An Efficient Probabilistic Methodology to Evaluate Web Sources as Data Source for Warehousing

Hariom Sharan Sinha[1], Saket Kumar Choudhary[2], Vijender Kumar Solanki[3]*

[1] Department of Computer Science & Engineering, Adamas University, Barasat, Kolkata, West Bengal (India)
[2] Department of Computer Science & Engineering, GITAM University, Bengaluru, Karnataka (India)
[3] Department of Computer Science & Engineering, CMR Institute of Technology, Hyderabad, Telangana (India)

Received 23 May 2022 | Accepted 1 February 2023 | Early Access 24 February 2023

Abstract

Internet is the largest source of data and the requirement of data analytics have fueled the data warehouse to switch from structured conventional Data Warehouse to complex Web Data Warehouse. The dynamic and complex nature of web poses various types of complexities during synthesis of web data into a conventional warehouse. Multi-Criteria-Decision Making (MCDM) is a prominent mechanism to select the best data for storing into the data-warehouse. In this article, a method, based on the probabilistic analysis of SAW and TOPSIS methods, has been proposed to select web data sources as data sources for web data warehouse. This method deals more efficiently with the dynamic and complex nature of web. Here, the result of the selection employs the analysis of both the methods (SAW and TOPSIS) to evaluate the probability of selection of respective score (1-9) for each feature. With these probability values, the probability of selection of the next web sources has been be determined. Moreover, using the same probability values, mean score and standard deviation of the scores of respective features of selected web sources have been deduced, which are further used to fix the standard score of each feature for selection of web sources. The standard score is a parameter of the proposed Mean-Standard-Deviation (MSD) method to check the suitability of web sources individually, whereas others do the same on comparative basis. The proposed method cuts down the cost of the repetitive comparison operation, once after computation of the Standard score using Mean and Standard deviation of each individual feature. Here, the respective value of the standard score of each feature is only compared with the score of each respective feature of the next web sources, so it reduces the cost of computation and selects the web sources faster as well.

Keywords

Mean-Standard-Deviation (MSD) Method, Multi-Criteria Decision Method (MCDM), Probabilistic Method, Standard Deviation of Score, Web Source.

DOI: 10.9781/ijimai.2023.02.012

I. Introduction

The evolution of Internet as well as the ease to share and to fetch the data from web have made web a magnificent platform of sources for information. Web is an independent platform to get and provide nearly all the types of information. At the mean time the requirement of data analytics for decision support system has obligated the data warehouse to deal with web data rather than traditional data, because the data from local data sources has turned insufficient for decision support systems. Despite being the data easily and publicly available on web, the web data cannot be queried and manipulated efficiently for data analytics as done in traditional Data Warehouse. So, the efficient way to use the data for data analytics is to exploit the warehouse technique rather than directly access the data. The Data warehouse main obligation is to collect information from various data sources to create repository and make integrated information available for Decision Support Systems. However, the exponential rising of web sources and complexity as well as dynamic nature of web have posed new challenges for data warehouse to deal with web-data from various and independent web sources [18], [22], [26], [27].

To find the suitable data to systematically incorporate it into a warehouse is an anticipating approach for data analytics. In order to collect the data for data warehouse, finding the relevant data on web is just as to find out needle in a haystack because of so many web sources [25], [26]. Besides, the dynamic nature of web data has made the situation more complicated and complex. So, the very first task for web warehouse is to find the relevant web sources as the data source for it. Thus, there is the requirement of evaluation of the relevancy and compatibility of the web sources. Various features must be entertained during evaluation of web sources. Zhu et al. have classified the features into three categories viz. web sources stability, web data quality and contextual issues of web data [25].

According to Zooknic statistics (http://www.zooknic.com/Domains/counts.html) on 15 December 2009, the total number of worldwide registered domains was 111,889,734, and these 111 million (around) websites are owned by government, private or individual organization and agencies, which causes complex (structured, semi-structured, unstructured) natured data designed in different (heterogeneous) styles. Besides so large number of websites, web sources are dynamic, the web data is updated frequently as well as even millions of new

* Corresponding author.
E-mail address: spesinfo@yahoo.com

web sources and web pages are being added every day on Internet. So already available sources may change or even disappear.

Another challenge is the quality of web-data because the web techniques are so opened and independent that web masters can fire whatever data they like on web. A big amount of data on the web is not properly examined, retrospected and percolated as done in conventional publications. Wrong, inconsistent, incomplete or vague data are easily available on web and even correct data are not properly presented. So, the quality of data on web is maverick. Third challenge is the context of data should fulfill the requirements of the user, because the availability of data on web is with the intention of browsing usually rather than for warehousing and analysis. So, the web-data must fulfill the requirements of web-warehousing, as relevancy of web data for analysis, easily extraction of necessitated data, all-important metadata (data definition, data format and derivation rules) etc. Probably these requirements may not be fulfilled.

Therefore, the designer of web-warehouse must build a set of features to evaluate the web sources to select the most suitable sources as data source for warehousing. In this article we will look into these challenges and discuss the methods for relevant web sources selection for warehousing. Firstly, a set of selection features is formulated and then evaluation of the web sources has been performed using Multi-Criteria Decision Method (MCDM) approach, (especially SAW and TOPSIS methods) with respect to these features. Again statistical and probabilistic analysis of the selected web sources has been done with respect to the score of the corresponding features. Then mean and standard deviation of the score of the corresponding features have been evaluated. Now using mean and standard deviation the relevancy of web source can be computed without any further relative comparison of web sources. Here only a fixed number of comparisons as the number of features and one more with the threshold value are required. So the computational complexity of the proposed method becomes constant.

The rest of the paper is organized as follows: Section II presents related work on web source selection as data source for warehousing and various approaches including MCDM (SAW and TOPSIS specially) which is based on evaluation of web sources. Section III explains the complexity during web sources selection and set of features for web source evaluation. Section IV explicates SAW and TOPSIS methods of MCDM, for selection of web sources as data source for warehouse. In Section V the proposed work has been explained. This section consists of three parts viz. statistical analysis of SAW and TOPSIS, Probability of selection of new web sources and Mean-Standard-Deviation (MSD) method based on mean and standard deviation. Section VI analyses the experimental setup and results of SAW, TOPSIS and MSD Methods and at last, Section VII presents the conclusion.

II. Related Work

During incorporating the data from web into warehouses, the dynamic and complex nature of web [2], [3], [6], [8], [9], [14] poses various challenges. Different approaches have been developed to overcome the challenges during warehousing the web data [2], [6], [17], [25], [29]. Doan et al. have explained XML technologies to extract, incorporate, store, query and analyze web data as well as their application to data warehouse [6]. Boussaid et al. haves proposed a UML (Unified Modeling Language) and XML model of warehousing along with the attributes of XML [2]. Hao Fan used HDM (Hypergraph Data Model) for warehousing the web data [8].

Another approach is comparative analysis of web sources to select the best one. In order to select the web source as data source for warehouse, quality of data available on web source is an important criteria of source selection. To define the quality of data, multiple features of web sources are entertained [20], [25]. So, the selection of a web source is multi features selection task [16]. To deal with the multi features selection problem [20], [25], Multi criterion Decision Making (MCDM) [13], [23], [33], [34] methods have been employed. With this method, on the basis of score and weight of features, the comparative analysis has been done. Having multiple criteria of decision, the MCDM approach is applicable in various real problems besides ranking of web sources [25] like [4] and many other problems. Le et al. [11] proposed a dynamic approach of web data warehousing using object oriented methodology to design the logical level for apprehending and presenting basic semantics of web sources and user requirements in a flexible and sensible way.

Moreover, Dong et al. have proposed a marginalism approach to select the web source. The marginalism approach is based on the marginalism principle of economics [12]. It restricts the selection of a new source till the marginal benefit is more than the marginal cost of integration. The marginal benefit is here the difference between benefit after and before the new source integration. Similarly marginal cost is the difference between the cost after and before the integration [7].

III. Features to Evaluate Web Sources

The evaluation features of web sources have been roughly classified into three major categories viz. web sources stability, web data quality and contextual issues of web data [18], [22], [25], [28].

A. Web Source Stability

This selection features can be further subcategorized into availability, durability, accessibility, and refreshing rate.

