International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6537
***Article Received**: 27 January 2023 **Revised**: 28 February 2023 **Accepted**: 06 March 2023*

Lancaster Stem Sammon Projective Feature Selection based Stochastic eXtreme Gradient Boost Clustering for Web Page Ranking

Mr. P. Sujai[1], Mrs. V. Sangeetha[2]
[1]Assistant Professor, SVR College of Commerce and Management Studies,
HSR Layout,Bangalore,
Karnataka,India.
spsujai2@gmail.com
[2]Assistant Professor, Government Arts and Science College,
Pappireddypatti, Tamilnadu,India.
sangee759@gmail.com

Abstract— Web content mining retrieves the information from web in more structured forms. The page rank plays an essential part in web content mining process. Whenever user searches for any information on web, the relevant information is shown at top of list through page ranking. Many existing page ranking algorithms were developed and failed to rank the web pages in accurate manner through minimum time feeding. In direction to address the above mentioned issues, Lancaster Stem Sammon Projective Feature Selection based Stochastic eXtreme Gradient Boost Clustering (LSSPFS-SXGBC) Approach is introduced for page ranking based on user query. LSSPFS-SXGBC Approach has three processes for performing efficient web page ranking, namely preprocessing, feature selection and clustering. LSSPFS-SXGBC Approach in account of the numeral of operator request by way of an input. Lancaster Stemming Preprocessed Analysis is carried out in LSSPFS-SXGBC Approach for removing the noisy data from the input query. It eradicates the stem words, stop words and incomplete data for minimizing the time and space consumption. Sammon Projective Feature Selection Process is carried out in LSSPFS-SXGBC Approach to select the relevant features (i.e., keywords) based on user needs for efficient page ranking. Sammon Projection maps the high-dimensional space to lower dimensionality space to preserve the inter-point distance structure. After feature selection, Stochastic eXtreme Gradient Boost Page Rank Clustering process is carried out to cluster the similar keyword web pages based on their rank. Gradient Boost Page Rank Cluster is an ensemble of several weak clusters (i.e., X-means cluster). X-means cluster partitions the web pages into 'x' numeral of clusters where each reflection goes towards the cluster through adjacent mean value. For every weak cluster, selected features are considered as the training samples. Subsequently, all weak clusters are joined to form the strong cluster for attaining the webpage ranking results. By this way, an efficient page ranking is carried out through higher accurateness and minimum time consumption. The practical validation is carried out in LSSPFS-SXGBC Approach on factors such ranking accurateness, false positive rate, ranking time and space complexity with respect to numeral of user query.

Keywords- web content mining, page rank, user query, strong cluster, X-means cluster.

I. INTRODUCTION

Web Mining is the process of using data mining methods for pull out the valuable information from the data on Web. PageRank is used by Google to identify the page importance on the web. Web used all incoming links to page as votes for efficient PageRank. Multi Criteria Indexing and Ranking Model (MCIR) was introduced in [1] depending on weighted documents and pages with one or more ranking features. The designed model attained better performance with the ability to rank the pages and documents. However, the page ranking time was not minimized by MCIR. An innovative approach was designed in [2] depending on semantic technologies. Semantic Search approach leveraged the knowledge graph representation for rating educational web content. But, the ranking accuracy was not reduced by designed approach.

A ranking method was presented in [3] through visual comparisons between the web pages through structure and vision-based structures. The web page visual structure was constructed with similarity through wireframe design. Nonetheless the error proportion was not minimalized through ranking method. Taylor Horse Herd Optimization (THHO)-based Deep Fuzzy Clustering (DFC) was introduced in [4] for web page recommendation model. An stimulating sub charts were obtained since web log record through Weighted-Gaston (W-Gaston) procedures. However, computational cost was not minimized by THHO-DFC.

An incremental C-Rank was introduced in [5] towards apprise the C-Rank values of particular slice of Web pages without any accurateness loss. C-Rank suffered after very high prices towards reproduce the dynamic also regular variations in World Wide Web. An attributed multiplex network was designed in [6] to determine node centrality. PageRank model performed ranking in Web pages network. An Modified PageRank Procedure was employed for monoplex systems through data. But, the space complexity was not condensed through intended algorithm.

A wrapping technique termed Proficiency Rank was introduced in [7] with the elective statistics towards compute the operator expertise for funders and lurkers. Proficiency Rank performed meaningful user ranking to compute the baselines depending on intrinsic and extrinsic information. However, the accuracy level was not improved by designed method. Internet webpage performance enhancement was carried out in [8] through Data Envelopment Analysis (DEA). The enhancement addressed in global ranked through their page load time forecast on dimensions and numeral of items. The decay of competence was carried out from reverse frontier viewpoint. But, the computational cost was not minimized by DEA.

The fuzzy reasoning was carried out in [9] through decision producer for choosing the appropriate suggestion in multi-source Dempster–Shafer (DS) related organization procedure. However, the ranking time was not minimized by DS based classification algorithm. A semantic web related document ranking structure was introduced in [10] with keywords and conceptual instances between the keywords. The applicable page was displayed happening highest of the web pages. But, the false optimistic amount was not decreased by designed scheme.

The problems identified from the above literature are lesser ranking accurateness, higher ranking time, higher false positive ratio, higher computational cost, higher computational complexity, higher space complication and more. In direction towards address these problems, Lancaster Stem Sammon Projective Feature Selection based Stochastic eXtreme Gradient Boost Clustering (LSSPFS-SXGBC) Approach is introduced.

The key involvement of the work is specified as follows:

- The aim of LSSPFS-SXGBC approach is towards accomplish effective page ranking based on user query. LSSPFS-SXGBC approach performed three processes namely preprocessing, feature selection and clustering.
- Lancaster Stemming Preprocessed Analysis in LSSPFS-SXGBC Approach removes the noisy data from the input query. It eliminates the stem words, stop words and incomplete data for minimizing the time and space consumption.
- Sammon Projective Feature Selection Process in LSSPFS-SXGBC Approach chooses the relevant features (i.e., keywords) based on user needs for efficient page ranking. Sammon Projection maps the high-dimensional space to lower dimensionality space to maintain the inter-point distance structure.
- Stochastic eXtreme Gradient Boost Page Rank Clustering process groups the similar keyword web pages based on their rank. Gradient Boost Page Rank Cluster is an ensemble of several weak clusters (i.e., X-means cluster). X-means cluster partitions the web pages into ‘x’ numeral of clusters anywhere each reflection goes on the way to the cluster through adjacent mean value. For every weak cluster, selected features are taken as the training samples. Subsequently, all weak clusters are joined to form the strong cluster for attaining the webpage ranking results. By this way, an efficient page ranking is accepted through higher accurateness also minimum time consumption.

