustering is applied to professionally cluster the users of polytechnic website depended on comparable browsing designs [3]. Bagging and improving technique forecast the actual data mining method for provoking the tie strength among users with their friends on this social network [4].

The main issue of many on-line web sites is the performance of numerous choices to the client at a time; this frequently consequences to strenuous and time overwhelming mission in ruling the right product or data on the site.

In this work, we establish in what way the popular k-means cluster algorithm can be joint with Neural Networks to attain enhancements in efficacy and also optimization. Cluster analysis is one among the main data mining approaches for knowledge discovery in great data bases. It is the procedure of grouping large data sets rendering to their resemblance. One among the simple clustering algorithms k-means is available over 50 years ago. Despite the element that thousands of clustering algorithms have been available since then, k-means is still extensively utilized.

K-means method

The K-means clustering algorithm is utilized to group objects on the basis of attributes/features into k (positive integer) number of groups by diminishing the sum of squares of distances between information and the consistent cluster centroid [5].

In data mining, k-means clustering that is popular for the clustering investigation is a vector quantization technique utilized in signal processing [6]. This algorithm frequently utilizes the Euclidean distance. K-means haphazardly selects k explanations from the dataset and usages them as the initial centers. It analyzes the novel centers to be the centroids of the explanations in the novel clusters. The algorithms have joined when assignments no longer change [7], [8].

ANNs in clustering analysis

Neural Network is also recognized as Artificial Neural Network. A data-processing standard that is enthused with the help of the biological nervous schemes is an Artificial Neural Network (ANN), like the brain. A neural network is also mentioned as Neuron computers, connectionist networks, parallel dispersed processors etc. Artificial neural network is a knowledge treating paradigm that stimulated from the biological nervous scheme.

Artificial Neural Networks (ANNs) comprises of a gathering of simple nonlinear calculating elements whose inputs and outputs are connected to form the network. The learning algorithms of ANNs can be alienated either: administered and unsupervised [9] [10]. ANN is a technique in those searching solutions can be completed optimized by training and testing specified by the neural network.

Advantages of Neural Network

1) Adaptive Learning 2) Real Time Operation 3) Inexpensive 4) Self-Organization 5) Fault tolerance via multiple data copies.

2. Literature Review

A Rapid Miner Linked Open Data extension method was proposed Petar Ristoski et al [11]. In that addition hooks into the influential data mining and investigation platform Rapid Miner, and compromises operators for retrieving Linked Open Data in Rapid Miner, permitting by means of it in urbane data analysis workflows without the requirement for expert knowledge in SPARQL or RDF. The postponement permissible for separately traveling the Web of Data by subsequent links, thereby decisive pertinent datasets on the fly, and also for assimilating overlying information found in dissimilar datasets.

A Data Mining and Knowledge Discovery in Databases were presented by Heiko Paulheim et al [12]. It utilized in research field disturbed with descending higher level insights from information. The errands achieved in that field were information concentrated and can often advantage by means of additional information from numerous sources. Consequently, numerous methods had been projected to syndicate Semantic Web data with the data mining and knowledge discovery procedure.

A large information method it categorizes huge amounts of dissimilar news articles into numerous categories (topic areas) on the basis of the text content of the articles these groups were continually updated with novel articles was presented by José Antonio Iglesias et al [13]. News websites yield thousands of articles casing an extensive spectrum of topics or groups which can be measured as a big data issue. For abstracting useful data, these news articles need to be administered by big data methods.

A Text Flows web-based text mining and natural language processing platform secondary workflow construction, sharing and implementation was presented by Janez Kranjc et al [14]. The platform permits visual construction of text mining workflows via a web browser and the implementation of the constructed workflows on a dispensation cloud. Text Flows is an adjustable organization for the structure and sharing of text processing workflows that can be reclaimed in numerous presentations.

The Data Mining Optimization Ontology (DMOP) it supports knowledgeable decision-making at numerous choice points of the data mining procedure was elucidated by C. Maria Keeta et al [15]. The ontology can be castoff by data miners and organized in ontology-driven data schemes. The main drive of DMOP had been industrialized in the automation of algorithm and model selection finished semantic meta-mining that creates usage of an ontology-based meta-analysis of comprehensive data mining procedures in the aspect of abstracting designs related with mining performance.

The temporal semantic annotations were presented by Zheng Xu et al [16]. It helps operators to learn and comprehend the unfamiliar or novel developed semantic relations among entities. The temporal semantic annotation construction assimilates the features from IEEE and Renlifang. A produced temporal semantic allow annotation of a semantic relation between entities by building its lexical-syntactic patterns, connection entities, context sentences, context communities and context graph.

An algorithm for mining weighted maximal recurrent item sets from incremental databases as proposed by Unil Yun et al [17]. It scans an assumed incremental database only once, it can not only behavior its mining operations appropriate for the incremental atmosphere but also abstract a lesser number of significant item sets associated with preceding methods this method also had a consequence on expert and intelligent schemes.