- Availability defines whether the specific site is up and in running mode, its response time and also reachability of the pages through the links.
- Durability defines the time period by which the data is made available on the website. Historical data may or may not be available on the website. So, the volatile data must be extracted and warehoused for the purpose of availability [20], [25].
- Accessibility checks whether the data has been accessed without breaching any authenticity norms (registration or password) during the automatic extraction for warehousing [20, 25].
- Refresh rate defines the timeliness by which the data is made available on the website, at the meantime fast refresh rate means volatile data is overwritten quickly, so must be extracted with the same rate to make it available for data analytics [20], [25], [28].

B. Web Data Quality

This selection feature can be further split into Origination, Objectivity, Accurateness, Completeness, and Metadata. Origination usually refers to data lineage, i.e. origin of the data. Objectivity concerns with deficiency of biasness in the data. Accurateness concerns with the accuracy of web data, i.e. error free data. Completeness concerns with the coverage, whereas Metadata concerns with the derivation rules and interpretation of web data [20], [25], [28], [30].

C. Contextual Issues of Web Data

This feature can be further split into three sub-categories viz. Relevancy, Timeliness, Layout. Relevancy is the most important feature to select the web source, as how much the specific data is relevant for data analytics. Timeliness concerns with how timely the data is made available on the website. Layout defines different formats of data presentation like XML, HTML, pdf, docs, pictures, audio, video or any other representation [20], [25].

IV. Evaluation and Selection of Web Source Using MCDM Methods (SAW and TOPSIS)

Zhu et al. [25] proposed four approaches to select the Web sources in compensatory methods viz. Simple Additive Weighing (SAW), Analytic Hierarchy Process (AHP), Data Envelopment Analysis (DEA) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Here SAW and AHP come under the scoring group, DEA under the concordance group while TOPSIS comes under the compromising group [13]. This section presents SAW and TOPSIS methods to statically analyze the source selection.

A. Simple Additive Weighing (SAW) Method

In this method for every feature of the web sources, some weight has been provided with the constraint that the sum of the weights of all the features must be 1. For example, four web sources *WS*1, *WS*2, *WS*3 and *WS*4, and twelve features have been assumed as shown in the following Table I.

TABLE I. Weights of Quality Features

Feature Symbol	Features	Weight
F1	Availability	0.07
F2	Durability	0.08
F3	Accessibility	0.09
F4	Refreshing Rates	0.07
F5	Origination	0.10
F6	Objectivity	0.07
F7	Accurateness	0.11
F8	Completeness	0.06
F9	Metadata	0.08
F10	Relevancy	0.10
F11	Timeliness	0.08
F12	Layout	0.09

In the SAW method, no standard scale has been defined for rating i.e. for giving a score, so it is defined by a decision maker. In this example, the minimum and maximum scale for score have been taken 1 and 9 respectively. Table II shows the performance score of the different web sources with respect to each feature.

TABLE II. Scores of the Different Web Sources With Regard to Each Feature

FS	WS1	WS2	WS3	WS4
F1	8	9	7	4
F2	6	1	9	8
F3	8	3	6	1
F4	4	3	2	2
F5	4	1	6	5
F6	7	7	6	1
F7	1	3	5	6
F8	4	8	7	9
F9	5	3	1	6
F10	8	4	1	3
F11	2	3	4	5
F12	5	8	5	4

Then

$$SAWi = \sum_{j=1}^{N} c_{ij} w_j; i = 1,2,\dots M \quad (1)$$

Where SAW_i: the SAW score of i^{th} web source; M: number of web sources; N: number of features; C_{ij}: score of i^{th} source in j^{th} feature; w_j: weight of j^{th} feature [4], [13], [15], [21]. Applying formula given in Eq. (1) to Table II, we find ranking score as SAW(W S1) = 5.09, SAW(W S2) = 4.19, SAW(W S3) = 4.83 and SAW(W S4) = 4.91. Here, Web Source WS1 is the best source for warehousing.

B. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) Method

This method was formulated by Hwang and Yoon as mentioned in the research article of Zhu et al. [25]. The fundamental approach of this method is to get an alternate solution in multi-dimensional computational area, such as; the solution is nearest to the ideal solution and farthest to the negative solution. The multi-dimensional computational area is defined by taking set of features as dimensions. Here the ideal solution is the positive extreme solution with a set of possible best synthetically scores with regard to each feature. Similarly, the negative ideal solution is the negative extreme solution with a set of possible worst scores. These two (ideal and negative ideal) solutions in computing area, are two points with extreme values as dimensions. This method has five steps to evaluate the best source [4], [12], [13], [21], [25]. They, with explanations taking the aforementioned example, are as follows:

1. Normalize the decision matrix.

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{M} x_{ij}^2}} \quad (2)$$

Where X_{ij} is the performance score of i^{th} Web Source in terms of j^{th} feature; M is the number of Web Sources. The values of the result are shown in Table III.

TABLE III. Normalized Decision Matrix

FS	WS1	WS2	WS3	WS4
F1	0.5521	0.6211	0.4830	0.2760
F2	0.4447	0.0741	0.6671	0.5930
F3	0.7628	0.2860	0.5721	0.0953
F4	0.6963	0.5222	0.3482	0.3482
F5	0.4529	0.1132	0.6794	0.5661
F6	0.6025	0.6025	0.5164	0.0861
F7	0.1187	0.3560	0.5934	0.7121
F8	0.2760	0.5521	0.4830	0.6211
F9	0.5394	0.3560	0.1187	0.7121
F10	0.8433	0.4216	0.1054	0.3162
F11	0.2722	0.4082	0.5443	0.6804
F12	0.3581	0.5729	0.3581	0.6445

2. Construct the weighted normalized decision Matrix.

$$WY = w_i y_{ij} \quad (3)$$

Where w_j is the weight of j^{th} feature (refer to Table I). The values of resultant matrix are shown in the Table IV.

TABLE IV. Weighted Normalized Matrix

FS	WS1	WS2	WS3	WS4
F1	0.0386	0.0435	0.0338	0.0193
F2	0.0356	0.0059	0.0534	0.0474
F3	0.0686	0.0257	0.0515	0.0086
F4	0.0487	0.0366	0.0244	0.0244
F5	0.0453	0.0113	0.0679	0.0566
F6	0.0422	0.0422	0.0361	0.0060
F7	0.0131	0.0392	0.0653	0.0783
F8	0.0166	0.0331	0.0290	0.0373
F9	0.0475	0.0285	0.0095	0.0570
F10	0.0843	0.0422	0.0105	0.0316
F11	0.0218	0.0327	0.0435	0.0544
F12	0.0322	0.0516	0.0311	0.0580

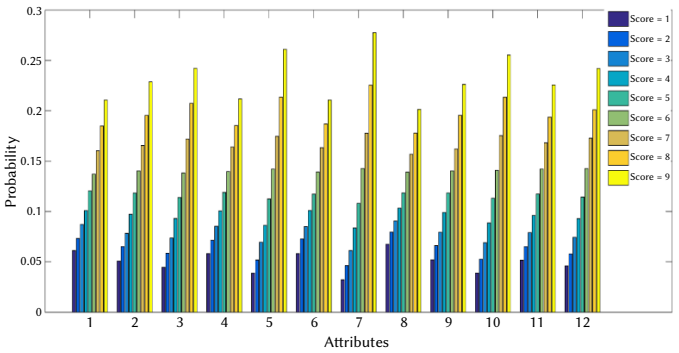

Fig. 1. Probability of selection with respective scores of each feature: SAW method.

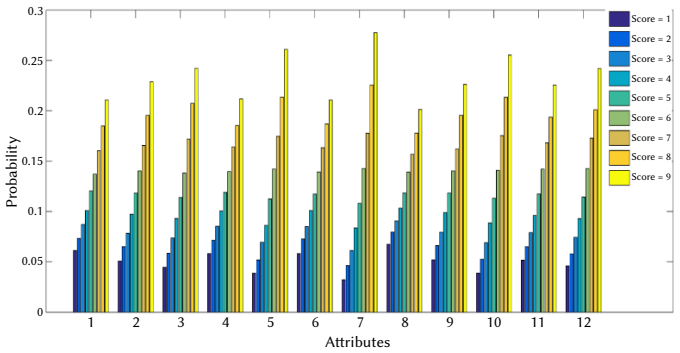

Fig. 2. Probability of selection with respective scores of each feature: TOPSIS method.