The structure of the article is organized as follows: The associated works of webpage ranking is discussed in section 2. In Section 3, particulars of the LSSPFS-SXGBC Approach are explained through a well-ordered architecture diagram. Section 4 discusses the practical setup through dataset. The result and deliberations are specified in the section 5. Finally, Section 6 concludes the paper.

II. ASSOCIATED WORKS

A ranking module was introduced in [11] with web page link structure to examine with Web structure mining (WSM) and their contented. It was examined with Web contented mining (WCM). But, error rate was not minimized by ranking module. A ranking algorithm was designed in [12] depending on page multi-attribute (PMA Rank). The designed algorithm employed index measuring procedure with pre-rank process for web pages. But, the computational cost was not decreased through designed algorithm.

The page ranking algorithm was introduced in [13] with integral part of search engines. The designed algorithm organized the web pages linked with queried TOI based on relevance level. The designed algorithm regulated search quality and operator experience for statistics recovery. However, the ranking accuracy was not improved by page ranking algorithm.

Multilingual information retrieval was carried out in [14] to identify the available relevant information rather than language used in query. The user was employed towards delivered documents in languages diverse since native one. But, the ranking time was not reduced by designed retrieval. A deep learning-based search ranking framework was designed in [15] to apprise the ranking design through collecting the real-time operator clickstream information. The log parser ingested the web logs that information folder user performance in real-time style. But, the computational cost was not minimized through designed framework.

A new intelligent framework was introduced in [16] for web page organization and re-ranking. The designed framework attained improved organization and re-ranking presentation with decrease space necessities also search time in web credentials. However, the computational complexity was not reduced by intelligent framework. Data Envelopment Analysis (DEA) was designed in [17] to present the list with most-retrieved in global ranked through page load time forecast on dimensions and numeral of items. The decay was carried out from inverse frontier perspectives. But, the computational cost was not decreased by DEA.

SALSA approach was presented in [18] to provide the sNorm(p) for well-organized ranking of web pages. The designed method was based on p-Norm since Vector Norm group for web pages ranking. Though, the time difficulty was not decreased through SALSA method. Machine learning related subsequent generation of web page ranking procedure called Advanced Cluster Vector Page Ranking algorithm (ACVPR) was introduced in [19]. ACVPR algorithm helped the user through provided that web page ranking towards gratify the adapted requirements. However, ACVPR and IMSS-P tool was not sophisticated to accomplish web search for adapted web links.

A microscopic model was introduced in [20] to random robot in swarm through executing the local actions. The optimization was carried out through this incited algorithm where the suitability function increased the PageRank score. But, the ranking accurateness was not enhanced by microscopic model.

III. METHODOLOGY

Page Ranking is groundbreaking algorithm used to recover the significant records and to lessen the user searching time. But, the ranking accurateness was not enhanced by existing methods. In direction to achieve this objective, a novel LSSPFS-SXGBC Approach is introduced with preprocessing, keyword selection, and classification.

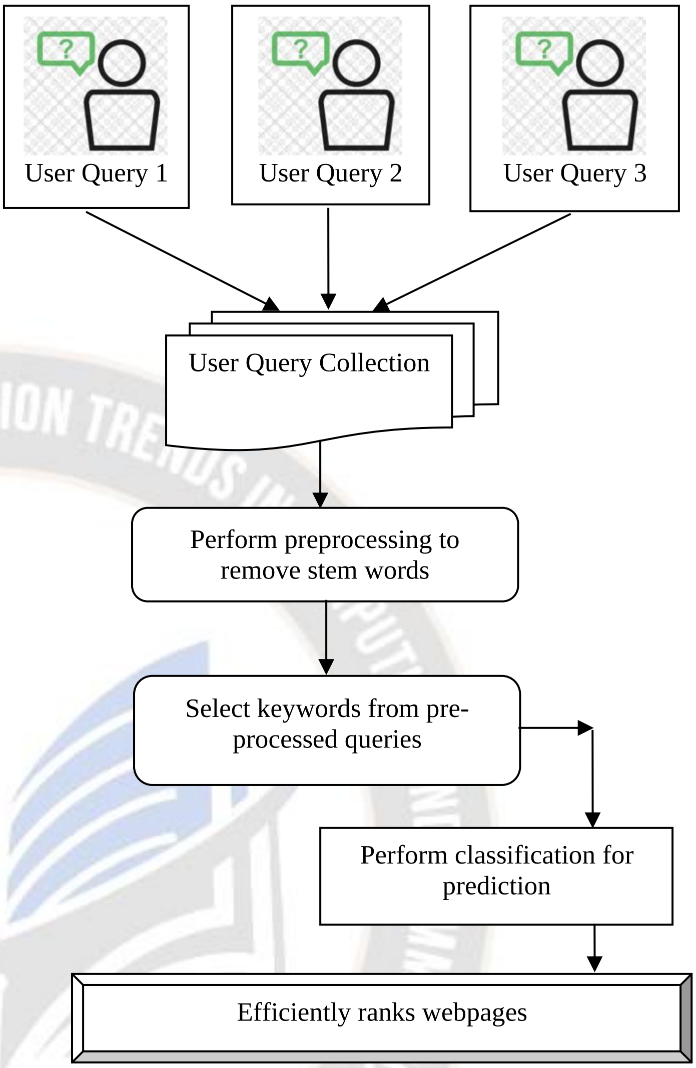
User Query 1
User Query 2
User Query 3
User Query Collection
Perform preprocessing to remove stem words
Select keywords from pre-processed queries
Perform classification for prediction
Efficiently ranks webpages

Figure 1 Architecture Design of Proposed LSSPFS-SXGBC Approach

Figure 1 indicates the architecture design of proposed LSSPFS-SXGBC approach to rank the web pages with respect to user query through preprocessing, keyword extraction, and classification. Initially, the numeral of user enquiries is engaged as an input. After that, input user queries are preprocessed using Lancaster Stemming process for removing the stop words and stem words. Sammon Projection in LSSPFS-SXGBC approach extracts the keywords from the preprocessed data. Then, the user queries are grouped with the extracted keywords through the Stochastic eXtreme Gradient Boost Page Rank Clustering. By this way, the web pages get ranked to minimize the time and space complexity.

3.1 Query preprocessing

Query pre-processing is the first step in LSSPFS-SXGBC approach to reduce the web page ranking time consumption. In pre-processing stage, the stop words and stem words are removed from the user query. In LSSPFS-SXGBC approach, the stopword removal is carried out using simple match coefficient and stem words are removed through Lancaster stemming algorithm.

