Indian Journal of Science and Technology, *Vol 9(20), DOI: 10.17485/ijst/2016/v9i20/91962, May 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Improvised TOPSIS Approach to Select Web Source as External Data Source for Web Warehousing

Hari Om Sharan Sinha*

SC & SS, Jawaharlal Nehru University, New Delhi - 110067, India; hariom.sinha@gmail.com

Abstract

Objective: The main objective of the paper to incorporate the external web-data efficiently to web-warehouse, as the evolution of web and the requisite of data analytics necessitate it for effective decision support system. **Methods/Statistical Analysis:** Since the data owned of any organization is insufficient for decision support system. Nevertheless dynamic and complex nature of web pose various challenges during selection of relevant web-data. So evaluation of web resources to select as external source for web-warehouse is the crucial phase during warehousing. Various Multi Criteria Decision Making (MCDM) approaches have been used for it. All these approaches evaluate the web resources on the basis of a set of features which define the relevancy of the resource. **Findings:** The main focus is on one of the approaches of MCDM viz. "Technique for Order Preference by Similarity to Ideal Solution" (TOPSIS) approach and also improvised the TOPSIS approach for efficient evaluation of the web resources. In traditional TOPSIS approach Euclidean distance has been measured to compute the proximity of real web-sources from Ideal web-sources. The Euclidean distance measure only the distances between the real and ideal web-resources but not the differences between them. In order to compute the differences between real and ideal web-resources Kullback-Leibler divergence method has been incorporated in the place of Euclidean distance method. **Application/Improvements:** The improvised TOPSIS computes symmetric as well as asymmetric distances to compute the differences, so efficient to compute the proximity in order to evaluation of web-resources.

Keywords: Improvised TOPSIS, Web-Data, Web-Warehouse, Web-Resources

1. Introduction

Nowadays web is prominent platform of both information sharing and retrieving. At the same time the data analytics compels the data warehouse to incorporate the web data for data analysis as the local data of specific organization is not sufficient for decision support system. As we know the data on the web is easily available and accessible but cannot be directly used efficiently for data analytics as done in conventional data warehouse[1-3]. The better solution is to club both the technologies for the data analytics as web technology provides enormous source of data and warehouse technology supports the data analysis. The data warehouse main task is accumulate the data for various sources and to design a repository with integrating the fetched data for data analysis. However the dynamic and complex nature of web as well as millions of resources available on web impose the constraints on conventional data warehouse while web data is used for warehousing.

For data analytics, it is more important task to find suitable data to incorporate consistently into warehouse[4-6]. In order to find suitable and consistent data for warehouse on web is like to search needle in a haystack as millions of web sources are usable on web[4,7]. Moreover the dynamic and complex nature of web data poses different challenges during web warehousing[4,6,8,9]. Thus for warehousing the foremost task is to ascertain the suitable web sources as

*Author for correspondence

data source for warehousing. For it, the relevancy of the web sources is evaluated on the basis of various features. Zhu et al. proposed three classes viz. web source stability, web data quality and contextual issues of web data[7] to categorize the features of web sources and also suggested Multi Criteria Decision Making (MCDM) approach[10] to evaluate the relevancy of web sources.

The first feature explains the challenges as, in addition to the numerous availability of web sources on web, the web sources have dynamic character i.e. web data changes frequently and even millions of new web sources are summing up routinely to web. Consequently present available web sources may alter or vanish[4,7] .

The second feature explains the quality of web data, as web is an open and independent platform. Thus a big amount of data available on web is not properly examined before sharing on web. So inconsistent, wrong, incomplete data, or ill structured data can be frequently envisioned on web[4,7] .

The third feature explains the context of data, as it also poses issues during warehousing of web data as data available on web is browsing centric rather than data analytics centric. Context of data imbibes not only the relevancy of data for warehousing but also the ease for extraction data and metadata like data definition, data derivation etc[4,7] .

So in order to design a web warehouse[3,4,6], a set of features of web sources must be built to evaluate their relevancy while selecting the web source as external data source for warehousing. Zhu et al. has suggested a set of features and also used MCDM approach[7,10] to evaluate the web sources to select as external source for warehouse. In this article we want to improvise the selection of web sources by including the Kullback Lieberal divergence[11,12] instead of Euclidian Distance measure in TOPSIS approach (One of MCDM approach) with respect to the evaluation features of web sources[13].