An enhanced tabu search (ETS) algorithm with progressive search features of multiple areas, adaptive tabu lists, dynamic tabu tenure, and multi-level aspiration standards was given by Peng Yeng Yin et al [18]. The website structure optimization (WSO) issue had abridged and it mechanically regroups the structure or content by learning from the consumers browsing behaviors, as such the use of the websites were enhanced.

3. Proposed Methodology

In our proposed technique, the annotation is enfolded for the web search solutions. So that, the method is investigated for an input query and thereby gathering the connected relevant search result records. Lastly, the wrapping of the finest relevant object with the search solution is achieved. The proposed method detects and wraps the annotation content by ranking the best label over the Neural Network method. The plan of the proposed methodology is assumed in the subsequent segment.

3.1. Framework of Proposed Methodology

The proposed technique to produce the annotation wrapper can be characterized by the subsequent seven steps namely,

- Data Extraction (Search Result Records) phase
- Feature Extraction phase
  - Dynamic Feature Weighting using PSO
- Clustering phase (Modified k-means)
- Score Calculation phase
- Annotation phase (Neural Network)
  - Ranking Labels
- Annotation wrapper generation phase

In the primary phase, the input query is assumed. For an assumed query, the search records were congregated from the web database. The information abstraction phase comprises the extraction of the elements that were gotten at the time of the search solutions. Also, the feature extraction stage comprises the resemblance computation for the records in the name of data similarity, content resemblance, presentation style resemblance, data type similarity, tag path resemblance and adjacency resemblance. Subsequent comes the clustering phase, where the comparable information was clustered by using the input features.

The architecture diagram to display the comprehensive view of the proposed method presenting the generation of annotation wrapper is specified as in the above-mentioned figure 1.