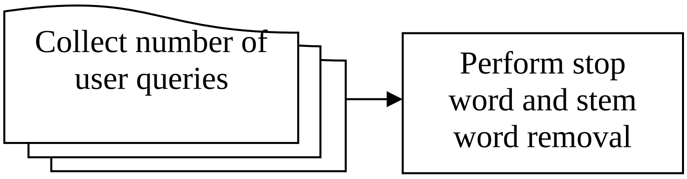
Collect number of user queries
Perform stop word and stem word removal

Figure 2 Query preprocessing

Figure 2 illustrates the process diagram of enquiry pre-processing. The method is take into account that, the numeral of web user enquiries '$wuq_1, wuq_2, wuq_3, \ldots wuq_n$'. For each web user query, the stop and stem word removal is carried out. The words from the web user query are '$wo_1, wo_2, wo_3, \ldots . wo_m$'. Each word in web user query is matched with list of existing stop words using simple match coefficient. The number of stop words comprised the collection of words. Simple match coefficient is a function used to measure the relation between query keywords and list of stop words. Simple match coefficient (SMC) is employed as,

$$SMC = \frac{wo_i}{N} \quad (1)$$

From (1), 'wo_i' symbolizes the list of stop words in web user query. 'N' represent the number of predefined stop words. 'SMC' value ranges from 0 to 1. Depending on the coefficient value, the stop words are identified and removed from user query.

3.1.1 Lancaster Stemming

Lancaster Word stemming process is second step performed in pre-processing step of proposed LSSPFS-SXGBC Approach. Word stemming process is employed to extract the root words through eliminating the suffixes from the word. The stems are used for finding the benefit of increasing recall through retrieving terms with same roots but different endings. When searching with the stems, it is common to recover many irrelevant terms that with similar roots but not related to the search objects. The stemmer includes the stemming algorithm and collection of the stemming guidelines. The standard guidelines present robust stemmer. Stemmer strength is a quality for index density and makes larger numeral of overstemming mistakes corresponding towards the numeral of understemming mistakes. The users who require lighter stemmer advance their individual set of guidelines. Stemmer is an iterative aspect one and guidelines denote the elimination or replacement of an conclusion. The replacement procedure evades the require for a divide phase in the procedure towards recode or deliver partial similar identification. Stemming process assistances towards uphold the competence. The guidelines are indexed through last letter of the conclusion for well-organized searching.

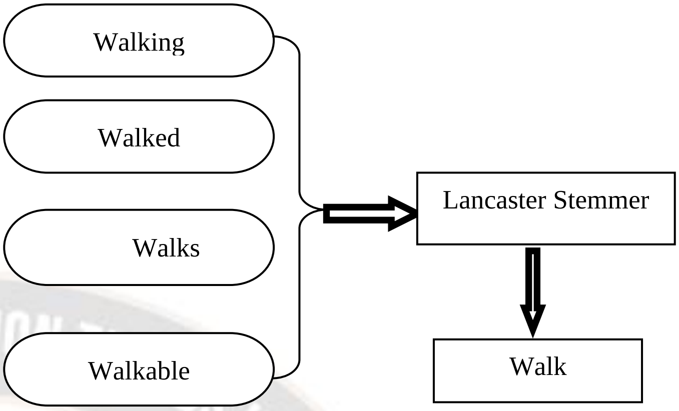

Figure 3 Lancaster Stemmer Process

Figure 3 explains the lancaster stemmer process. By using lancaster stemmer process, the word ends with 'ing', 'ed', 's', and 'able' are termed as the suffixes that are eliminated from the query and attain the origin word 'Walk'. In big data query processing, the words consume large amount of time towards discovery the user stimulating web pages. Consequently, the technique accomplishes the preprocessing to enhance query processing through minimum time. Jaro Similarity Matching is an approximate string-matching process. The matching process contains substring approximately equal to the given pattern. When substring and pattern are within specified distance of each other, then Jaro Similarity Matching algorithm considered as an equal. The Jaro Similarity Matching is given as,

$$JS = \frac{\max\left(|wo_1|, |wo_2|\right)}{2} \quad (2)$$

From (2), 'JS' denotes the jaro similarity metric. 'JS' similarity metric attains the output values between zero and one. With help of the matching value, the accurate word-stemming process is carried. When both the string length is same, the words stem process is not carried out in accurate manner. When the string length is not matched, word stem is not performed in accurate manner.

3.2 Sammon Projective Feature Selection Process

In LSSPFS-SXGBC approach, sammon projective process maps the features from high-dimensional space to the lower dimensionality space. Sammon projective feature selection process is non-linear approach where minimization is carried out through iterative gradient descent method. The numeral of iterations is determined by experimentation and convergent resolutions are not assurance. Sammon mapping is the successful nonlinear metric multidimensional scaling on the method of stress function. The diagrammatic representation of the sammon projective is given in figure 4.

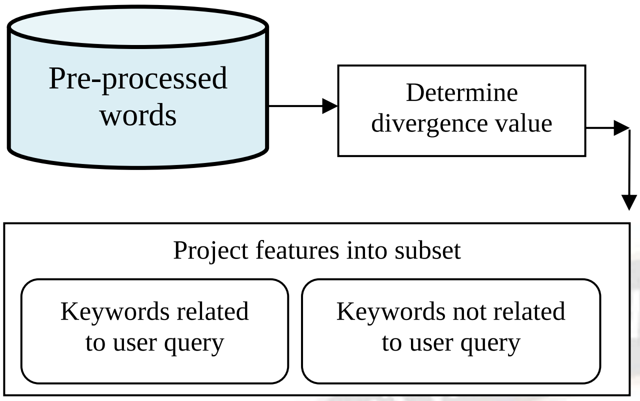
Pre-processed words
Determine divergence value
Project features into subset
Keywords related to user query
Keywords not related to user query

Figure 4 Sammon Projective Keyword Selection Process

Sammon mapping is enhanced by increasing the stress function by means of left Bregman divergence and right Bregman divergence.

$$BD = \frac{1}{\sum_{a<b} wo_{ab}^*} \sum_{a<b} \frac{(wo_{ab}^* - wo_{ab})}{wo_{ab}^*} \quad (3)$$

Let us consider that the number of words in user query is '$wo_1, wo_2, wo_3, \ldots . wo_m$'. By applying divergence, the distance is resolute among words and web user query. The divergence rate differs from '0' to '1'. Afterward that, the threshold rate is pre-set towards map the input significands from the database into any subsets. Subsequently, the sammon mapping outcome is given as,

$$SM \rightarrow \begin{cases} BM > th\,; Selected\ as\ keyword \\ BM < th\,; Not\ selected\ as\ keyword \end{cases} \quad (4)$$

From (4), 'SM' represents the sammon mapping function. 'th' denotes the threshold. The word that gets more diverged from web user query is considered as the non-keyword. Or else, the words are considered as the keywords. This in turn, the keywords are selected from the web user query. The algorithmic process of Sammon Projective Feature Selection in LSSPFS-SXGBC approach is given as,

```
// Algorithm 1: Sammon Projective Feature Selection
Input: Number of words '$wo_1, wo_2, wo_3, \ldots . wo_m$'
Output: Selected relevant keywords
1. Begin
2.   Number of words '$wo_1, wo_2, wo_3, \ldots . wo_m$' engaged as an input
3.   For individual feature '$wo_j$'
4.      Compute the divergence
5.      Map the features into subsets related on divergence rate
6.      Choice the applicable keywords from database
7.   End for
8. End
```

The algorithmic steps of sammon projective feature selection are explained in LSSPFS-SXGBC approach with the feature choice procedure. Primarily, the numeral of words is taken in account as an input. At that point, the divergence value is determined for each word. Depending on divergence values, the applicable keywords are selected to perform efficient web page ranking.