Rest of the paper is organized as: Section 2 explains the various features for evaluation of web sources. Section 3 elaborates MCDM approach and especially TOPSIS approach comprehensively. Section 4 presents the significance of Kullback-Leibler Divergence method. It helps in precise evaluation of web sources. Section 5 explicates the proposed work. Section 6 demonstrate the experimental setup and result analysis. The last section, Section 7 concludes the work.

2. Evaluation features of Web Sources

The features of evaluation is roughly categorized into three major groups: web source stability, web data quality and context of web data. These groups are further classified into subgroups to further refine the characteristics of web sources, as web source stability into Availability, Durability, Accessibility, Refreshing rate; web data quality into Origination, Objectivity, Accurateness, Completeness, Metadata; and context of web data into Relevancy, Timeliness, Layout[7]. Each feature has some weight and some performance score to evaluate the web sources. The sum of the weights of all the features is always equal to one i.e. $\sum_{j}^{n}=1\ Wj=1$ and there is no standard scale to assign the values to performance score. In this article twelve features have been incorporated with their "Weights" as shown in Table 1 and values of performance "Scores" have been taken between the ranges 1 to 9 randomly as shown in Table 2. In next section the TOPSIS (An MCDM approach) has been explained[7,10,14] comprehensively.

3. Multi Criteria Decision Making (MCDM) Approach

MCDM is a part of Operation Research discipline[10], in which the multiple criteria are being entertained explicitly for decision making. There are many conflicting

Table 1 Weight of Quality Features7

Feature Symbol	Features	Weight
F1	Availability	0.07
F2	Durability	0.08
F3	Accessibility	0.09
F4	Refreshing rate	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

Table 2 Score of Features[7]

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

Table 3 Decision Matrix[14]

Features WS	F1	F2	F3	.. ……..	… …….	… … ……	FN
1	X_{11}	X_{12}	X_{13}	…… ……	………..	…… …….	X_{1N}
2	X_{21}	X_{22}	X_{23}	…… ……	………..	…… …….	X_{2N}
3	X_{31}	X_{32}	X_{33}	…… ……	………..	…… …….	X_{3N}
.							
.							
M	X_{M1}	X_{M2}	X_{M3}	…… ……	………..	…… …….	X_{MN}

criteria that requires to evaluate for decision making to solve many real problems.

MCDM has two types of methods: Non Compensatory and Compensatory. Non Compensatory method does not allow tradeoff among attributes. An unsatisfactory value of one attribute cannot be counter balanced by promising values of other attributes[7,14]. Here each attribute has to qualify on its own basis. Whereas Compensatory methods allow tradeoff among attributes. The partial decrease in the value of one attribute is compensated by increase of the value of one or more attributes. Compensatory methods have been divided into four class of methods: **Scoring Methods, Compromising Methods, Concordance Methods and Evidential Reasoning Approach.** Simple Additive Weighting (SAW) and Analytic Hierarchy Process (AHP) **belong to Scoring Methods, TOPSIS** belongs to Compromising Methods and Data Envelopment Analysis (DEA) belongs to Concordance Methods. **Evidential Reasoning Approach** is latest development in MCDM approach different all the above three methods[10,14]. It uses extended decision matrix instead of decision matrix for MCDM approaches. Here our focus is on TOPSIS approach so we pass over all other approaches.

3.1 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) Approach

It is a MCDM approach, originally formulated by Hwang and Yoon in 1981 as a substitute of "*Elimination and Choice Translating Reality*" (ELECTRO) approach presented by Benayoun et al. in 1966. Further improvement was done by Yoon in 1987 and Hwang et al. in 1993. This approach is based on the concept, that the selected alternative solution should be closest to the Positive Ideal Solution and farthest to the Negative Ideal Solution. These two ideal solutions are extreme points in the computing space. In traditional TOSIS approach, the Euclidean distance measure is used it evaluate the relative proximity between alternate solutions and Positive Ideal Solution and preference order of alternative solutions is made on the basis of relative proximity[10,14]. This method consists of five steps, which is illustrated with following example. There is a decision matrix represented as in Table 3.