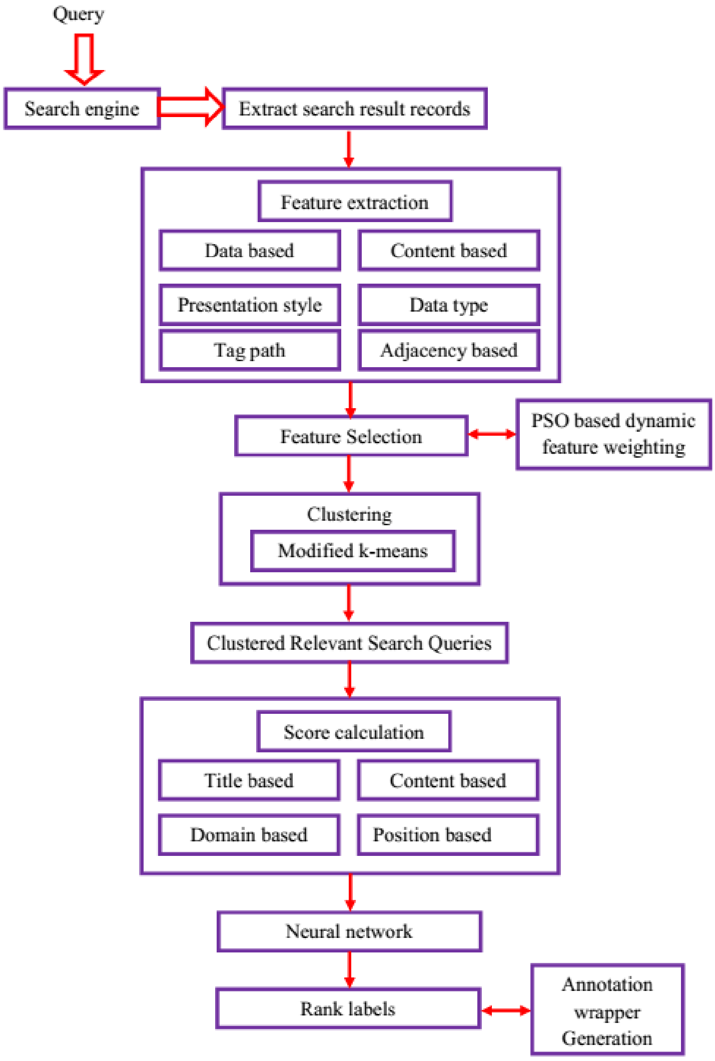


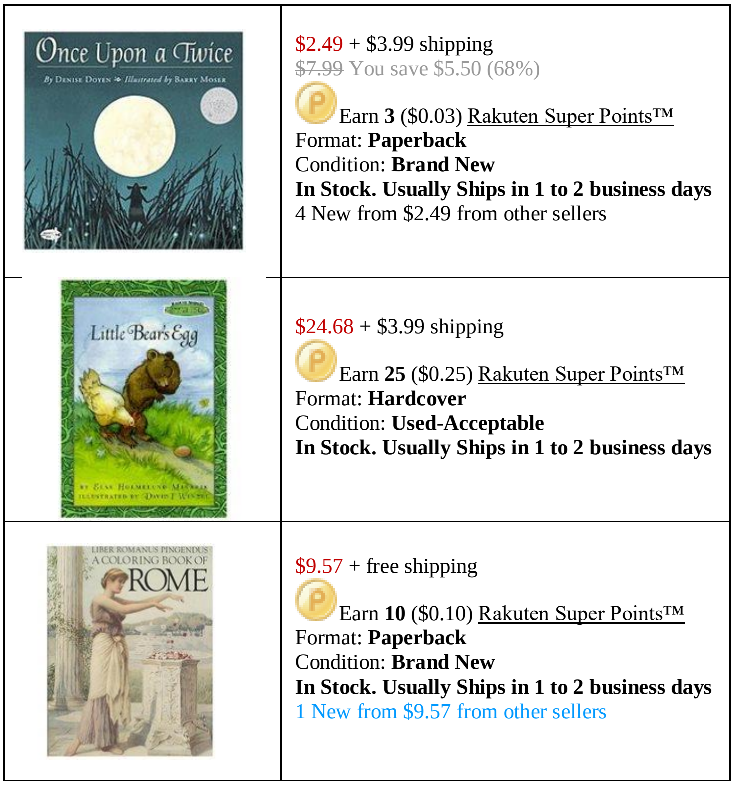
Figure 1: Proposed Architecture for Annotation Wrapping

For the abstracted records, the most suitable record is assessed by grouping the alike elements. At the time, the some weights will be allocated dynamically based on the individual importance of each feature attribute. In our method, the weights were produced dynamically through the PSO algorithm on the basis of the clustering accuracy. At this time, the clustering is achieved by Modified k-means algorithm.

After grouping by modified K-means method, the relevant and irrelevant groups are identified. Once the relevant cluster is found, the corresponding search queries are alone picked and certain important scores are calculated. The similar elements were computed through the similarity computation process taking place between the query and the web search results. And then, based on the similarity scores, the ranking of the labels takes place. Lastly, the best labels were designated for wrapping as annotation by the Neural Network and are enfolded as annotation in the Annotation Wrapper generation stage.

II. Data (Search Result Records) Extraction phase

At this phase, the data extraction flinches with inputting a search query. Frequently, the web search solutions in the extraction of search result record from the web database. At this time, a few among the search archives regained at the time of the search of query “book” specified as follows.



$2.49 + $3.99 shipping
~~$7.99~~ You save $5.50 (68%)
Earn **3** ($0.03) Rakuten Super Points™
Format: **Paperback**
Condition: **Brand New**
In Stock. Usually Ships in 1 to 2 business days
4 New from $2.49 from other sellers

$24.68 + $3.99 shipping
Earn **25** ($0.25) Rakuten Super Points™
Format: **Hardcover**
Condition: **Used-Acceptable**
In Stock. Usually Ships in 1 to 2 business days

$9.57 + free shipping
Earn **10** ($0.10) Rakuten Super Points™
Format: **Paperback**
Condition: **Brand New**
In Stock. Usually Ships in 1 to 2 business days
1 New from $9.57 from other sellers

III. Feature Extraction phase

The feature extraction process can be performed to regulate the characteristic property of an object. At this time, in our technique the convinced features such as data content, presentation style, information type, tag path and adjacency were abstracted. Additionally, the feature similarity is calculated between the neighboring information units subsequent the feature abstraction of the search solution records.

For the data units x_1 and x_2, the data content similarity $sC(x_1,x_2)$, information type similarity $sD(x_1,x_2)$, presentation style similarity $sP(x_1,x_2)$, tag path similarity $sT(x_1,x_2)$ and adjacency similarity $sA(x_1,x_2)$ can be found using the following representations [19]:

$sC(x_1,x_2)=FV_{x_1}\bullet FV_{x_2}/\|FV_{x_1}\|*\|FV_{x_2}\|$ .Where, FV_x represents the frequency vector of data units of x
.

$sD(x_1,x_2)=LCS(e_1,e_2)/MAX(m(e_1),m(e_2))$ .Where, $LCS(e_1,e_2)$ is the longest common sequence of the sequence e_1 and e_2 of the data types of x_1 and x_2 respectively. And, $m(e)$ is the number of element types in the data unit.

$sP(x_1, x_2) = \sum_{i=1}^{6} MS_i / 6$. Here, MS_i is the score of the i^{th} style feature.

$sT(x_1, x_2) = 1 - EDT(t_1, t_2)/(m(TP_1) + m(TP_2))$. Where, $m(TP)$ is the number of tag paths in the data unit.

$sA(x_1, x_2) = (s'(x_1^{\,p}, x_2^{\,p}) + s'(x_1^{\,s}, x_2^{\,s}))/2$. Where, x^p and x^s are the preceding and succeeding data units of x.

Combining, all the features, the similar datas with more importance is grouped. Here, the proposed approach selects the data based on the importance needed to be given in the time of clustering. In [19], the feature weighting between two data units is calculated by,

$$s(x_1, x_2) = w_1 * sC(x_1, x_2) + w_2 * sP(x_1, x_2) + w_3 * sD(x_1, x_2) \\ + w_4 * sT(x_1, x_2) + w_5 * sA(x_1, x_2)$$

But, the manual weight allocation does not work for all type of datas. So, dynamic weight selection based on the data type is highly required. Therefore, we have selected the importance of each features based on the dynamic weights allocation using PSO algorithm.