TABLE V. Probability of Selection With Respective Scores of Each Feature: SAW Method

Score	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
1	0.0193	0.0144	0.0090	0.0186	0.0059	0.0192	0.0035	0.0258	0.0137	0.0059	0.0136	0.0091
2	0.0278	0.0205	0.0149	0.0285	0.0113	0.0276	0.0077	0.0366	0.0207	0.0102	0.0211	0.0150
3	0.0408	0.0328	0.0247	0.0406	0.0202	0.0406	0.0151	0.0496	0.0334	0.0202	0.0334	0.0265
4	0.0579	0.0504	0.0424	0.0587	0.0347	0.0589	0.0292	0.0690	0.0502	0.0343	0.0497	0.0435
5	0.0820	0.0749	0.0667	0.0837	0.0582	0.0814	0.0522	0.0882	0.0769	0.0597	0.0770	0.0681
6	0.1142	0.1102	0.1040	0.1134	0.0964	0.1160	0.0914	0.1193	0.1097	0.0982	0.1100	0.1053
7	0.1589	0.1591	0.1595	0.1582	0.1547	0.1582	0.1525	0.1527	0.1591	0.1590	0.1602	0.1565
8	0.2126	0.2262	0.2355	0.2131	0.2461	0.2141	0.2494	0.2026	0.2256	0.2413	0.2216	0.2350
9	0.2864	0.3115	0.3434	0.2853	0.3726	0.2841	0.3989	0.2562	0.3108	0.3711	0.3134	0.3411

3. Fix the positive extreme and negative extreme solutions.

   Positive extreme solution: $PESj = max(w_j y_{ij})$ (4)

   Negative extreme solution: $NESj = min(w_j y_{ij})$ (5)

   where i = 1, 2, ...

   PES = (0.0435, 0.0534, 0.0686, 0.0487, 0.0679, 0.0422, 0.0783, 0.0373, 0.0570, 0.0843, 0.0544, 0.0580); and NES = (0.0193, 0.0059, 0.0086, 0.0244, 0.0113, 0.0060, 0.0131, 0.0166, 0.0095, 0.0105, 0.0218, 0.0322);

4. Determine the Euclidean distance of both virtual solutions.

   $$DPES_i = \sqrt{\sum_{j=1}^{N}(PES_j - w_j y_{ij})^2} \quad (6)$$

   $$DNES_i = \sqrt{\sum_{j=1}^{N}(w_j y_{ij} - NES_j)^2} \quad (7)$$

   In current example it takes the values (DP ESW S1 = 0.0858, DP ESW S2 = 0.1100, DP ESW S3 = 0.0987, DP ESW S4 = 0.0950) and (DNESW S1 = 0.1217, DNESW S2 = 0.0717, DNESW S3 = 0.1085, DNESW S4 = 0.1135).

5. Compute the relative closeness for the ideal solution.

   $$C_i = \frac{DNES_i}{DPES_i + DNES_i}; 0 \le C_i \le 1 \quad (8)$$

   and here the measure of relative closeness are found as: CW S1 = 0.5864, CW S1 = 0.3946, CW S1 = 0.5237, CW S1 = 0.5444.

Thus, Web Source WS1 is the best source for warehousing.

V. Proposed Work

In this article the proposed work consists of three parts. In first part statistical analysis of SAW and TOPSIS has been performed and the probability of selection with respective scores of each feature has been determined. In the second part, the probability of selection of a new web source has been determined using the probability of the respective scores of features. In the third part, improvised method (MSD method) has been proposed which is more efficient to handle the dynamic and complex behavior of the web.

A. Statistical Analysis of SAW and TOPSIS

In the statistical analysis, we have entertained all the twelve features [5], [10], [31], [32]. After execution of the Matlab implementation of both methods (SAW & TOPSIS) repeatedly around 105 times, we have selected 105 web sources, every time the best one out of 500 random sources. After that, the probability of selection of web sources respective to each score (1 to 9) for each feature has been determined using histogram methodology [19], [24]. The calculated value of the probabilities for both methods is shown in Table V and Table VI. The pictorial representation of the probability of selection with respective scores of each feature in both methods are illustrated in Fig. 1 and Fig. 2. As the figures show in both methods, as the score of the feature increases the probability of selection also increases irrespective of the weight.

Now, we are calculating the mean score and the standard deviation of score of each feature of the selected web sources for both the methods by employing the formulae:

Mean: $M_{score}(i) = \sum_{j=1}^{9} p(WS_i(j))WS_i(j)$ (9)

Std. Deviation: $DS_{score}(i) = \sqrt{\sum_{j=1}^{9} p(WS_i(j))(WS_i(j))^2}$ (10)

The values of mean score and standard deviation [19], [24] of score are shown in Table VII and Table VIII respectively. The pictorial representation of mean score and standard deviation of the score, as shown in the Fig. 3 and Fig. 4, illustrate that there is minor variation in the mean score of respective features of the selected web sources while there is a significant variation in the standard deviation of respective features of the selected web sources for both methods. As Fig. 4 shows the value of standard deviation in SAW method is greater than what we get in the TOPSIS method. It depicts that TOPSIS method is more efficient than SAW method while selecting the web sources.

TABLE VI. PROBABILITY OF SELECTION WITH RESPECTIVE SCORES OF EACH FEATURE: TOPSIS METHOD

Score	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
1	**0.0135**	0.0055	0.0014	0.0141	0.0002	0.0137	0.0000	0.0273	0.0052	0.0003	0.0054	0.0014
2	0.0257	0.0130	0.0055	0.0257	0.0017	0.0266	0.0003	0.0430	**0.0135**	0.0016	0.0140	0.0053
3	0.0451	0.0289	0.0158	0.0469	0.0077	0.0462	**0.0028**	0.0610	0.0290	0.0073	0.0298	0.0166
4	0.0731	0.0556	0.0393	0.0730	0.0245	0.0713	0.0138	0.0844	0.0560	0.0253	0.0565	**0.0384**
5	0.1002	0.0911	**0.0779**	0.1014	0.0587	0.1018	0.0425	0.1110	0.0916	0.0597	**0.0896**	0.0760
6	0.1389	0.1348	0.1269	0.1377	0.1158	**0.1370**	0.1007	0.1355	0.1332	0.1162	0.1354	0.1262
7	0.1734	**0.1819**	0.1866	0.1726	0.1891	0.1713	0.1865	**0.1599**	0.1837	0.1881	0.1815	0.1884
8	0.2027	0.2265	0.2452	**0.2038**	0.2666	0.2055	0.2849	0.1824	0.2262	0.2680	0.2277	0.2495
9	0.2274	0.2628	0.3014	0.2250	**0.3358**	0.2266	0.3686	0.1954	0.2616	**0.3353**	0.2602	0.2982

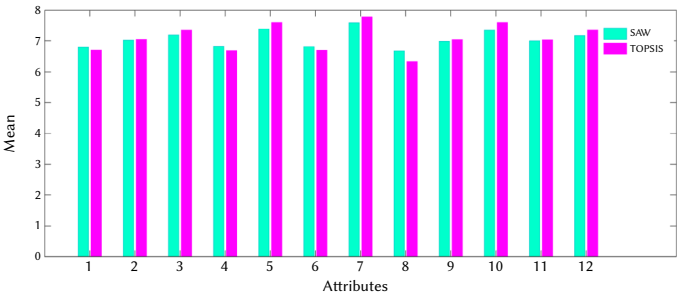

Fig. 3. Mean of the scores of features of selected resources.

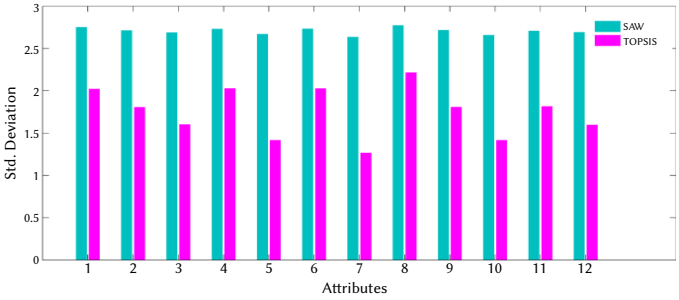

Fig. 4. Standard deviation of the scores of features of selected resources.