Here "WS" represents Web Source, "Fi" represents i^{th} feature of evaluation, "M" is the number of web sources "N" is the number of features and X_{ij} is the performance score of the j^{th} feature for i^{th} web source.

Table 4 Weighted Normalized Matrix[14]

Features WS	F1	F2	F3	.. ……..	… …….	… … ……	FN
1	W_1Y_{11}	W_2Y_{12}	W_3Y_{13}	…… ……	……….	…… …….	W_NY_{1N}
2	W_1Y_{21}	W_2Y_{22}	W_3Y_{23}	…… ……	……….	…… …….	W_NY_{2N}
3	W_1Y_{31}	W_2Y_{32}	W_3Y_{33}	…… ……	……….	…… …….	W_NY_{3N}
.							
.							
M	W_1Y_{M1}	W_2Y_{M2}	W_3Y_{M3}	…… ……	……….	…… …….	W_NY_{MN}

Step 1: Normalize the Decision Matrix:

$$Y_{ij} = \frac{X_{ij}}{\sqrt{\sum_1^M X_{ij}^2}} \qquad (1)$$

Where X_{ij} is the performance score of ith Web Source in terms of jth feature; M is the number of Web Sources and N is the number of features.

Step 2: Construct the weighted normalized decision Matrix:

$$\mathrm{WY} = \mathrm{W_j Y_{ij}} \qquad (2)$$

Where W_j is the weight of jth feature, such as $\sum W_j = 1$.

The resultant matrix of Weighted Normalized Matrix is as shown in Table 4.

Step 3: Fix the Positive Ideal and Negative Ideal solution

$$\text{Positive Ideal Solution} : \mathrm{PISj} = \max\left(\mathrm{W_j Y_{ij}}\right) \qquad (3)$$

$$\text{Negative Ideal Solution} : \mathrm{NISj} = \min\left(\mathrm{W_j Y_{ij}}\right) \qquad (4)$$

Step 4: Determine the distance measure, from alternative solutions to positive ideal solution as:

$$DPIS_i = \sqrt{\sum_1^N (PIS_j - W_i Y_{ij})^2} \qquad (5)$$

and from alternative solutions to negative ideal solution as:

$$DNIS_i = \sqrt{\sum_1^N (W_i Y_{ij} - PIS_j)} \qquad (6)$$

Step 5: Compute the relative proximity to the ideal solution[14]

$$P_i = \frac{DNIS_i}{DPIS_i + DNIS_i} \quad Where\ 0 \le P_i \le 1 \qquad (7)$$

Clearly $P_i = 1$ if $\mathrm{WS_i} = \mathrm{DPIS_i}$ and $P_i = 0$ if $\mathrm{WS_i} = \mathrm{DNIS_i}$. Larger the P_i value shows the WS_i is closer to PIS and farther to NIS. The corresponding web source (WS) having largest P_i values is the best solution[7,14].

In Traditional TOPSIS approach to compute the proximity, Euclidean distance[15] is measured between the alternative solutions and ideal solutions. It does not include the proximity of alternative solutions on the vertical line of two Ideal solutions, thus does not reflect the full merit of the alternative solutions in all other dimensions during evaluation. Euclidean Distance measure only computes linear distance between alternative and ideal solutions but not differences between two the two resources. Moreover random and dynamic nature of web sources also arise challenge in precise computation. So probabilistic approach is more effective for evaluation of web sources. So to improve the result of traditional TOPSIS the "Euclidean Distance measure" has been replaced by Kullback-Leibler divergence. Kullback-Leibler Divergence computes[11,12,15] the differences between two probability distributions[13,16].

4. Kullback Leibler Divergence (KLD)

It is asymmetric distance measure approach to compute the difference between the two probability distributions. For the probability distributions P and Q, where P represents the true distribution or precisely calculated distribution of data and Q represents the descriptive or approximate distribution of P. The difference degree of two random system of "n" dimensions having discrete probability distribution P and Q can be defined with KLD approach as[13,16-18]:

$$D = \sum_{i=1}^n P_i \log \frac{P_i}{Q_i} \qquad (8)$$

For continuous random systems having probability distribution P and Q the difference with KLD approach is defined as:

$$D = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}d(x) \qquad (9)$$

Where p(x) and q(x) represent the probability densities of probability distribution P and Q respectively[13,16-18].