3.3 Stochastic eXtreme Gradient Boost Page Rank Clustering

Collaborative learning is a type of machine learning technique where number of weak learners is trained to attain better performance results. Boosting is a machine learning technique that alters the weak learner into form strong learner. The weak learner is a cluster with the correlated results. As a result, boosting method is selected for performing the clustering process. Here in LSSPFS-SXGBC approach, X-means clustering is considered as the base learner. The stochastic extreme gradient boost clustering is carried out in LSSPFS-SXGBC approach to combine the weak learner into form the strong learner for refining the clustering results. The gradient boosting computes the residues of previous replicas and joins organized to attain the target outcome. Extreme gradient boost (XGB) is used to control the over-fitting and to reduce the complexity for attaining the better clustering performance. LSSPFS-SXGBC approach uses an ensemble of X-means clustering to rank the web pages.

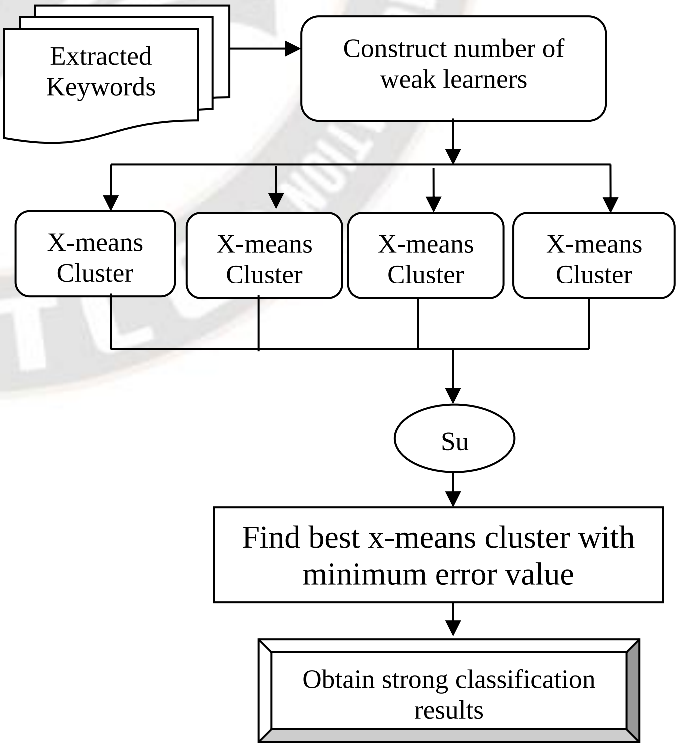
Extracted Keywords
Construct number of weak learners
X-means Cluster
X-means Cluster
X-means Cluster
X-means Cluster
Su
Find best x-means cluster with minimum error value
Obtain strong classification results

Figure 5 Flow process of Strong Classification Process

Figure 5 explains the flow process of strong classification procedure. Let us taken into account that the numeral of extracted keywords '$kw_1, kw_2, kw_3 \ldots\ldots, kw_n$' from web user query. Extreme gradient boosting classifier categorizes the web patterns by the help of base learners. The Extreme gradient boosting classifier usages a training set $\{(x_1, Clu_1), (x_2, Clu_2), (x_3, Clu_3), \ldots (x_n, Clu_n)\}$ where 'x_i' denotes an input i.e. keywords, 'Clu_i' represents the cluster output. The number of keywords is extracted from user query. Then, 'X' number of clusters and their weight value are allocated. The comparison among the keywords and cluster weight is identified using Tanimoto Indexing. Depending on the comparation value, the keywords are allocated to the exact group. It is formulated as,

$$kw_i = kw_{1,} kw_2, kw_3 \ldots.., kw_n \tag{5}$$

From (5), 'kw_i' represent the set of keywords. By applying the LSSPFS-SXGBC approach, the 'X' number of clusters is initialized. It is computed as,

$$Clu_j = clu_1, clu_2, clu_3, \ldots. clu_x \tag{6}$$

From (6), 'Clu_i' denotes a cluster set and '$clu_1, clu_2, clu_3, \ldots. clu_x$' denotes '$X$' number of clusters. Then, the weight value is allocated arbitrarily to individual cluster. It is defined as,

$$weight_{Clu} \rightarrow \{clu_1, clu_2, clu_3, \ldots. clu_x\} \tag{7}$$

From (7), '$weight_{Clu}$' symbolizes the weight value to every cluster. Afterward weight allocation towards individual cluster, the web pages get ranked. Clustering is a dividing of web pages into dissimilar groups based on the keywords. Each group is termed as the cluster which comprises web pages that are parallel towards one another also dissimilar to other groups. The Tanimoto Comparation coefficient is used in LSSPFS-SXGBC approach for measuring the similarity between the webpage with extracted keywords and weight value of cluster is identified as follows,

$$TS(kw_i, Clu_j) = \frac{|kw_i \cap Clu_j|}{|kw_i| + |kw_i| + |kw_i \cap Clu_j|} \tag{8}$$

From (8), the intersection symbol '$\cap$' represents a mutual necessity among the keywords and weight value of cluster. $|kw_i|$ and $|Clu_j|$ symbolizes the cardinalities of the two sets. Tanimoto comparation coefficient value ranges between the values '0' and '1'. The cluster weight value and keywords are similar when the output value achieved is '1'. But the, two-weight value is not similar, the output value is '0'. Afterward computation the similarity value, the more similar data points are gathered into the cluster. The data point is not belongs towards any of the cluster, then weighted iterative x-means clustering enhances cluster works by repetitively trying subdivision, by help of Bayesian probability criterion. This helps towards allocate the whole data points into the specific cluster through the possibility. The criterion is definite as,

$$BPC = -2 \, log \, Prb \, (kw_i | weight_{Clu}) + \, log(n) \tag{9}$$

From (9), 'BPC' represent the Bayesian probability criterion function. 'Prb' symbolizes the probability operation and 'n' symbolizes the numeral of keywords. $Prb(kw_i | weight_{Clu})$ symbolizes the make the most of value of the probability function of the keywords and cluster. The probability illustrates the probability of keywords in webpage to get grouped into particular cluster. Accordingly, all the web pages are assembled into any cluster related on their weight value. This procedure gets iterated while waiting for the webpages are grouped. This in turn assistances towards enhance the clustering accurateness as well as decrease the time consumption. The flow chart of Tanimoto Indexed X-means Clustering is described as follows,