TABLE VII. MEAN SCORE OF SELECTED WEB SOURCES FEATURES

Feature	SAW Method	TOPSIS Method	Average
F1	6.9152	6.7088	6.8120
F2	7.1176	7.0549	7.0862
F3	7.3302	7.3479	7.3390
F4	6.9106	6.6940	6.8023
F5	7.5019	7.5914	7.5467
F6	6.9118	6.7043	6.8081
F7	7.6434	7.7822	7.7128
F8	6.6763	6.3394	6.5078
F9	7.1139	7.0502	7.0820
F10	7.4954	7.5920	7.5437
F11	7.1146	7.0426	7.0786
F12	7.3102	7.3509	7.3305

TABLE VIII. STANDARD DEVIATION OF SCORE OF SELECTED WEB SOURCES FEATURES

Feature	SAW Method	TOPSIS Method	Average
F1	2.0876	2.0190	2.0533
F2	1.9693	1.8050	1.8871
F3	1.8289	1.6014	1.1715
F4	2.0863	2.0272	2.0567
F5	1.7157	1.4148	1.5653
F6	2.0839	2.0265	2.0552
F7	1.5990	1.2630	1.4310
F8	2.2028	2.2141	2.2084
F9	1.9672	1.8063	1.8836
F10	1.9694	1.8136	1.8915
F11	1.8430	1.5952	1.7191
F12	1.3420	1.7920	1.8121

B. Probability of Selection of a New Web-Source

As there is a large number of web sources available on web as well as an exponential growth of web sources, an efficient methodology to select the web sources for web warehousing is required. One way is to calculate the probability of selection of each new web source to evaluate the relevance of specific web source. In order to calculate the probability of selection of a web source by any comparative method (like SAW and TOPSIS methods), the probability of respective score and weight of each feature are required and the formula to calculate the probability is as follows [1]:

$$p_s(WS) = \sum_{i=1}^{n} W_i p_i(WS_i(j)) \quad (11)$$

Where

$$\sum_{i=1}^{n} W_i = 1 \quad (12)$$

$$\sum_{i=1}^{m} p_i(WS_i(j)) = 1 \quad (13)$$

Here, n is the number of features of the web source (WS) and (1, 2... m) are the scores of the features. $p_i\,(WS_i\,(j))$ is the probability of selection of i^{th} feature having score value j. In this article n = 12 and m = 9.

A higher value of the probability shows the specific web source is more relevant, so it is recommendable to be used as data source for warehouse. Here the individual web source relevancy can be assessed by probability without any comparison.

For example, if a new web source (WS) has the score of all twelve features as {1, 7, 5, 8, 9, 6, 3, 7, 2, 9, 5, 4} and the weight of each feature is as mentioned in Table I, then its probability of getting selected (in TOPSIS method) is as follows:

Here, WS(1) = 1, WS(2) = 7, WS(3) = 5, WS(4) = 8, WS(5) = 9, WS(6) = 6, WS(7) = 3, WS(8) = 7, WS(9) = 2, WS(10) = 9, WS(11) = 5 and W S(12) = 4. The probability of selection of the respective feature with respect to the score in TOPSIS method is given in TABLE VI (highlighted in bold) which are:

p1 (WS (1)) = 0.0135, p2 (WS (2)) = 0.1819, p3 (WS (3)) = 0.0779, p4 (WS (4)) = 0.2038, p5 (WS (5)) = 0.3358, p6 (WS (6)) = 0.1370,

p7 (WS (7)) = 0.0028, p8 (WS (8)) = 0.1599, p9 (WS (9)) = 0.0135, p10 (WS (10)) = 0.3353, p11 (WS (11)) = 0.0896, p12 (WS (12)) = 0.0384.

Now, by applying the formula (11), the probability of selection of the web source (WS) is **0.1351**.

Similarly, the probability of selection in the TOPSIS method can also be calculated.

C. The MSD Method

The proposed MSD method is an enhancement of the already defined methods to handle the complex and dynamic nature of the web. It is based on probabilistic analysis of MCDM methods. The proposed method consists of two parts: (i) fixing the standard score *Sscore* of each feature for selection, and (ii) checking the suitability of the coming web sources. The steps of the method have been elaborated in the following algorithm.

Algorithm:

(i) Fixing the standard score:

1. Determine the mean score (Mscore(i)) and standard deviation of score (Sscore(i)) of i^{th} feature from selected web sources, where i = {1, 2, ..., m}.

2. Determine the standard score (Sscore(i)) for i^{th} feature using formula:

$$\text{Sscore}(i) = \text{Mscore}(i) - \text{SDscore}(i) \quad (14)$$

3. Set the Sscore(i) as the selection parameter for i^{th} feature.

(ii) Check the suitability:

1. Set the threshold value Th: (where Th ≤ m).

2. For each feature of a web source (WS) calculate:

if $\text{WS}(i) \geq \text{Sscore}(i)$

$\text{Suitability}(i) = 1;$

otherwise,

$\text{Suitability}(i) = 0;$

3. If $\sum_{i=1}^{m} \text{Suitability}\,(i) \geq \text{Th}$, then the web source is suitable to select, otherwise rejected.

Here the standard score of each feature is derived from the mean score and the standard deviation of the scores, using the probability values of the respective score of each features employing SAW and TOPSIS methods. Now the standard score of the respective feature is used as parameter to check the suitability of the web source with respect to that feature. If a new web source has the number of features (whose score is greater than the respective standard score) more than the threshold value, then the web source is selected otherwise rejected. Here the threshold value is only to check the number of features whose value is more than the standard score.

VI. Experimental Setup and Result Analysis

For the implementation of all the three methods (SAW, TOPSIS and MSD), we have used Matlab12a, Windows 8 (64 bit Operating System), Intel CITM) i3-4005U CPU @ 1.70 GHz. In order to determine the standard score (*Sscore*), we calculate the average of the mean scores and average of the standard deviation of the scores of each feature of the selected web sources employing SAW and TOPSIS methods, and results are shown in the TABLE VII and TABLE VIII respectively. Using the aforementioned algorithm of MSD method, the worthy web sources have been selected as shown in TABLE IX, here NFS stands for 'None Found Suitable'. For implementation and analysis, we have taken fourteen data-sets and each data-set consists of twenty randomly generated web sources with some score value for each feature. All these data-sets are given in the Appendix I.

The results and comparative analysis of all the three methods as shown in Table IX, show the effectiveness of the proposed MSD method while dealing with the complex and dynamic nature of web. The MSD method also shows improvisation during selection of web sources in comparison with SAW and TOPSIS methods in the following way:

- MSD method assures the suitability of web sources individually, whereas SAW and TOPSIS methods find the best one, on relative comparison basis.
- SAW and TOPSIS methods will select available single web source by default without any evaluation. However, the MSD method either selects or rejects depending on whether the threshold value is met or not.
- When the data-set consists of worthy web sources, the MSD method either agrees or disagrees with the SAW and TOPSIS methods due to the involvement of weight of features, as shown in Table IX for Data-sets 1, 2, 3, 4, 5, 6 and for Data-sets 7, 8, 11 respectively.
- SAW and TOPSIS methods usually select one web source (the best one) while the MSD method may select more than one suitable web sources in a single execution as shown in TABLE IX for Data-sets 3, 4, 5, 6, 10, 13 and 14. So it is effective to handle the dynamics of web.
- If all the new web sources are bad, both SAW and TOPSIS methods will select the best one from all the bad, but the MSD method will reject all of them as shown in in TABLE IX for Data-sets 9 and 12.