- *Dynamic Weight Allocation for features*

The similarity computation comprises computing convinced functions such as the data content similarity, performance style resemblance, information type similarity, tag path similarity and adjacency similarity. Once, the features are extracted, the weights are allocated initially to all the feature set. Then, for finding the importance of each features, the initial value allocated for the weight has to be attuned to produce maximum cluster accuracy throughout the clustering phase. In our system, the weight is produced dynamically by engaging the PSO algorithm.

The significance is calculated for five features by tuning the weights of each feature dynamically using PSO. The weight values are produced and are updated with the fresher weights during every iteration. Throughout update, the clustering outcome is checkered at each phase, so as to accomplish a maximum accuracy of clustered outcomes.

- *Particle Swarm Optimization*

Particle Swarm Optimization algorithm is an evolutionary computation algorithm interested with the help of the social behavior of bird flocking and fish schooling. PSO is a populace supported optimization algorithm, wherever, the population is primed as a particle. Formerly, the best position of the particle is unwavering and updated. The local best and the universal best positions were strong-minded for all the particles.

The local best is strong-minded with the help of the particle by the incessant following of the particle. Likewise, the worldwide best is gained from the best position that has been determined as best by any particle. Originally, the local best and the worldwide best positions were primed arbitrarily and on the basis of the fitness values, the results will be substituted by the newer result produced from its community positions.

The phases complicated in the PSO algorithm is specified as follows,

Step 1: Initialize parameters

Primarily, the particle and its corresponding velocity are primed within the search space arbitrarily. At this time, the initial result is produced from the set of weights allocated at the time of the feature weighting phase.

Let us seeing the particle at k^{th} position be, $R_k = [r_{k1}, r_{k2}, ..., r_{kD}]$. As, each particle transfers with a velocity assumed as, $V_k = [v_{k1}, v_{k2}, ..., v_{kD}]$. For every particle, the particle best and also the worldwide best position are assessed.

Step 2: Initialize particle best ($sbest$)

The $sbest$ is the best position established with the help of the particle. The $sbest$ for the particle at k^{th} position is arbitrarily produced that is given through the subsequent representation.

$$S_k = [s_{k1}, s_{k2}, ..., s_{kD}] \quad (1)$$

Step 3: Initialize global best ($hbest$)

The worldwide best position is also primed haphazardly. The worldwide best ($hbest$) is measured to be the best position among all the particles. The $hbest$ is indicated as follows:

$$S_h = [s_{h1}, s_{h2}, \dots, s_{hD}] \tag{2}$$

Step 4: Fitness Evaluation

Formerly, the fitness of each result is intended and the result related to the fittest value is substituted. At this time, the fitness assessment is made over the subsequent equation as,

$$FF = \max\ (clustering\, accuracy) \tag{3}$$

Where, the clustering accuracy is calculated from the data clusters produced after clustering into groups arbitrarily for the input query based on the weighted features. Furthermore, the fitness function, $\max\ (clustering\, accuracy)$ denotes the maximum accuracy of relevant correlated clusters.

Here, the fitness is assessed for the arbitrarily prepared result (i.e. the clustering accuracy is found for cluster output, for which the clustering is made based on the feature attributes with randomly allocated weights). If the current $sbest$ is establish to be fittest than the initial result, then the velocity of the initial result is updated with the $sbest$.

The flowchart of the PSO algorithm is specified as above in figure 2.

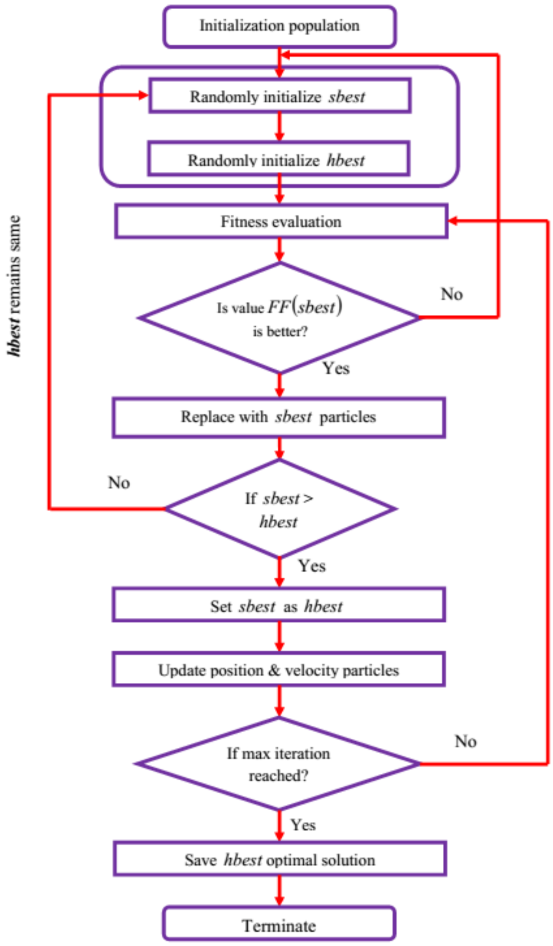

Initialization population
Randomly initialize sbest
Randomly initialize hbest
Fitness evaluation
Is value FF(sbest) is better?
No
Yes
Replace with sbest particles
No
If sbest > hbest
Yes
Set sbest as hbest
Update position & velocity particles
If max iteration reached?
No
Yes
Save hbest optimal solution
Terminate
hbest remains same