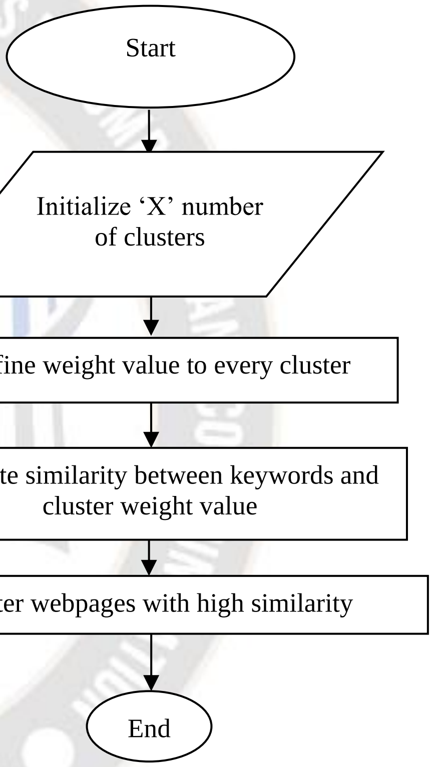
Start
Initialize 'X' number of clusters
Define weight value to every cluster
Compute similarity between keywords and cluster weight value
Cluster webpages with high similarity
End

Figure 6 Flowchart of Tanimoto Indexed X-means Clustering

Figure 6 illustrates the flowchart of Tanimoto Indexed X-means Clustering process in LSSPFS-SXGBC approach towards enhance the clustering procedures. The clusters and weight values are adjusted arbitrarily. Afterward that, the comparation is measured for each cluster weight value and keywords. This assistances towards divider all data points into 'X' clusters. The any of the webpages transfer into the group, then the Bayesian probability is calculated. As a outcome, complete webpages are grouped into a specific cluster. But the clustering results were not improved by existing weak learners. Therefore, a stochastic extreme gradient boost cluster identify the loss function of complete x-means cluster to form strong cluster for improving the webpage ranking accurateness. The output of individual x-means cluster is the summation of every individual cluster. It is expressed as,

$$y_a = \sum_{i=1}^{n} Clu_j(kw_i) \tag{10}$$

From (10), 'y_a' symbolizes the actual output of strong cluster. 'Clu_j' represent the base cluster output. The gradient boost cluster is illustrated as,

$$objective\ function = Training\ loss + regularization\ term \tag{11}$$

From (11), the objective function is determined. The regularization term controls complexity. XGBoost cluster reduced the training loss to avoid over-fitting problems. Then in LSSPFS-SXGBC approach, XGBoost cluster determined the pseudo residuals. The residual is considered as the gradient of loss function. The Gaussian loss function is measured as following equation,

$$\sigma_l(kw_i) = \frac{1}{2}\left(y_a - Clu_j(kw_i)\right)^2 \tag{12}$$

From (12), 'σ_l' symbolizes the Gaussian loss function. 'y_a' represent the actual output. '$Clu_j(kw_i)$' denotes the forecast output. Depending on the loss function, the pseudo residues are computed for every iteration. It is determined as,

$$\rho = -\left[\frac{\partial(y_a, \sigma_l(kw_i))}{\partial\ \sigma_l(kw_i)}\right] \tag{13}$$

From (13), 'ρ' represent the pseudo residuals. The base learner is appropriate to pseudo residues. The greatest gradient descent step-size is computed for identifying the least loss function of base cluster. It is formulated as,

$$GDS = arg\,\underset{\rho}{min} \sum_{i=1}^{n} [y_a, \sigma_{l-1}(kw_i) + objective\ function * Clu_j(kw_i)] \tag{14}$$

From (14), 'GDS' symbolizes the gradient descent step-size. The dispute minimum (arg min) function is employed to identify the least error of weak learner. The predictive model is modernized to rank the web pages. The updating design is described as,

$$y_a = \sum_{i=1}^{n} \sigma_{l-1}(kw_i) + objectivefunction * Clu_j(kw_i) \tag{15}$$

From (15), 'y_a' symbolizes the strong cluster results. The attained strong cluster output is help to rank the web pages. Accordingly, an collaborative of weak cluster is used to create strong cluster for excellently ranking webpages. By this way, the proposed LSSPFS-SXGBC approach ranks the web pages in efficient way. The algorithmic process of Stochastic eXtreme Gradient Boost Page Rank Clustering is formulated as,

```
// Algorithm 2:_Stochastic eXtreme Gradient Boost Page Rank Clustering
Input: Number of extracted keywords
Output: Rank web pages
Begin
1.For each training set $kw_i$
2.      Measure similarity between keywords and web pages
3.      Form x-number of cluster based on the similarity value to group the web pages
4.      For each iteration
5.          Determine the pseudo residuals
6.          Find the best x-means cluster
7.          Update the cluster model
8.          Attain strong cluster outcomes
9.      End for
10. End for
End
```

Algorithm 2 explains the ensemble cluster process towards rank the web pages related on user query. Initially, the extracted keywords are considered as an input for X-means cluster. In that cluster, 'X' number of clusters and their weight values are adjusted and the comparation between the extracted keywords and cluster weight value is determined for grouping the more related web pages. Stochastic Extreme gradient boost clustering calculates the loss functions of all x-means cluster for providing strong cluster results. Therefore LSSPFS-SXGBC approach effectively ranks the web pages related on user query.

IV. Experimental Setup

The experimental analysis of proposed in LSSPFS-SXGBC approach and conventional methods namely, MCIR (Mohamed Attia et al., 2022) and innovative approach (Carla Limongelli et al., 2022) are executed in the java software language. The incorporation of ACM dataset (http://ir.dcs .gla.ac.uk/resources/test_collections/cacm/) and Cranfield dataset http://ir.dcs.gla.ac.uk/resources/test_collections/cran/ are taken as input to retrieve more relevant web pages for conducting experiments. CACM dataset comprises the gathering of web-based documents, input enquiries and stops word lists in dissimilar files. Cacm.all file includes the text of documents then the common_words file includes the stop words. The text file comprises 64 user enquiries. Cranfield dataset comprises two such as 1400 gatherings and 200 gatherings. The dataset comprises the (cran.all) documents of web pages, queries (cran.qry), then applicable valuations (i.e. cranqrel). The cran.qry includes the 365 enquiries employed for web page ranking based on preprocessing, and keyword extraction.

ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6537
***Article Received**: 27 January 2023 **Revised**: 28 February 2023 **Accepted**: 06 March 2023*

V. Result And Discussion

The evaluation of the proposed LSSPFS-SXGBC approach and two existing techniques specifically MCIR [1] and innovative approach [2] are compared with certain parameters, namely ranking accuracy, false-positive rate, and ranking time and memory consumption.