TABLE IX. Selected Web Sources Features

Data Set	SAW Method	TOPSIS Method	MSD Method
1	7	7	7
2	9	9	9
3	4	4	4, 8
4	6	6	6, 15
5	16	15	15, 16
6	5	5	5, 10, 20
7	4	3	14
8	17	17	15
9	16	16	NFS
10	10	8	6, 9, 12
11	11	16	13
12	10	8	NFS
13	15	17	11, 17, 19
14	9	10	9, 17

VII. Conclusion

In this article, statistical analysis has been performed on SAW and TOPSIS methods to study the behavior of both methods and also propose an efficient method based on this statistical study. In statistical analysis, the probability of the scores of each feature in both methods enforces that, as the value of the probabilities increases, the chance of selection increases. Using these probability values of the score of the features, the probability of selection of a new web source can be calculated by eq. (11). Furthermore, the mean of the score of a feature in both methods is almost the same but there is significant variation in standard deviation of the scores of the respective features. It shows the TOPSIS method is more effective than SAW to select the web sources. SAW and TOPSIS methods always yield the best one among all the available web sources on comparative basis without checking the quality of web sources, while the MSD method deals individually with each web source and assures its quality while selecting. So, if there is single web source, it is selected by default by both the methods because there is no other source to compare, but not in the MSD method.

Once after the computation of Mean Score and Standard Deviation Score, there is no further comparisons of feature score as in SAW and TOPSIS methods for selection of web-sources. So, the proposed method is more efficient in selection of web-sources where the data is updated frequently.

The proposed MSD method is only based on standard scores of each feature so gets rid from manually/randomly fixing weight of features as well as comparison among the web sources. Thus checking the quality of web sources individually in the MSD method makes it more efficient to deal with the dynamic and complex nature of web. Moreover, the computation cost in both methods is always higher than the MSD method due to the involvement of comparison operations to select the best one but it is linear for the proposed method and it will based on the number of evaluation features of the web-sources. If the number of the features are n then the computational cost is $O(n)$.

VIII. Future Work

A lot of enhancement is still required to design the effective web warehouse. Further research is needed to analyze the sensitiveness of the selected web sources when various critical factors are changed simultaneously. The MCDM approaches for selecting suitable web-data sources have a number of methods to evaluate the suitability. The proposed work is based on the aggregated study of SAW and TOPSIS methods. Various other MCDM methods like TOPSIS-COMET, COCOSO, and MABRAC [34] may be statistically investigated and incorporated with the proposed MSD method to improve the suitability of the selection of web-data sources for the data-warehouse storage. Moreover, the contents over the web sources change randomly and dynamically. So our focus in the future is to identify the updated relevant data over the selected web sources with minimum latency in order to update the web warehouse.

Appendix

Data Set: 01

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	7	6	8	7	8	1	4	1	1	9	7	6
WS2	5	3	8	9	9	8	2	3	4	9	5	6
WS3	1	3	3	4	5	4	5	3	2	4	7	5
WS4	3	2	9	4	5	6	1	9	8	4	8	1
WS5	4	4	6	2	9	9	8	9	7	7	3	5
WS6	2	9	1	4	6	3	3	1	7	7	9	1
WS7	1	8	4	5	7	9	7	6	9	8	6	8
WS8	5	4	5	5	6	8	8	1	8	6	3	6
WS9	4	7	6	2	5	8	7	4	5	1	3	7
WS10	8	3	5	2	2	6	6	5	1	6	9	6
WS11	5	9	8	4	6	5	3	8	1	1	4	6
WS12	8	6	2	8	2	4	5	6	5	7	2	9
WS13	5	2	9	2	2	8	9	3	9	6	9	2
WS14	1	1	8	5	8	6	3	9	8	6	6	5
WS15	1	6	4	2	2	8	2	7	9	3	9	8
WS16	2	3	6	8	6	9	3	6	9	7	2	5
WS17	4	1	3	7	4	8	7	9	4	7	8	7
WS18	7	2	1	6	6	8	3	4	3	4	5	2
WS19	2	1	5	4	7	5	6	4	8	1	6	7
WS20	3	7	8	5	3	8	2	6	2	8	1	9

Data Set: 02

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	7	6	5	7	8	4	3	2	7	6	6	1
WS2	1	6	3	5	1	3	7	1	8	4	9	1
WS3	4	9	7	6	7	6	3	6	5	6	6	2
WS4	5	2	9	1	6	4	9	1	1	4	5	5
WS5	7	8	6	3	7	2	5	7	1	2	6	1
WS6	2	4	3	1	9	9	4	2	2	1	9	1
WS7	5	9	9	8	2	5	6	3	8	2	5	1
WS8	1	8	1	1	1	5	6	9	4	4	4	6
WS9	8	7	6	5	8	3	8	2	8	8	9	1
WS10	4	2	9	2	5	2	8	8	2	4	9	6
WS11	3	1	5	6	1	8	2	1	7	6	9	6
WS12	1	2	9	6	1	3	1	3	2	4	7	1
WS13	1	5	3	7	2	5	3	9	8	8	6	3
WS14	9	1	5	3	2	5	6	3	5	7	7	4
WS15	2	8	9	9	9	9	4	2	9	2	4	2
WS16	9	9	2	8	3	3	9	8	4	4	8	1
WS17	4	4	8	1	6	3	1	6	9	6	2	5
WS18	9	5	2	3	8	3	9	8	4	4	8	1
WS19	4	1	4	9	3	6	7	9	1	1	9	7
WS20	1	3	8	1	7	4	6	1	9	6	9	8

Data Set: 03

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	4	7	4	2	6	4	9	2	5	7	3	1
WS2	2	7	3	2	6	2	8	5	6	7	4	8
WS3	7	7	3	6	8	4	9	9	1	5	5	3
WS4	2	8	3	7	9	8	7	4	6	9	9	8
WS5	9	6	1	8	1	4	3	6	7	9	1	2
WS6	1	5	9	1	8	7	7	1	6	9	7	7
WS7	1	3	5	2	6	3	7	2	5	5	9	8
WS8	8	9	6	5	2	6	7	2	8	9	8	7
WS9	8	3	3	4	2	8	7	1	9	4	2	3
WS10	6	5	9	4	7	2	3	3	3	6	9	4
WS11	7	7	9	1	6	5	5	3	8	6	4	4
WS12	5	3	9	7	6	2	7	4	3	8	8	3
WS13	6	8	6	4	9	2	6	2	1	3	8	4
WS14	2	5	7	7	6	1	9	9	1	6	4	1
WS15	5	9	4	5	1	1	2	5	3	6	6	5
WS16	1	4	5	1	2	6	9	4	3	6	8	5
WS17	1	9	3	3	1	4	7	3	1	3	1	1
WS18	4	9	7	9	2	6	3	4	2	3	1	6
WS19	6	6	8	9	7	8	6	4	4	2	4	1
WS20	9	1	1	4	6	7	5	9	7	1	2	2

Data Set: 04

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	3	9	7	1	9	5	9	3	8	8	5	6
WS2	7	1	6	6	2	1	3	4	4	5	5	9
WS3	9	3	2	3	5	3	9	3	3	1	7	1
WS4	8	5	1	5	9	2	5	2	4	9	9	2
WS5	6	3	4	6	6	1	6	6	4	4	9	4
WS6	6	5	5	1	6	9	9	8	8	7	9	7
WS7	9	8	2	1	5	1	7	2	1	2	6	9
WS8	4	5	8	3	6	8	3	9	7	8	9	1
WS9	3	9	6	5	9	5	1	7	8	9	2	2
WS10	6	2	1	6	4	2	1	5	3	8	4	4
WS11	3	4	9	3	3	5	8	9	6	7	1	7
WS12	6	3	5	9	4	6	4	6	3	4	5	5
WS13	1	6	1	8	3	1	4	2	6	7	7	7
WS14	9	1	1	5	5	9	5	8	1	6	2	2
WS15	2	1	9	9	6	6	9	8	9	6	4	7
WS16	6	3	8	6	5	5	4	3	6	4	1	6
WS17	4	1	6	3	4	8	6	3	4	7	3	1
WS18	7	2	1	8	8	3	1	8	1	2	5	9
WS19	1	4	3	5	3	8	4	5	8	1	5	9
WS20	1	1	5	3	4	3	2	1	6	6	7	2