Figure 2: Flowchart of PSO

Step 5: Velocity Updation

The velocity updation is completed through the subsequent equation:

$$V_{kl}^{i+1} = wv_{kl}^{i} + f_1 v_{1k}^{i}(s_{kl}^{i} - r_{kl}^{i}) + f_2 v_{2k}^{i}(s_{gl}^{i} - r_{gl}^{i}) \tag{4}$$

Where,

$k = 1,2,...,n;\ l = 1,2,...,D;$

Step 6: Particle Position updation

With the velocity strong-minded, the current position of the particle is updated with the help of the best position.

$$R_{kl}^{i+1} = r_{kl}^{i} + V_{kl}^{i+1} \tag{5}$$

Step 7: Swarm updation

Another time, the fitness is assessed for the current best position and also the neighborhood result. As per the better fitness values, the $sbest$ is updated. At every stage of $sbest$ updation, the $sbest$ is associated with the $hbest$ (global best). When the $sbest$ is better than the current $hbest$, the $sbest$ will be substituted at the place of $hbest$.

Step 8: Termination

The calculation and updation of $hbest$ takes place until the dissolution criteria (or) the maximum generation has touched.

Finally, the optimized weights are produced dynamically with the PSO method by checking with the clustering accuracy. The dynamic weights calculated will be utilized for clustering the similar type of datas.

IV. Clustering phase (Modified k-means)

K-means algorithm is simple, easy to use, and efficiency to execute. Therefore, it is suitable to process a huge amount of high dimensional and continuous numerical data. But, the data features do not precisely represent all the data in this cluster, when the data in one cluster are quite different. However, due to the drawbacks of existing k-means method, we have utilized the Modified k-means clustering algorithm. Here, the clustering is achieved with the dynamically weighted features found based on the accuracy of clustering procedure. In this step, the applicable information was gathered into clusters of information units (i.e. relevant and irrelevant). The stages taking place in the clustering algorithm is specified as follows,

Step 1: Accomplish information objects haphazardly whereas each object expresses primary cluster center.

Step 2: Regulate distance within each unit in one set and each clustering center, the unit is situated into the subset of the nearest clustering center.

Step 3: Compute the average of units in each clustering subset as its novel clustering center.

Step 4: Recurrence steps (2 & 3) until the objective performances converges.

At last, the relevant groups are alone selected after clustering the data. Then, the scores were considered for the comparable data units inorder to rank them using ANN.

V. Score Calculation phase

Following to clustering, the scores were calculated to the pertinent data units. The pertinent data were ranked on the basis of the score calculation in the name of content based, domain based, position based and title constructed scheming for the concepts. The content based, domain based, position based and title based score calculations are given in the following equations.

Here, the title based value of k^{th} unique link $T_k(q)$ is given as,

$$T_k(q) = \sum_{i=1}^{n} \left\{ \left(\frac{(T_k\_x_i) - \max(T\_x_i) + 1}{\max(T\_x_i)} \times w_Q \right) + \left(\sum_{j=1}^{b} \frac{(T_k\_x_i)N_j - \max(T\_x_i N_j) + 1}{\max(T\_x_i N_j)} \times w_N \right) \right\} \tag{6}$$

In the above equation, $(T_k\_x_i)$ is the number of occurrences of i^{th} query word in the title T of k^{th} link, $\max(T\_x_i)$ is the maximum number of occurrence of i^{th} query word in the title of whole unique links, $(T_k\_x_i)N_j$ is the number of occurrence of j^{th} meaning of i^{th} query word in the title T of k^{th} link, $\max(TEdu_i N_j)$ is maximum number of occurrence of j^{th} meaning of i^{th} query word in the title of whole unique links, 'n' is the total number of query word, and b represents the total number of meaning of i^{th} query word, w_Q the weight value of the query word and w_N is the weight value of the meaning word of the query word.

Also, the content based score calculation $C_k(q)$ is given as,

$$C_k(q)=\sum_{i=1}^{n}\left(\frac{C_k\_x_i}{\max(C\_x_i)}\times w_Q+\sum_{j=1}^{b}\frac{C_k\_x_iN_j}{\max(C\_x_iN_j)}\times w_N\right) \tag{7}$$

In the above equation,; and $C_k\_x_i$ is the number of occurrence of i^{th} query word x_i in the k^{th} unique link, $\max(C\_x_i)$ is the maximum number of occurrence of i^{th} query word x_i in, $C_k\_x_iN_j$, the number of occurrence of j^{th} synonym of i^{th} query word x_i in the content of k^{th} unique link; and $\max(C\_x_iN_j)$ is the maximum number of occurrence of j^{th} synonym of i^{th} query word x_i in the content of k^{th} unique link.