5.1 Analysis of Ranking Accurateness

Ranking Accurateness is definite as the ratio of web pages that are appropriately ranked with respect to numeral of user queries. The ranking accuracy is formulated as,

$$RA = \left[\frac{Q_n\ (P_{cr})}{Q_n}\right] * 100 \qquad (16)$$

From (16), 'RA' denotes the ranking accuracy. 'Q_n' symbolize the number of user queries, '$Q_n(P_{cr})$' symbolizes the number of queries where webpages are correctly ranked. The ranking accurateness is computed in relations of percentage (%). When the ranking accurateness is higher, the technique is supposed to be more competent.

Table 1 Tabulation for Ranking Accuracy

Number of User Queries (Number)	Ranking Accuracy (%)		
	Proposed LSSPFS-SXGBC approach	MCIR	Innovative Approach
40	95	90	85
80	93.75	87.5	82.5
120	95.83	89.17	85
160	96.25	88.13	83.75
200	95	89	85
240	97.5	87.5	84.17
280	97.5	89.29	86.07
320	96.88	87.5	85
360	96.94	90.28	86.11
400	97.5	88	82.5

Table 1 designates the ranking accurateness of three dissimilar approaches proposed LSSPFS-SXGBC approach and existing techniques namely MCIR [1] and innovative approach [2] from two different datasets, namely ACM dataset and Cranfield dataset. The various number of user queries are collected from database. For conducting experiments, let us consider number of user queries is 120. The ranking accuracy attained using the LSSPFS-SXGBC approach is '95.83%' whereas the ranking accuracy attained using MCIR and Innovative Approach are '89.17%' and '85%' respectively. The performance of ranking accurateness of the proposed LSSPFS-SXGBC approach is considerably improved when associated towards the conventional methods. The simulation chart through diverse ranking accuracy outcomes is shown in the figure 7.

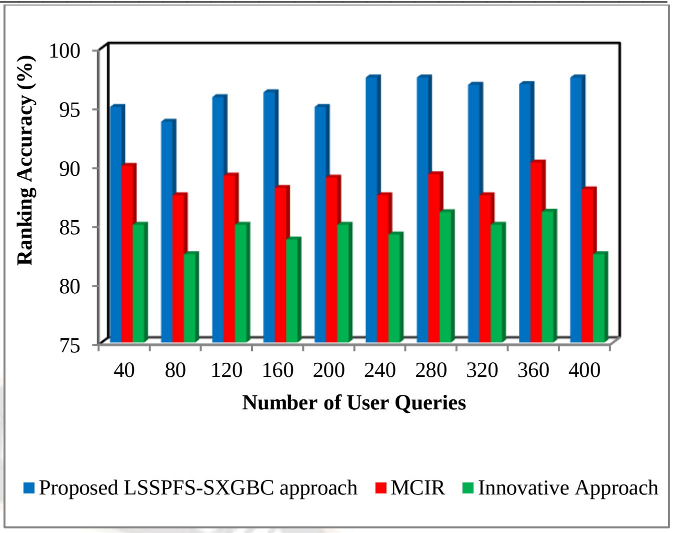

Figure 7 Measurement of Ranking Accuracy

Figure 7 illustrates the impact of ranking accuracy by different number of queries ranging from 40, 80, 120…400. From the figure, it is observed that ranking accuracy gets increased or decreased with different number of user queries. For experimentation, 10 iterations are carried out with three different methods. This significant enhancement is attained through Lancaster Stemming Preprocessed Analysis for removing the noisy data from the input and Sammon Projective Feature Selection Process to select the relevant features based on user needs for page ranking. In addition, Stochastic eXtreme Gradient Boost Page Rank Clustering process groups the web pages for efficient webpage ranking. This in turn, the webpage ranking process is performed with higher accuracy. As a result, the LSSPFS-SXGBC approach increased the ranking accuracy by 9%, and 14% when associated to the existing MCIR model [1] and innovative method [2] correspondingly.

5.2 Analysis of False Positive Rate

False Positive Rate is definite as the ratio of numeral of web pages that are wrongly ranked with respect to the numeral of user queries. It is determined as,

$$R_{FP} = \left(\frac{Q_n\ (P_{Icr})}{Q_n}\right) * 100 \qquad (17)$$

From (17), 'R_{FP}' symbolizes the false positive rate, 'Q_n' symbolizes the numeral of user queries. '$Q_n\ (P_{Icr})$' represents the numeral of queries where the web pages are incorrectly ranked. The false positive rate is calculated in relations of percentage (%). When the false positive rate is lesser, the technique is supposed to be more competent.

Table 2 Tabulation of False Positive Rate

Numeral of User Queries (Number)	False Positive Rate (%)		
	Proposed LSSPFS-SXGBC approach	MCIR	Innovative Approach
40	5	10	15
80	6.25	12.5	17.5
120	4.17	10.83	15
160	3.75	11.87	16.25
200	5	11	15
240	2.5	12.5	15.83
280	2.5	10.71	13.93
320	3.12	12.5	15
360	3.06	9.72	13.89
400	2.5	12	17.5

Table 2 explains the false positive rate of three dissimilar approaches, specifically proposed LSSPFS-SXGBC approach and existing techniques namely MCIR [1] and innovative approach [2] from ACM dataset and Cranfield dataset. The various number of user queries are collected from two input database. For conducting experiments, let us consider numeral of user enquiries is 360. The false positive rate attained help the LSSPFS-SXGBC approach is ‘3.06%’ whereas the false positive rate attained helping MCIR and Innovative Method are ‘9.72%’ and ‘13.89%’ respectively. The function of false positive rate of the proposed LSSPFS-SXGBC approach is significantly decreased when associated towards the conventional methods. The simulation chart through dissimilar false positive rate outcomes is illustrated in the figure 8.

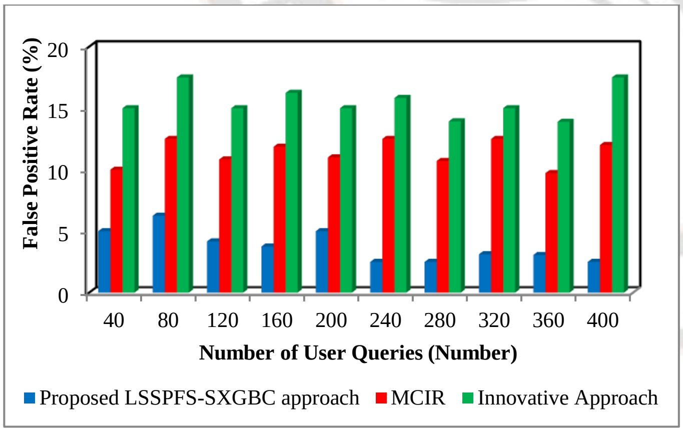

Figure 8 Measurement of False Positive Rate

Figure 8 illustrates the false positive rate by dissimilar numeral of queries ranging from 40, 80, 120…400. From the figure, it is clear that false positive rate gets increased or decreased with numeral of enquiries. For investigation, 10 iterations are performed and comparison is carried out with three different methods. This significant enhancement is attained through query preprocessing. After collecting the input queries, the proposed LSSPFS-SXGBC approach eliminates the stop and stem words. The proposed LSSPFS-SXGBC approach uses the jaro similarity matching with substring approximately equal to the given pattern. The exchange method evades the necessity for separate phase towards recode or deliver partial matching. By this way, the webpage ranking process is performed through minimum false positive rate. As a result, the LSSPFS-SXGBC approach reduced the false positive rate by 66%, and 76% when associated towards the existing MCIR model [1] and innovative approach [2] correspondingly.