Data Set: 05

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	4	4	1	3	3	2	6	7	7	3	1	1
WS2	9	2	9	8	3	3	2	1	9	7	3	7
WS3	4	2	6	8	2	3	5	9	9	3	8	7
WS4	7	1	6	3	2	8	5	8	4	7	1	2
WS5	1	8	5	9	4	2	8	6	2	8	6	8
WS6	2	8	5	2	9	7	6	8	4	1	9	2
WS7	1	9	5	8	8	8	2	2	8	9	9	9
WS8	2	4	7	8	7	7	2	2	6	2	9	3
WS9	2	5	6	5	6	8	6	1	6	2	1	7
WS10	5	6	2	2	3	7	2	7	9	6	4	6
WS11	2	7	4	1	3	8	5	3	3	3	3	2
WS12	5	9	6	2	1	4	7	7	3	9	4	3
WS13	7	6	5	7	8	9	6	5	3	6	9	2
WS14	8	9	2	5	7	5	3	3	9	9	3	9
WS15	7	9	8	5	7	8	7	9	8	2	6	5
WS16	9	9	8	9	3	9	4	8	9	1	8	9
WS17	2	3	3	2	8	8	1	5	2	9	7	2
WS18	2	9	2	2	5	4	7	2	8	7	9	6
WS19	9	2	8	4	2	6	7	1	4	1	1	8
WS20	9	8	3	2	8	5	1	7	6	1	8	4

Data Set: 06

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	8	8	1	8	8	2	3	2	4	7	6	2
WS2	3	4	8	9	6	7	3	3	8	8	2	6
WS3	4	6	5	2	9	5	5	8	9	8	2	7
WS4	4	9	9	5	6	6	5	5	4	9	1	3
WS5	7	8	8	7	8	4	6	5	7	8	7	8
WS6	9	8	2	1	2	4	6	8	9	5	6	2
WS7	2	9	4	1	8	7	2	9	9	5	8	3
WS8	7	2	1	4	9	9	4	1	1	1	1	6
WS9	5	7	4	2	3	8	4	4	1	5	3	8
WS10	7	7	4	9	9	6	1	5	6	6	1	8
WS11	8	8	3	9	7	9	3	7	1	1	6	6
WS12	3	7	9	9	7	6	9	9	2	4	3	1
WS13	9	3	5	9	7	7	9	7	1	3	7	9
WS14	1	1	7	7	8	9	8	3	8	6	2	8
WS15	6	2	6	7	5	3	2	9	2	4	5	4
WS16	9	3	8	4	7	1	5	9	9	4	1	8
WS17	2	8	3	1	5	8	1	5	7	1	1	7
WS18	4	9	7	4	9	4	7	7	8	3	5	3
WS19	3	5	4	5	2	7	8	6	4	3	7	2
WS20	6	9	1	5	8	9	9	9	3	7	6	8

Data Set: 07

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	9	1	1	6	6	7	3	4	5	1	3	7
WS2	4	3	1	2	3	2	1	1	3	8	9	3
WS3	7	7	7	6	8	1	8	9	4	6	5	1
WS4	2	8	4	7	8	7	4	6	3	5	7	9
WS5	1	5	6	5	5	5	6	3	3	5	7	9
WS6	6	9	3	2	4	5	6	1	2	5	8	8
WS7	7	4	7	4	9	3	1	3	2	6	3	9
WS8	3	3	4	6	1	3	6	9	5	6	3	5
WS9	6	5	7	3	4	9	3	2	1	2	7	2
WS10	3	2	9	8	5	5	3	5	9	8	2	5
WS11	8	2	1	2	6	7	5	9	2	2	8	7
WS12	4	3	5	3	1	2	6	2	2	4	2	7
WS13	7	7	1	9	2	1	8	5	7	3	9	8
WS14	6	6	7	6	6	2	6	6	7	3	7	7
WS15	9	9	9	7	2	6	1	6	4	1	2	6
WS16	2	7	1	8	2	6	8	6	1	9	8	3
WS17	6	6	5	1	1	9	3	3	1	5	9	4
WS18	1	5	5	4	2	1	8	4	4	5	6	8
WS19	5	3	7	7	5	2	7	8	8	5	4	4
WS20	8	7	3	2	8	4	6	9	5	4	4	1

Data Set: 08

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	8	2	2	6	1	1	4	7	9	9	8	3
WS2	2	3	3	1	9	3	6	1	1	4	3	5
WS3	5	6	5	4	6	5	9	1	9	5	9	9
WS4	3	4	3	9	5	3	8	8	9	8	9	6
WS5	6	6	5	9	9	9	8	4	9	2	9	1
WS6	5	3	8	3	7	2	1	6	7	9	1	4
WS7	3	9	6	3	5	3	8	7	7	3	6	8
WS8	5	7	1	8	4	5	4	9	1	7	4	4
WS9	1	7	2	4	4	6	3	1	2	5	1	7
WS10	6	4	4	8	2	8	1	2	8	6	9	7
WS11	8	6	9	5	5	6	7	2	8	7	7	3
WS12	7	5	6	5	9	8	8	8	1	3	1	9
WS13	5	9	2	4	2	6	1	2	7	5	7	2
WS14	7	1	4	4	7	5	3	5	3	2	3	8
WS15	8	5	3	5	8	8	9	7	3	9	9	4
WS16	5	7	4	7	9	4	1	9	8	6	2	7
WS17	9	9	8	7	6	1	9	6	4	8	6	4
WS18	6	6	5	4	6	9	4	3	6	5	9	5
WS19	8	6	3	5	7	7	8	5	1	8	4	8
WS20	5	6	6	4	3	6	3	5	7	5	8	2

Data Set: 09

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	7	7	4	9	8	5	4	3	4	8	2	6
WS2	3	5	6	1	5	6	4	7	1	5	9	6
WS3	1	1	9	4	4	3	8	3	8	7	7	5
WS4	8	4	9	1	7	2	2	2	6	7	9	2
WS5	7	4	3	3	5	7	6	5	5	5	7	5
WS6	9	6	2	7	3	4	9	4	6	8	2	7
WS7	9	2	4	7	1	5	6	9	3	5	3	4
WS8	5	1	5	3	6	7	1	3	7	9	1	7
WS9	9	5	1	3	7	6	7	6	4	8	1	7
WS10	5	7	8	1	6	5	3	6	8	2	3	9
WS11	9	6	9	1	5	7	7	6	4	3	2	3
WS12	1	9	5	7	5	7	4	6	8	2	7	4
WS13	2	3	2	3	3	3	5	3	8	5	9	4
WS14	7	2	7	5	4	8	2	2	1	1	4	1
WS15	4	3	7	9	7	6	4	1	9	9	8	6
WS16	4	8	5	7	9	6	4	1	9	9	8	6
WS17	6	7	2	5	3	1	1	8	4	7	5	4
WS18	1	8	8	6	8	8	1	4	3	8	7	4
WS19	4	7	8	6	2	2	6	8	6	5	7	5
WS20	1	6	2	3	1	8	9	9	8	6	4	3

Data Set: 10

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	8	8	9	3	2	8	8	2	6	1	7	1
WS2	3	3	5	2	8	8	7	3	4	1	8	6
WS3	2	7	4	8	5	1	6	2	4	6	5	8
WS4	7	4	6	7	2	8	4	3	1	2	1	3
WS5	9	3	8	2	1	4	2	2	5	5	6	6
WS6	9	8	8	8	3	6	4	5	6	6	9	2
WS7	2	5	7	4	1	1	8	6	9	3	5	8
WS8	2	4	9	3	9	8	2	4	7	9	5	8
WS9	5	6	2	8	8	5	1	5	8	7	9	7
WS10	9	2	9	9	3	3	8	9	5	4	7	9
WS11	2	7	9	3	2	5	4	2	6	3	5	5
WS12	7	3	8	9	7	8	2	8	6	9	1	7
WS13	5	3	3	1	3	7	8	3	3	4	5	7
WS14	8	5	3	6	1	8	1	6	2	1	7	4
WS15	6	8	5	8	4	9	3	5	3	2	1	9
WS16	4	4	9	2	4	6	8	1	3	6	1	8
WS17	5	6	5	9	5	4	3	5	3	5	6	3
WS18	9	2	7	6	7	7	7	6	2	3	4	6
WS19	9	6	2	1	9	3	7	4	8	4	4	5
WS20	6	3	6	1	7	4	3	3	5	2	5	9