Moreover, the domain based score value of k^{th} domain score value $D_k(q)$ is given as,

$$D_k(q)=\log_{10}\left(\frac{2y-1+acc_s}{2y}\right) \tag{8}$$

In the above equation, acc_s represents the number of unique links with same domain name and y represents the number of search engines used.

Moreover, the position based score is given as,

$$P_k(q)=\frac{y*r-\left(\sum_{l=1}^{y}P(q)\right)}{y*r} \tag{9}$$

In the above equation, r is number of links we have taken for our process from each search engine; and $P(q)$, the rank of a link in a particular search engine.

The score scheming of the analogous information units was then used to interpret the best score with the web result page. The best score is strong-minded from all the four approaches taken to compute the score. Here, the ranking of the score values is accomplished through the Neural Network method.

VI. Annotation phase (Neural Network)

For the given query, the pertinent related information was recovered and was allocated with suitable score values. Then, the best label is originated by ranking the score values allocated for the retrieved data. The topmost ranked data is evaluated by means of the Neural Network process. Artificial neural Network (ANN) is an artificial intelligence technique that copies the human mind's operation and calculation performance, and can mirror the basic linear or non-linear connections among input and target information. Obtaining learning from an outside source, ANN stores data in its inner processing units and conveys them by methods for the communication of the transfer functions and the association parameters between nearby layers.

The neural networks originally, train itself with the input information. Then, the training process of the information for the ranking of labels was used further at the time of forecast of best labels for the input query information. Here, the back propagation (BP) based training algorithm is used. The steps involved in the BP algorithm are given in the following section.

Back propagation learning algorithm steps

i. The input and hidden layer, hidden and output layer weights of the neural network are initialized randomly.

ii. Learning the network according to the input and the corresponding target $\{x_1, x_2 \cdots x_n\}$.

iii. The current output of the network is determined by following them,

$$x_n^{(actual)} = \alpha_n + \sum_{m=2}^{M} wt_{nm}\, x_n^{(m)} \tag{4}$$

Where, α_n is the bias function of the n^{th} node

$$x_n^{(m)} = \frac{1}{1+\exp(-wt_{nm}x_n - wt_{om})} \tag{5}$$

iv. Calculate the back propagation error of the target.

$$Er_n^{\Delta} = x_n^{(desired)} - x_n^{(actual)} \tag{6}$$

Where, $x_n^{(desired)}$ is the network target of the n^{th} node and $x_n^{(actual)}$ is the current output of the network.

v. The new weights of the each neurons of the network are updated by, $wt' = wt + \Delta wt$. Here, wt' is the new weight, wt is the previous weight and Δwt is the change of weight of each output. The change of weight is determined by follow,

$$\Delta wt = \delta . x_n . Er_n^{\Delta} \tag{7}$$

Where, δ is the learning rate (0.2 to 0.5).

vi. Repeat the above steps till the Er^{Δ} gets minimized $Er^{\Delta} < 0.1$.

Once the training procedure is completed, the network is trained to rank the data units using the calculated scores based on the trained data.

Annotation wrapper generation phase

Annotators can be of countless forms predominantly, FA, QA, IA and CA, where, FA, QA, IA and CA characterize the frequency annotator, query annotator, In-text prefix/ suffix annotator and common information annotator. At this time, the annotation wrapper is created with the best data unit found from ranking them through ANN. The wrapper cohort can be accomplished by using the annotation wrapper generation rule. The annotation wrapper can equip with the label, prefix, suffix, and unit index information composed of the search solution records. The annotation rule is given by,

$$attribute_i = < label_i, prefix_i, suffix_i, seperators_i, unitindex_i >$$

Furthermore, the annotation is utilized to link the web page to another web page for that the best ranking is created. So, one can make usage of the annotation enfolded together with web search result and thus permitting the user with a more suitable manner to gain knowledge from their obligatory area.

4. Results and Discussions

We assess the proposed technique approach over five domains (music, job, book, game and movie) and gadget it in java platform. Its goal is to display our proposed process performance is better than the available method.

For input query image first data extraction occurs then its feature was abstracted. Particle Swarm Optimization algorithm was utilized in dynamically allocating their importance in clustering. The datas were clustered with the help of Modified k-means algorithm. for the clustered sets (i.e. only relevant group of datas) the scores were considered and then they were ranked. The best labels were nominated for wrapping as annotation by the Neural Network.

The proposed annotation method is compared with the traditional frequency annotator (FA), query annotator (QA), In-text prefix/ suffix annotator (IA) and common information annotator (CA). Moreover, the comparison is also made with [19].