5. Proposed Work

In this article we improvised the TOPSIS approach by replacing the "Euclidean Distance Measure" by "Kullback Leibler Divergence" method for evaluation of web sources for warehousing. For web sources evaluation during selection as external source for warehousing all the aforementioned features have been taken into account. As shown in the table 3 there are "M" web sources and "N" features for evaluation of web sources. Except step 4, all other steps of Improvised TOPSIS follow same way as follow in conventional TOPSIS approach. In step 4 to determine the distance the Euclidean distance method is replaced by the Kullback Leibler Divergence method as shown below[11,12].

Step 4: Determine the distance measure

From alternative solution to Positive Ideal Solution is as:

$$DPIS_i = \sum\nolimits_j^n = 1\ PIS_j\ \log\frac{PIS_j}{W_jY_{ij}} \qquad (10)$$

Similarly from alternative solutions to negative ideal solution is as:

$$DNIS_i = \sum\nolimits_j^n = 1\ NIS_j\ \log\frac{NIS_j}{W_jY_{ij}} \qquad (11)$$

Here, in the Equations (5) and (6) the Distance measure is replaced by the equation (8).

6. Experimental Setup and Result Analysis

For implementation of the approach we have used Matlab 14b, Windows 7 (64 bit Operating System) , Intel (R) Core(TM) i5- 4210U CPU @ 1.70 GHz. For step wise result analysis we have taken two data sets and each data set has twenty web sources having random values of performance scores for their features. Both the datasets have been given in Appendix A, and the step wise result analysis of TOPSIS and Improvised TOPSIS has been given below. Here the values of M=20 and N= 12 for the variables mentioned in Table 3. For few datasets the selection of best source is different and for few datasets the selection of best source is same as shown the result on the result of Dataset 1 and Dataset 2 respectively. However the result of Improvised TOPSIS effectively measures the differences in context of multidimensional features.

For The data set 1:

Step 1: Normalized Decision Matrix

0.2624	0.2725	0.3258	0.2505	0.3724	0.0724	0.1789	0.3311	0.1113	0.0895	0.2343	0.0390
0.0875	0.3114	0.2036	0.1879	0.3724	0.2534	0.0894	0.2838	0.0371	0.0448	0.3123	0.1171
0.3061	0.3503	0.2036	0.1879	0.0828	0.1086	0.1342	0.2365	0.1854	0.4029	0.1171	0.3123
0.2624	0.3503	0.2851	0.2505	0.2069	0.1448	0.1789	0.1892	0.1113	0.1343	0.1952	0.3514
0.0437	0.1946	0.1629	0.1252	0.1655	0.3258	0.2236	0.1419	0.3338	0.2686	0.2343	0.2733
0.0437	0.1168	0.1222	0.3131	0.2069	0.2896	0.0447	0.2365	0.3338	0.4029	0.0390	0.2733
0.0875	0.0389	0.2443	0.4384	0.1241	0.1448	0.2236	0.3311	0.1484	0.0895	0.1952	0.1171
0.0437	0.1946	0.3258	0.1879	0.0414	0.1086	0.1342	0.3311	0.0371	0.0895	0.2343	0.2343
0.1749	0.2335	0.1629	0.5010	0.1655	0.2172	0.1342	0.2838	0.2596	0.1791	0.0781	0.3123
0.3498	0.2725	0.0814	0.0626	0.0828	0.3258	0.1342	0.3311	0.2967	0.4029	0.3123	0.1562
0.2624	0.0389	0.1629	0.2505	0.0414	0.3258	0.0894	0.2838	0.3338	0.1791	0.3514	0.3514
0.2186	0.2335	0.1222	0.0626	0.3724	0.2172	0.4025	0.0946	0.0371	0.1343	0.3123	0.0390

0.3498	0.1946	0.1629	0.1252	0.1655	0.1086	0.4025	0.2365	0.1854	0.0895	0.1952	0.1171
0.0437	0.0778	0.3258	0.0626	0.3724	0.2896	0.3578	0.1892	0.2225	0.1791	0.1171	0.1952
0.3935	0.3503	0.1629	0.1252	0.2897	0.1448	0.3130	0.0473	0.2596	0.1791	0.2733	0.0390
0.0437	0.2335	0.1629	0.1252	0.0414	0.3258	0.0894	0.0946	0.2596	0.0895	0.1171	0.2733
0.2186	0.1557	0.1629	0.1879	0.2897	0.0362	0.1789	0.0946	0.2225	0.1791	0.3514	0.2343