Data Set: 11

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	3	3	7	5	9	7	6	9	4	1	3	6
WS2	2	2	8	9	3	8	5	8	4	3	2	3
WS3	6	8	6	4	4	7	3	8	8	3	2	1
WS4	9	7	5	6	6	2	8	9	1	2	6	1
WS5	7	4	2	4	3	3	7	9	9	7	4	9
WS6	2	6	2	2	9	3	7	1	5	5	1	1
WS7	2	4	8	8	2	9	8	9	3	2	3	6
WS8	3	5	4	9	7	8	3	8	5	9	6	4
WS9	1	2	4	5	1	7	5	1	7	1	5	6
WS10	7	6	5	3	3	8	9	3	2	6	8	8
WS11	5	5	5	9	8	4	9	9	6	1	8	8
WS12	2	4	4	2	2	8	2	5	6	4	9	7
WS13	6	8	8	7	3	7	9	5	2	3	9	8
WS14	2	1	4	5	7	6	6	1	9	1	1	4
WS15	2	7	2	8	3	1	1	5	2	3	9	8
WS16	6	1	6	1	6	9	5	4	8	9	9	7
WS17	3	6	9	5	7	7	3	4	3	7	2	1
WS18	5	4	2	4	7	3	5	1	5	5	6	1
WS19	2	4	8	1	6	1	3	7	9	5	1	8
WS20	1	7	4	9	4	5	3	9	6	5	8	6

Data Set: 12

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	2	8	3	1	3	9	6	8	1	4	5	8
WS2	8	2	4	1	7	5	7	2	9	8	8	6
WS3	1	1	5	5	3	6	8	5	6	5	6	9
WS4	4	6	1	5	9	1	4	7	2	6	2	6
WS5	3	1	4	5	3	6	7	3	4	3	6	4
WS6	9	8	3	2	2	9	5	5	8	1	2	6
WS7	8	9	5	4	5	5	4	4	5	6	4	5
WS8	8	3	8	8	8	4	4	8	5	9	5	1
WS9	9	3	4	7	4	1	7	7	4	1	8	6
WS10	8	1	5	4	9	6	3	9	2	2	6	9
WS11	8	5	1	9	4	9	9	9	3	1	8	5
WS12	3	9	9	1	3	7	9	6	3	8	2	6
WS13	7	7	2	2	3	4	3	1	5	5	5	6
WS14	6	3	4	9	9	6	4	4	6	5	9	3
WS15	3	6	6	1	3	6	9	2	4	9	1	7
WS16	4	8	7	4	4	7	8	9	9	4	6	2
WS17	9	2	2	3	3	6	8	5	3	5	1	7
WS18	1	6	5	1	3	5	6	8	6	7	4	2
WS19	7	8	9	2	4	9	8	6	3	1	7	2
WS20	2	7	7	4	9	6	5	5	3	5	1	7

Data Set: 13

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	8	2	2	6	1	1	4	7	9	9	8	3
WS2	2	3	3	1	9	3	6	1	1	4	3	5
WS3	5	6	5	4	6	5	9	1	9	5	9	9
WS4	3	4	3	9	5	3	8	8	9	8	9	6
WS5	6	6	5	9	9	9	8	4	9	2	9	1
WS6	5	3	8	3	7	2	1	6	7	9	1	4
WS7	3	9	6	3	5	3	8	7	7	3	6	8
WS8	5	7	1	8	4	5	4	9	1	7	4	4
WS9	1	7	2	4	4	6	3	1	2	5	1	7
WS10	6	4	4	8	2	8	1	2	8	6	9	7
WS11	8	6	9	5	5	6	7	2	8	7	7	3
WS12	7	5	6	5	9	8	8	8	1	3	1	9
WS13	5	9	2	4	2	6	1	2	7	5	7	2
WS14	7	1	4	4	7	5	3	5	3	2	3	8
WS15	8	5	3	5	8	8	9	7	3	9	9	4
WS16	5	7	4	7	9	4	1	9	8	6	2	7
WS17	9	9	8	7	6	1	9	6	4	8	6	4
WS18	6	6	5	4	6	9	4	3	6	9	5	9
WS19	8	6	3	5	7	7	8	5	1	8	4	8
WS20	5	6	6	4	3	6	3	5	7	5	8	2

Data Set: 14

Features	1	2	3	4	5	6	7	8	9	10	11	12
WS1	3	9	2	5	5	6	2	9	8	1	6	1
WS2	6	7	1	9	8	1	8	4	9	5	6	8
WS3	2	3	2	2	6	4	1	3	4	2	3	5
WS4	9	7	1	9	3	4	2	5	1	5	8	7
WS5	7	3	3	4	5	7	2	2	1	6	4	3
WS6	8	6	1	5	9	4	9	4	8	1	9	1
WS7	5	3	7	4	8	5	5	8	9	8	2	2
WS8	9	5	9	2	8	6	1	3	7	1	8	6
WS9	8	7	2	8	9	4	8	7	9	6	8	1
WS10	2	7	6	2	7	8	9	6	5	5	5	9
WS11	9	4	4	8	1	5	5	8	5	6	7	9
WS12	7	2	2	4	9	8	4	5	6	8	8	4
WS13	3	1	2	4	5	3	6	4	9	2	3	9
WS14	8	1	2	5	5	8	3	6	9	2	7	6
WS15	9	9	9	4	1	2	7	6	1	2	9	8
WS16	7	6	5	1	8	2	9	5	6	8	9	4
WS17	6	6	8	6	7	1	2	7	7	7	8	1
WS18	2	1	2	3	5	2	6	4	2	9	5	5
WS19	9	2	3	6	7	8	6	2	1	2	1	3
WS20	8	1	8	4	7	9	6	4	5	4	1	9

References

[1] S. I. Amari, H. Nagaoka, and D. Harda, "Methods of information geometry. Translation of mathematical monographs," *Oxford University Press*, 2000. ISBN: 978-1-4704-4605-**5** https://bookstore.ams.org/mmono-191

[2] O. Boussaid, J. Darmont, F. Bentayeb, and S. Loudcher, "Warehousing complex data with from the web," *International Journal of Web Engineering and Technology*, vol. 4, no. 4, pp. 408-433, 2008. doi: 10.1504/IJWET.2008.019942.

[3] O. Boussaid, A. Tanasescu, F. Bentayeb, and J. Darmont, "Integration and dimensional modeling approaches for complex data warehousing," *Journal of Global Optimization*, vol. 37, pp. 571-591, 2007. https://doi.org/10.1007/s10898-006-9064-6.

[4] T. Y. Chen, "Comparative analysis of SAW and TOPSIS based on interval valued fuzzy sets: Discussion on score functions and weights constraints," *Expert Systems with Applications*, vol. 39, pp. 1848–1861, 2012. doi: 10.1016/j.eswa.2011.08.065.

[5] J. L. Devore, "Probability and statistics for engineering and the sciences," *Cengage Learning*, 2012. ISBN: 978-8131518397.

[6] A. Doan, A. Halevy, and Z. Ives, "Principles of data integration," *Elsevier*, 2012. ISBN: 978-0-12-416044-6.

[7] X. L. Dong, B. Saha, and D. Srivastava, "Less is more: Selecting sources wisely for integration," *Proceeding of the VLDB Endowment*, vol. 6, pp. 37-48, 2012. doi: https://doi.org/10.14778/2535568.2448938 .

[8] H. Fan, "Investigating a heterogeneous data integration approach for data warehousing," *PhD Thesis*, School of Computer Science & Information Systems, Birkbeck College, University of London, 2005. Accessed: Jan. 15, 2023. [Online]. Available: https://www.dcs.bbk.ac.uk/site/assets/files/1025/haofan.pdf

[9] R. D. Hackathorn, "Web framing for the data warehouse," *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, 1999. ISBN: 978-1558605039.

[10] J. L. Johnson, "Probability and statistics for computer science," *Wiley*, 2008. ISBN: 978-0470383421.

[11] D. Le, J. Rahayu, and E. Pardede, "Dynamic approach for integrating web data warehouses," *Computational Science and Its Applications*, ICCSA-2006, Springer, 2006. ISBN: 0302-9743.