4.1. Performance Measures

The performance of Proposed and prevailing method performances are considered in the name of precision and recall, two performance events utilized to associate different algorithms, namely, precision and recall. They are distinct correspondingly with a specific word was follows

$$\text{Precision (w)} = \frac{\text{Total number of images correctly annotated with w}}{\text{Total numberof images automatically annotated with w}} \qquad (6)$$

$$\text{Recall (w)} = \frac{\text{Total number of images correctly annotated withw}}{\text{Total numberof images manually annotated with w}} \qquad (7)$$

Precision has subsequent terms to the degree they are frequency annotator, query annotator, In-text prefix/ suffix annotator and mutual knowledge annotator. Although recall also has title based calculation, domain based calculation, position based calculation and content based calculation.

TABLE 3: EXISTING FA, QA, IA AND CA METHOD'S LABELING PERFORMANCE

DOMAINS	EXISTING METHOD							
	PRECISION				RECALL			
	FA	QA	IA	CA	FA	QA	IA	CA
BOOK	82.9	88.5	90.0	90.1	55.9	80	82.4	80
JOB	82.5	88.7	89.9	89.9	55.4	80.3	82	80.1
MUSIC	82.8	88.9	90	90	55.6	80.1	82.3	80.1
GAME	83.0	88.6	89.9	90.1	55.7	80.1	82.3	78.3
MOVIE	83.2	89.2	90.3	89.9	56.3	80.2	81.2	78.2
Average	82.8	88.7	90	90	55.7	80.1	82	79.3

TABLE 4: PROPOSED METHOD LABELING PERFORMANCE

DOMAINS	PROPOSED METHOD							
	PRECISION				RECALL			
	TC	DC	PC	CC	TC	DC	PC	CC
BOOK	83.1	88.8	90.3	90.3	55.9	80.2	82.4	86.1
JOB	87.6	88.8	90.2	90.2	55.5	87.4	82.3	80.4
MUSIC	88.9	89	97.1	96.2	55.8	80.2	82.4	87.2
GAME	86.2	96.7	93.2	94.2	58.8	89.3	89.5	78.5
MOVIE	89	89.4	90.2	90	56.5	80.4	89.5	78.5
Average	86.96	90.54	92.2	92.18	56.5	83.5	85.22	82.14

Discussion

The performance of the planned technique and the available method are deliberated in Table 1 and Table2. In table.1, FA, QA, IA and CA characterize the frequency annotator, query annotator, In-text prefix/ suffix annotator and communal knowledge annotator. In table.2, TC (Title calculation score), DC (Domain Calculation score), PC (Position calculation score), and CC (Content calculation score) signify the title based calculation, domain based calculation, and position based calculation and content based calculation. In table.1 the average performance of 4 annotators namely frequency annotator, query annotator, In-text prefix/suffix annotator and communal knowledge annotator is assumed that is associated with the performance of the proposed technique namely title based calculation, domain based calculation, position based calculation and content based calculation.

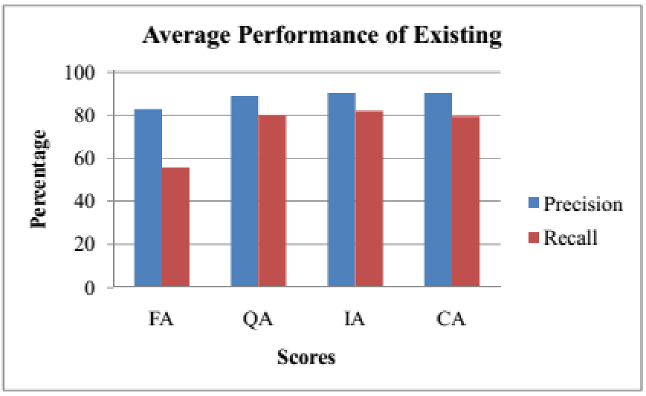


Figure 1: Average performance graph of existing method

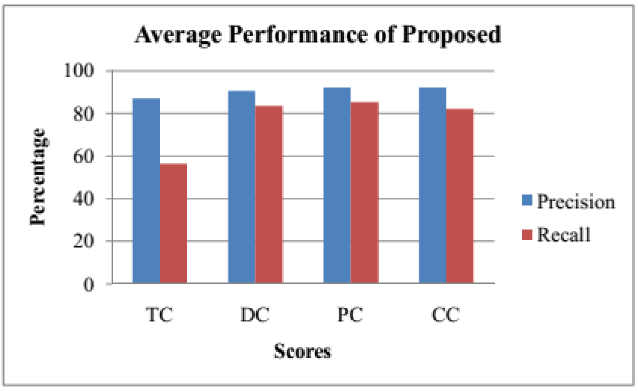


Figure 2: Average performance graph of proposed method

Discussion

From the values specified in the table, two graphs is strategized for the assessment of the performance of the proposed technique and the available technique by taking its precision and recall values. From the graph, it is evident that the exactness and recall of our proposed method are higher than the available method in all the cases. Thus the function of the proposed method performs better than the available method. Once the precision and recall of the proposed technique are higher than the available technique, it is well accomplished of labeling the records in the search engine and thus decreasing the time consumption in examining a specific file.