Step 2: weighted normalized decision Matrix

0.0214	0.0062	0.0330	0.0088	0.0083	0.0177	0.0246	0.0057	0.0208	0.0313	0.0031	0.0211
0.0153	0.0187	0.0220	0.0132	0.0083	0.0152	0.0344	0.0028	0.0119	0.0179	0.0187	0.0211
0.0061	0.0031	0.0183	0.0044	0.0248	0.0127	0.0098	0.0085	0.0208	0.0269	0.0031	0.0070
0.0184	0.0218	0.0293	0.0175	0.0372	0.0051	0.0197	0.0199	0.0089	0.0090	0.0187	0.0035
0.0061	0.0249	0.0183	0.0132	0.0372	0.0177	0.0098	0.0170	0.0030	0.0045	0.0250	0.0105
0.0214	0.0280	0.0183	0.0132	0.0083	0.0076	0.0148	0.0142	0.0148	0.0403	0.0094	0.0281
0.0184	0.0280	0.0257	0.0175	0.0207	0.0101	0.0197	0.0114	0.0089	0.0134	0.0156	0.0316
0.0031	0.0156	0.0147	0.0088	0.0166	0.0228	0.0246	0.0085	0.0267	0.0269	0.0187	0.0246
0.0031	0.0093	0.0110	0.0219	0.0207	0.0203	0.0049	0.0142	0.0267	0.0403	0.0031	0.0246
0.0061	0.0031	0.0220	0.0307	0.0124	0.0101	0.0246	0.0199	0.0119	0.0090	0.0156	0.0105
0.0031	0.0156	0.0293	0.0132	0.0041	0.0076	0.0148	0.0199	0.0030	0.0090	0.0187	0.0211
0.0122	0.0187	0.0147	0.0351	0.0166	0.0152	0.0148	0.0170	0.0208	0.0179	0.0062	0.0281
0.0245	0.0218	0.0073	0.0044	0.0083	0.0228	0.0148	0.0199	0.0237	0.0403	0.0250	0.0141
0.0184	0.0031	0.0147	0.0175	0.0041	0.0228	0.0098	0.0170	0.0267	0.0179	0.0281	0.0316
0.0153	0.0187	0.0110	0.0044	0.0372	0.0152	0.0443	0.0057	0.0030	0.0134	0.0250	0.0035
0.0245	0.0156	0.0147	0.0088	0.0166	0.0076	0.0443	0.0142	0.0148	0.0090	0.0156	0.0105
0.0031	0.0062	0.0293	0.0044	0.0372	0.0203	0.0394	0.0114	0.0178	0.0179	0.0094	0.0176
0.0275	0.0280	0.0147	0.0088	0.0290	0.0101	0.0344	0.0028	0.0208	0.0179	0.0219	0.0035
0.0031	0.0187	0.0147	0.0088	0.0041	0.0228	0.0098	0.0057	0.0208	0.0090	0.0094	0.0246
0.0153	0.0125	0.0147	0.0132	0.0290	0.0025	0.0197	0.0057	0.0178	0.0179	0.0281	0.0211

Step 3: Fix the positive ideal and negative ideal solution

PIS = (0.0275 0.0280 0.0330 0.0351 0.0372 0.0228 0.0443 0.0199 0.0267 0.0403 0.0281 0.0316)
NIS = (0.0031 0.0031 0.0073 0.0044 0.0041 0.0025 0.0049 0.0028 0.0030 0.0045 0.0031 0.0035)

Step 4: Determine the distance measure

For TOPSIS:
DPIS = (0.0592 0.0552 0.0725 0.0595 0.0684 0.0567 0.0525 0.0552 0.0643 0.0647 0.0720 0.0552 0.0607 0.0643 0.0633 0.0601 0.0571 0.0549 0.0749 0.0564)
DNIS = (0.0549 0.0504 0.0393 0.0561 0.0526 0.0591 0.0552 0.0545 0.0580 0.0450 0.0412 0.0552 0.0627 0.0571 0.0615 0.0540 0.0616 0.0607 0.0397 0.0504)