[12] A. Marshall, "Principles of Economics," Prometheus Books, 1890. Accessed: Jan. 15, 2023. [Online]. Available: https://eet.pixel-online.org/files/etranslation/original/Marshall,%20Principles%20of%20Economics.pdf

[13] B. H. Massam, "Massam. Multi-criteria decision making (mcdm) techniques in planning," *Progress in planning*, vol. 30, no. 1, pp. 1–84, 1988.

[14] A. Mehedintu, I. Buligiu, and C. Pirvu, "Web-enabled data warehouse and data webhouse," *Revista Informatica Economica nr*, vol. 1, no. 45, pp. 96-102, 2008. https://core.ac.uk/download/pdf/6612753.pdf

[15] A. Memariani, A. Amini, and A. Alinezhad, "Sensitivity analysis of simple additive weighting method (saw): the results of change in the weight of one attribute on the final ranking of alternatives," *Journal of Industrial Engineering*, vol. 4, pp. 13-18, 2009.

[16] F. Naumann, "Data fusion and data quality," 1998.

[17] J. M. Perez, R. Berlanga, M. J. Aramburu, and T. B. Pedersen, "Integrating data warehouses with web data: A survey," *IEEE Transactions on Knowledge and Engineering*, vol. 20, no. 7, pp. 940-955, 2008. doi: 10.1109TKDE.2007.190746.

[18] S. Rizzi, A. Abello, J. Lechtenborger, and J. Trujillo, "Research in data warehouse modeling and design: dead or alive?" *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP* (DOLAP '06), pp. 3-10. *IEEE Computer Society*, 2006. doi: 10.1145/1183512.1183515.

[19] S. Ross, "Introduction to Probability Models," *Academic Press/Elsevier*, 2012. ISBN: 978-0-12-407948-9.

[20] R. Simanaviciene and L. Ustinovichius, "Quality-driven integration of heterogeneous information systems," *Informatik-Berichte*, vol. 117, pp. 1-21, 1999. https://www.vldb.org/conf/1999/P43.pdf

[21] R. Simanaviciene and L. Ustinovichius, "Sensitivity analysis for multiple criteria decision making methods: Topsis and saw," *Procedia Social and Behavioral Sciences*, vol. 2, pp. 7743-7744, 2010.

[22] X. Tan, D. C. Yen, and X. Fang, "Web warehousing: Web technology meets data warehousing," *Technology in Society*, vol. 25, no. 131-148, 2003.

[23] E. Triantaphyllou, B. Shu, S. Sanchez, and T. Ray, "Multi-criteria decision making: an operations research approach," *Encyclopedia of Electrical and Electronics Engineering*, vol. 15, pp. 175-186, 1998.

[24] K. S. Trivedi, "Probability and Statistics with Reliability, Queuing and Computer Science Applications," *Wiley*, 2013.

[25] Y. Zhu and A. Buchmann, "Evaluating and selecting web sources as external information resources of a data warehouse," *Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE202)*, pp. 140-160. IEEE Computer Society, 2002. doi: 10.1109/WISE.2002.1181652.

[26] G. Xu, "The Construction Site Management of Concrete Prefabricated Building by ISM-ANP Network Structure Model and BIM Under Big Data Text Mining," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 4, pp. 138-145, 2020. doi: 10.9781/ijimai.2020.11.013

[27] S. Kumar, V. K. Solanki, S. K. Choudhary, A. Selamat and R. G. Crespo, "Comparative Study on Ant Colony Optimization (ACO) and K-Means Clustering Approaches for Jobs Scheduling and Energy Optimization Model in Internet of Things (IoT)," International Journal of Interactive Multimedia and Artificial Intelligence (Special Issues on Soft Computing), vol. 6, no. 1, pp. 107-116, 2020. doi: 10.9781/ijimai.2020.01.003.

[28] S. Zhang, L. Genga, H. Yan, H. Nie, X. Lu and U. Kaymak, "Towards Multi-perspective Conference Checking with Fuzzy Sets," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 5, pp. 134-141, 2021. doi: 10.9781/ijimai.2021.02.013.

[29] Y. Wu, L. Zhang, G. Ding, T. Xue and F. Zhang, "Modeling of Performance Creative Evaluation Driven by Multimodal Affective Data," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 7, pp. 90-100, 2021. doi: 10.9781/ijimai.2021.08.005.

[30] D. Burgos, "Ritual and Data Analytics: A Mixed-Methods Model to Process Personal Belief," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 1, pp. 52-61, 2021. doi: 10.9781/ijimai.2021.07.002.

[31] S. K. Choudhary, K. Singh and V. K. Solanki, "Spiking Activity of LIF Neuron in Distributed Delay Framework," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 7, pp. 70-76, 2016. doi: 10.9781/ijimai.2016.3710 .

[32] I. Lopez-Plata, C. Exposito-Izquierdo, E. Lalla-Ruiz, B. Melian-Batista, J. Marcos-Vega, "A Greedy Randomized Adaptive Search With Probabilistic Learning for solving the Uncapacitated Plant Cycle Location Problem," International Journal of Interactive Multimedia and Artificial Intelligence, 2022, (In Press), doi: 10.9781/ijimai.2022.04.003.

[33] N.S. Houari & N. Taghezout, "An Efficient Tool for the Experts' Recommendation Based on PROMETHEE II and Negotiation: Application to the Industrial Maintenance," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 6, pp.67-77, 2021, doi: 10.9781/ijimai.2021.01.002.

[34] A. Baczkiewicz, B. Kizielewicz, A. Shekhovtsov, J. Watrobski & W. Salabun, "Methodical Aspects of MCDM Based E-Commerce Recommender System," Journal of Theoretical and Applied Electronic Commerce Research, vol. 192, pp. 4991-5002, 2021. doi: https://doi.org/10.1016/j.procs.2021.09.277.

Hariom Sharan Sinha

Hariom Sharan Sinha has obtained M.Tech, PhD (CSE) from JNU New Delhi. He is working as an associate professor in the Department of Computer Science and Engineering at Adamas University, Barasat, Kolkata, West Bengal, India. He has more than nine years of experience in teaching and eleven years of experience in research.

Saket Kumar Choudhary

Saket Kumar Choudhary is an assistant professor in CSE, GITAM University, Bengaluru, Karnatka, India. He has obtained his master degrees in Mathematics from the University of Allahabad, Allahabad, India in 2005, Master of Computer Application (MCA) from UPTU, Lucknow, India in 2010, Master of Technology (M.Tech) from Jawaharlal Nehru University, New Delhi, India in 2014. He is Ph.D (Computer Science and Technology) School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2017. His research interest includes mathematical modeling and simulation, dynamical systems, computational neuroscience: modeling of single and coupled neurons, computer vision, digital image processing, machine learning and artificial intelligence.

Vijendra Kumar Solanki

Vijender Kumar Solanki, Ph.D., is an Associate Professor in Department of Computer Science & Engineering, CMR Institute of Technology (Autonomous), Hyderabad, TS, India. He has more than 15 years of academic experience in network security, IoT, Big Data, Smart City and IT. Prior to his current role, he was associated with Apeejay Institute of Technology, Greater Noida, UP, KSRCE (Autonomous) Institution, Tamilnadu, India & Institute of Technology & Science, Ghaziabad, UP, India. He has attended an orientation program at UGC-Academic Staff College, University of Kerala, Thiruvananthapuram, Kerala &Refresher course at Indian Institute of Information Technology, Allahabad, UP, India. He has authored or co-authored more than 75 research articles that are published in journals, books and conference proceedings. He has edited or co-edited 12 books in the area of Information Technology. He teaches graduate & post graduate level courses in IT at ITS. He received Ph.D in Computer Science and Engineering from Anna University, Chennai, India in 2017 and ME, MCA from Maharishi Dayanand University, Rohtak, Haryana, India in 2007 and 2004, respectively and a bachelor's degree in Science from JLN Government College, Faridabad Haryana, India in 2001.He is Editor in International Journal of Machine Learning and Networked Collaborative Engineering (IJMLNCE) ISSN 2581-3242, Associate Editor in International Journal of Information Retrieval Research (IJIRR), IGI-GLOBAL, USA,ISSN: 2155-6377 | E-ISSN: 2155-6385 also serving editorial board members with many reputed journals. He has guest edited many volumes, with IGI-Global, USA, InderScience & many more reputed publishers.