5.3 Analysis of Ranking Time

Ranking time is described as the quantity of time consumed for ranking the web pages based on user enquiry. The ranking time is determined as,

$$RT = Q_n * T\ (webpage\ ranking\ for\ one\ user\ query)$$
(18)

From (18), ‘RT’ symbolizes the ranking time, ‘Q_n’ symbolizes the numeral of user enquiries. ‘T’ specifies the quantity of time engaged to process one user enquiry.

Table 3 Analysis of Ranking Time

Number of User Queries (Number)	Ranking Time (ms)		
	Proposed LSSPFS-SXGBC approach	MCIR	Innovative Approach
40	11	14	18
80	13.2	17.6	20.2
120	16	21.6	25
160	18	25.6	28.4
200	21	28	33.7
240	23.5	31.2	38
280	25	33.6	41.2
320	28	36.8	45.1
360	31.8	38.88	47.9
400	34	42	49.6

Table 3 demonstrates the overall performance outcomes of ranking time consumption. For conducting experiments, let us taken into account the number of user queries is 240. The ranking time consumed using the LSSPFS-SXGBC approach is ‘23.5ms’ whereas the ranking time consumed using MCIR [1] and Innovative Approach [2] are ‘31.2ms’ and ‘38ms’ respectively. The simulation chart through dissimilar ranking time outcomes is demonstrated in the figure 9.

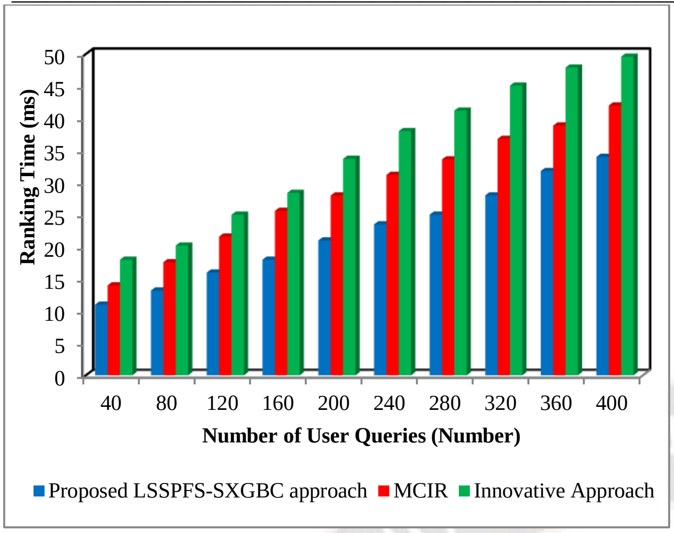

Figure 9 Measurement of Ranking Time

Figure 9 illustrates the impact of ranking time by different number of queries ranging start the value 40, 80, 120…400. From the figure, it is observed that ranking time consumption gets increased with increasing number of user queries. For experimentation, 10 iterations are conducted with three different techniques. This enhancement is achieved through Lancaster Stemming Preprocessed Analysis for removing the noisy data (i.e., stem words and stop words) from the input. Then, Sammon Projective Feature Selection Procedure is accepted to choice the applicable features based on the user requirements for webpage ranking. This in turn, the webpage ranking process is performed through least time consumption. As a result, the LSSPFS-SXGBC approach reduced the ranking time consumption by 24%, and 36% when associated to the existing MCIR model [1] and innovative method [2] correspondingly.

5.4 Analysis of Space Complexity

Space complexity is definite as the quantity of space consumed for ranking the web pages depending on the user query. The ranking time is determined as,

$$SC = Q_n * Space\ consumed\ (\ webpage\ ranking\ for\ one\ user\ query) \quad (18)$$

From (18), ‘SC’ denotes the space complexity, ‘Q_n’ symbolizes the number of user queries. When the space difficulty is smaller, the technique is supposed to be more competent.

Table 4 Analysis of Space Complexity

Numeral of User Queries (Number)	Space Complexity (MB)		
	Proposed LSSPFS-SXGBC approach	MCIR	Innovative Approach
40	9.2	11.2	13.81
80	11	13.6	15.4
120	12.14	14.4	17.28
160	14.6	16.8	19.48
200	15.8	18	20.68
240	17.2	20.4	23.18
280	19.49	22.4	24.82
320	20.36	24	26.16
360	21.40	25.2	28.28
400	23.54	26	29.58

Table 4 illustrates the overall functioning outcomes of space complexity. For conducting experiments, let us taken into account the number of user queries is 280. The space complexity using the LSSPFS-SXGBC approach is ‘17.2MB’ whereas the ranking time consumed using MCIR [1] and Innovative Approach [2] are ‘ 20.4MB ’ and ‘ 23.18 ’ respectively. The performance of space complexity of the proposed LSSPFS-SXGBC approach is considerably minimized when associated to the conventional methods. The simulation chart through dissimilar space complexity outcomes is represented in the figure 10.

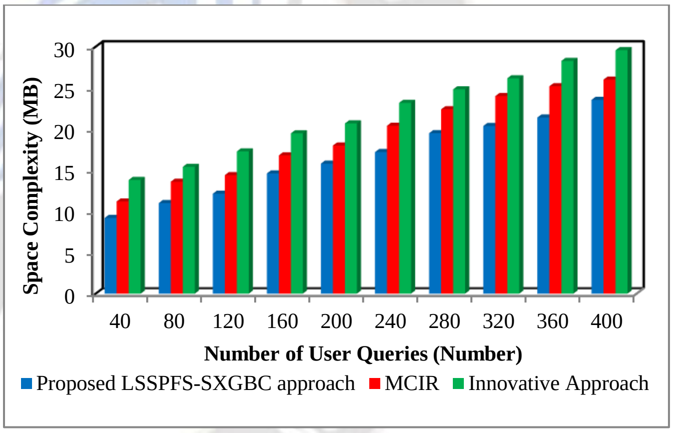

Figure 10 Measurement of Space Complexity

Figure 10 explains the impact of space complexity by different numeral of queries ranging from 40, 80, 120…400. From the figure, it is clear that space complexity gets increased with increasing number of user queries. For experimentation, 10 iterations are carried out with three different methods. The functioning of ranking time of the proposed LSSPFS-SXGBC approach is considerably minimized when associated to the conventional methods. This enhancement is attained through Lancaster Stemming Preprocessed Analysis for eliminating the stem words and stop words since the input user query. Sammon Projective Feature Selection Process is chooses the relevant features based on the user needs for efficient webpage ranking. This in turn, the webpage ranking process is performed through least space complexity. As a outcome, the LSSPFS-SXGBC approach minimize the space complexity by 15% and 25% when associated towards the existing MCIR model [1] and innovative approach [2] respectively.