Moreover, the overall precision, recall of the proposed method is compared with the traditional methods and [19]. The values are tabulated in table 5.

TABLE 5: OVERALL PERFORMANCE OF PROPOSED AND EXISTING METHOD

Metrics	Proposed	Existing	[19]
Precision	90.47	87.875	89.1725
Recall	76.84	74.275	75.5575

From the above table, it is clear that the proposed annotation method, annotates the best results when compared to the existing methods.

CONCLUSION

In this work, the annotation is produced from the web database for the consumer query input. The proposed method is investigated with the web database comprising search result records with the information congregated from numerous websites for the similar query. Besides, in order to display the effectiveness of our proposed technique, the performance of the proposed and also the available approaches were assessed and then associated concurrently. At this time, the available annotation approaches such as frequency annotator, query annotator, In-text prefix/ suffix annotator and communal knowledge annotator is associated with the proposed annotation produced by ANN with numerous score calculation. Additionally, the performance measure like precision and recall is single-minded for both the methods to evaluate the performance of our proposed technique and reference [19]. The result outcome shows 90.47 precision for the proposed approach, which is better than the existing methods.

References

[1] Mohamed M. Mostafa and Ahmed A. El-Masry, "Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques", International Journal of Information Management, vol. 33, no. 4, pp. 627-641, 2013.

[2] Shyi-Ming Chen and Po-Jui Sue, "Constructing concept maps for adaptive learning systems based on data mining techniques", Expert Systems with Applications, vol. 40, no. 7, pp. 2746-2755, 2013.

[3]G. Shivaprasad, N.V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining", Procedia Computer Science, vol. 54, pp. 327-334, 2015.

[4] Mohammad Karim. Sohrabi and Soodeh. Akbari, "A comprehensive study on the effects of using data mining techniques to predict tie strength", Computers in Human Behavior, vol. 60, pp. 534-541, 2016.

[5] Sungjune. Park, Nallan C. Suresh and Bong-Keun. Jeong, "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm", Data & Knowledge Engineering, vol. 65, no. 3, pp. 512-543, 2008.

[6] Süleyman Savaş. Durduran, "Automatic classification of high resolution land cover using a new data weighting procedure: The combination of k-means clustering algorithm and central tendency measures (KMC–CTM)", Applied Soft Computing, vol. 35, pp. 136-150, 2015.

[7] Mostafa. Ghodousi, Ali Asghar. Alesheikh and Bahram. Saeidian, "Analyzing public participant data to evaluate citizen satisfaction and to prioritize their needs via K-means, FCM and ICA", Cities, vol. 55, pp. 70-81, 2016.

[8] Anil K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[9] R. Kuo, J. Liao and C. Tu, "Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce", Decision Support Systems, vol. 40, no. 2, pp. 355-374, 2005.

[10] Shu-Hsien. Liao, Pei-Hui. Chu and Pei-Yuan. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, vol. 39, no. 12, pp. 11303-11311, 2012.

[11] Petar Ristoski, Christian. Bizer and Heiko. Paulheim, "Mining the Web of Linked Data with Rapid Miner", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 35, pp. 142-151, 2015.

[12] Petar Ristoski and Heiko Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 36, pp. 1-22, 2016.

[13] José Antonio. Iglesias, Alexandra. Tiemblo, Agapito. Ledezma and Araceli. Sanchis, "Web news mining in an evolving framework", Information Fusion, vol. 28, pp. 90-98, 2016.

[14] Matic. Perovšek, Janez. Kranjc, Tomaž. Erjavec, Bojan. Cestnik and Nada. Lavrač, "TextFlows: A visual programming platform for text mining and natural language processing", Science of Computer Programming, vol. 121, pp. 128-152, 2016.

[15] C. Maria Keeta, Agnieszka Ławrynowiczb, Claudia d'Amatoc, Alexandros Kalousisd, Phong Nguyene, Raul Palmaf, Robert Stevensg and Melanie Hilarioh, "The Data Mining OPtimization Ontology", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 32, pp. 43-53, 2015.

[16] Zheng. Xu, Xiangfeng. Luo, Shunxiang. Zhang, Xiao. Wei, Lin. Mei and Chuanping. Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines", Future Generation Computer Systems, vol. 37, pp. 468-477, 2014.

[17] Unil. Yun and Gangin. Lee, "Incremental mining of weighted maximal frequent item sets from dynamic databases", Expert Systems with Applications, vol. 54, pp. 304-327, 2016.

[18] Peng-Yeng. Yin and Yi-Ming. Guo, "Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining", Applied Mathematics and Computation, vol. 219, no. 24, pp. 11082-11095, 2013.

[19] Sabitha, V., and S. K. Srivatsa. "Machine Learning Technique Based Annotation in Web Database Search Result Records with Aid of Modified K-Means Clustering (MKMC)." Research Journal of Applied Sciences, Engineering and Technology 10, no. 8 (2015): 853-862.