For Improvised TOPSIS
DPIS = (0.3018 0.2689 0.4396 0.2874 0.3592 0.2474 0.2213 0.2611 0.3423 0.3400 0.4143 0.2370 0.2696 0.3080 0.3466 0.2846 0.2915 0.2689 0.4267 0.2766)
DNIS = (-0.0606 -0.0626 -0.0431 -0.0629 -0.0530 -0.0653 -0.0694 -0.0624 -0.0531 -0.0563 -0.0487 -0.0662 -0.0592 -0.0580 -0.0513 -0.0602 -0.0615 -0.0609 -0.0438 -0.0607)

Step 5: relative proximity to the ideal solution

TOPSIS Proximity (P) =
(0.4811 0.4777 0.3518 0.4854 0.4344 0.5105 0.5127 0.4967 0.4742 0.4101 0.3640 0.4999 0.5079
0.4700 0.4926 0.4732 0.5191 0.5254 0.3465 0.4719)
Improvised TOPSIS Proximity (P) =
(-0.2515 -0.3032 -0.1087 -0.2803 -0.1731 -0.3586 -0.4573 -0.3143 -0.1838 -0.1985 -0.1332 -0.3879 -0.2811
-0.2321 -0.1737 -0.2681 -0.2672 -0.2926 -0.1144 -0.2808)

Best Selection in TOPSIS = 18

Best Selection in Improvised TOPSIS = 3

For Dataset 2:

Step 1: Normalized Decision Matrix

0.1690	0.2969	0.1926	0.3397	0.1478	0.1122	0.0811	0.2959	0.2154	0.4340	0.1172	0.0934
0.1690	0.2227	0.3467	0.0377	0.0370	0.2992	0.2433	0.0658	0.1795	0.0964	0.3126	0.3734
0.0845	0.2598	0.0385	0.0755	0.2218	0.2618	0.1622	0.2959	0.1436	0.3376	0.0391	0.2334
0.1268	0.1113	0.3467	0.2265	0.2587	0.2618	0.2028	0.0986	0.1795	0.1447	0.3517	0.1400
0.2535	0.0742	0.1156	0.1132	0.3326	0.2244	0.2028	0.2302	0.3231	0.0964	0.1563	0.3267
0.3381	0.2598	0.3467	0.1510	0.1848	0.2618	0.2433	0.2630	0.1795	0.1929	0.3126	0.1867
0.1268	0.3340	0.2311	0.1887	0.1478	0.0374	0.0811	0.0986	0.2872	0.0482	0.1563	0.0467
0.0423	0.1113	0.1926	0.3397	0.1109	0.2992	0.2028	0.2959	0.0359	0.1447	0.0781	0.1400
0.2535	0.2969	0.3081	0.1887	0.2957	0.2244	0.0811	0.1315	0.0359	0.0482	0.0391	0.1867
0.0845	0.0742	0.0385	0.0377	0.2218	0.1122	0.1622	0.1973	0.1077	0.0482	0.3517	0.1400
0.0845	0.2227	0.0770	0.0755	0.3326	0.2244	0.2433	0.2630	0.3231	0.1447	0.1563	0.0934
0.2535	0.3340	0.2696	0.1510	0.3326	0.0748	0.2028	0.0986	0.3231	0.3858	0.3126	0.3734
0.2958	0.2969	0.1926	0.2265	0.1478	0.1122	0.3650	0.2959	0.2513	0.0482	0.1954	0.2334
0.2958	0.2227	0.1926	0.1887	0.0370	0.1870	0.2433	0.2630	0.2513	0.3858	0.1563	0.1400
0.3381	0.2227	0.1926	0.3397	0.3326	0.2244	0.3650	0.2959	0.2872	0.1929	0.3517	0.4201
0.3381	0.0371	0.0385	0.3397	0.1478	0.2992	0.3650	0.1315	0.1436	0.1929	0.1563	0.1867
0.1268	0.2227	0.2311	0.2265	0.0370	0.0748	0.2433	0.1315	0.1077	0.1929	0.1563	0.1400
0.2113	0.1113	0.1541	0.3019	0.1109	0.3366	0.1217	0.1973	0.1436	0.3376	0.0781	0.2801
0.1690	0.1113	0.3081	0.3019	0.2957	0.1122	0.1622	0.2630	0.2872	0.0964	0.3126	0.0934
0.2958	0.2598	0.1926	0.1510	0.2587	0.3366	0.1622	0.2630	0.2872	0.1929	0.1563	0.1400