VI. CONCLUSION

In this paper, LSSPFS-SXGBC approach is introduced for webpage ranking based on user query. The numeral of user queries is engaged as input. Then, stop words and stem words are eliminated through query preprocessing to minimize the ranking time and space complexity. With preprocessed words, LSSPFS-SXGBC approach extracts the relevant keywords using Sammon Projective Feature Selection process. After that, the ranking process is performed using Stochastic eXtreme Gradient Boost Page Rank Clustering process. The designed process groups the web pages for efficient webpage ranking through higher accurateness. The investigational evaluation is approved using dissimilar functioning metrics such as ranking accurateness, false-positive rate, ranking time and space complexity versus a numeral of queries collected from the dataset. The quantitative analysis confirms that the LSSPFS-SXGBC approach has attained higher accuracy of webpage ranking with lesser time consumption as well as space complexity when associated towards other conventional approaches.

REFERENCES

[1] Mohamed Attia, Manal A. Abdel-Fattah and Ayman E. Khedr, "A proposed multi criteria indexing and ranking model for documents and web pages on large scale data", Journal of King Saud University - Computer and Information Sciences, Elsevier, Volume 34, Issue 10, Part A, November 2022, Pages 8702-8715

[2] Carla Limongelli, Matteo Lombardi, Alessandro Maran, and Davide Taibi, "A Semantic Approach to Ranking Techniques: Improving Web Page Searches for Educational Purposes", IEEE Access, Volume 10, 2022, Pages 68885 - 68896

[3] Ahmet Selman Bozkir and Ebru Akcapinar Sezer, "Layout-based computation of web page similarity ranks", International Journal of Human-Computer Studies, Elsevier, Volume 110, February 2018, Pages 95-114

[4] N Jayalakshmi, V Sangeeta and Appala Srinuvasu Muttipati, "Taylor Horse Herd Optimized Deep Fuzzy clustering and Laplace based K-nearest neighbor for web page recommendation", Advances in Engineering Software, Elsevier, Volume 175, January 2023, Pages 1-15

[5] Jangwan Koo, Dong-Kyu Chae, Dong-Jin Kim and Sang-Wook Kim, "Incremental C-Rank: An effective and efficient ranking algorithm for dynamic Web environments", Knowledge-Based Systems, Elsevier, Volume 176, 15 July 2019, Pages 147-158

[6] Leandro Tortosa, Jose F. Vicent and Gevorg Yeghikyan, "An algorithm for ranking the nodes of multiplex networks with data based on the PageRank concept", Applied Mathematics and Computation, Elsevier, Volume 392, 1 March 2021, Pages 1-15

[7] Sergio Jimenez, Fabio N Silva, George Dueñas and Alexander Gelbukh, "ProficiencyRank: Automatically ranking expertise in online collaborative social networks", Information Sciences, Elsevier, Volume 588, April 2022, Pages 231-247

[8] Késsia Nepomuceno, Thyago Nepomuceno and Djamel Sadok, "Measuring the Internet Technical Efficiency: A Ranking for the World Wide Web Pages", IEEE Latin America Transactions, Volume 18, Issue 06, June 2020, Pages 1119 - 1125

[9] Moitrayee Chatterjee and Akbar Siami Namin, "A fuzzy Dempster–Shafer classifier for detecting Web spams", Journal of Information Security and Applications, Elsevier, Volume 59, June 2021, Pages 1-18

[10] P. Chahal, M. Singh and S. Kumar, "An Efficient Web Page Ranking for Semantic Web", Journal of the Institution of Engineers (India): Series B, Springer, Volume 95, 2014, Pages 15–21

[11] Prem Sagar Sharma, Divakar Yadav, and R. N. Thakur, "Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey", Mobile Information Systems, Hindawi Publishing Corporation, Volume 2022, 2022, Pages 1-19

[12] Mohammed Rashad Baker and M. Ali Akcayol, "A novel web ranking algorithm based on pages multi-attribute", International Journal of Information Technology, Springer, Volume 14, 2022, Pages 739–749

[13] Shubham Goel, Ravinder Kumar, Munish Kumar and Vikram Chopra, "An efficient page ranking approach based on vector norms using sNorm(p) algorithm", Information Processing & Management, Elsevier, Volume 56, Issue 3, May 2019, Pages 1053-1066

[14] P.V. Vidya, P.C. Reghu Raj and V. Jayan, "Web Page Ranking Using Multilingual Information Search Algorithm - A Novel Approach", Procedia Technology, Elsevier, Volume 24, 2016, Pages 1240-1247

[15] Yun Li, Yongyao Jiang, Chaowei Yang, Manzhu Yu, Lara Kamal, Edward M. Armstrong, Thomas Huang, David Moroni and Lewis J. McGibbney, "Improving search ranking of geospatial data based on deep learning using user behavior data", Computers & Geosciences, Elsevier, Volume 142, September 2020, Pages 1-15

[16] Syed Ahmed Yasin and P. V. R. D. Prasada Rao, "Enhanced CRNN-Based Optimal Web Page Classification and Improved Tunicate Swarm Algorithm-Based Re-Ranking", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 30, Issue 05, 2022, Pages 813-846

[17] Késsia Nepomuceno, Thyago Nepomuceno and Djamel Sadok, "Measuring the Internet Technical Efficiency: A Ranking for the World Wide Web Pages", IEEE Latin America Transactions, Volume 18, Issue 06, June 2020, Pages 1119 - 1125

[18] Shubham Goel, Ravinder Kumar, Munish Kumar and Vikram Chopra, "An efficient page ranking approach based

***Article Received**: 27 January 2023 **Revised**: 28 February 2023 **Accepted**: 06 March 2023*

on vector norms using sNorm(p) algorithm", Information Processing and Management, Elsevier, Volume 56, 2019, Pages 1053–1066

[19] Dheeraj Malhotra and O.P. Rishi, "IMSS-P: An intelligent approach to design & development of personalized meta search & page ranking system", Journal of King Saud University - Computer and Information Sciences, Elsevier, November 2018, Pages 1-16

[20] M. Coppola, J. Guo, E. Gill and G. C. H. E. de Croon, "The PageRank algorithm as a method to optimize swarm behavior through local analysis", Swarm Intelligence, Springer, Volume 13, Issue 3–4, December 2019, Pages 277–319