Step 2: weighted normalized decision Matrix

0.0118	0.0238	0.0173	0.0238	0.0148	0.0079	0.0089	0.0178	0.0172	0.0434	0.0094	0.0084
0.0118	0.0178	0.0312	0.0026	0.0037	0.0209	0.0268	0.0039	0.0144	0.0096	0.0250	0.0336
0.0059	0.0208	0.0035	0.0053	0.0222	0.0183	0.0178	0.0178	0.0115	0.0338	0.0031	0.0210
0.0089	0.0089	0.0312	0.0159	0.0259	0.0183	0.0223	0.0059	0.0144	0.0145	0.0281	0.0126
0.0177	0.0059	0.0104	0.0079	0.0333	0.0157	0.0223	0.0138	0.0258	0.0096	0.0125	0.0294
0.0237	0.0208	0.0312	0.0106	0.0185	0.0183	0.0268	0.0158	0.0144	0.0193	0.0250	0.0168
0.0089	0.0267	0.0208	0.0132	0.0148	0.0026	0.0089	0.0059	0.0230	0.0048	0.0125	0.0042
0.0030	0.0089	0.0173	0.0238	0.0111	0.0209	0.0223	0.0178	0.0029	0.0145	0.0063	0.0126

0.0177	0.0238	0.0277	0.0132	0.0296	0.0157	0.0089	0.0079	0.0029	0.0048	0.0031	0.0168
0.0059	0.0059	0.0035	0.0026	0.0222	0.0079	0.0178	0.0118	0.0086	0.0048	0.0281	0.0126
0.0059	0.0178	0.0069	0.0053	0.0333	0.0157	0.0268	0.0158	0.0258	0.0145	0.0125	0.0084
0.0177	0.0267	0.0243	0.0106	0.0333	0.0052	0.0223	0.0059	0.0258	0.0386	0.0250	0.0336
0.0207	0.0238	0.0173	0.0159	0.0148	0.0079	0.0401	0.0178	0.0201	0.0048	0.0156	0.0210
0.0207	0.0178	0.0173	0.0132	0.0037	0.0131	0.0268	0.0158	0.0201	0.0386	0.0125	0.0126
0.0237	0.0178	0.0173	0.0238	0.0333	0.0157	0.0401	0.0178	0.0230	0.0193	0.0281	0.0378
0.0237	0.0030	0.0035	0.0238	0.0148	0.0209	0.0401	0.0079	0.0115	0.0193	0.0125	0.0168
0.0089	0.0178	0.0208	0.0159	0.0037	0.0052	0.0268	0.0079	0.0086	0.0193	0.0125	0.0126
0.0148	0.0089	0.0139	0.0211	0.0111	0.0236	0.0134	0.0118	0.0115	0.0338	0.0063	0.0252
0.0118	0.0089	0.0277	0.0211	0.0296	0.0079	0.0178	0.0158	0.0230	0.0096	0.0250	0.0084
0.0207	0.0208	0.0173	0.0106	0.0259	0.0236	0.0178	0.0158	0.0230	0.0193	0.0125	0.0126

Step 3: Fix the positive ideal and negative ideal solution

PIS = (0.0237 0.0267 0.0312 0.0238 0.0333 0.0236 0.0401 0.0178 0.0258 0.0434 0.0281 0.0378)
NIS = (0.0030 0.0030 0.0035 0.0026 0.0037 0.0026 0.0089 0.0039 0.0029 0.0048 0.0031 0.0042)

Step 4: Determine the distance measure

For TOPSIS:

DPIS = (0.0565 0.0567 0.0576 0.0524 0.0548 0.0425 0.0721 0.0651 0.0663 0.0727 0.0593 0.0332 0.0531 0.0499 0.0304 0.0570 0.0609 0.0537 0.0571 0.0494)

DNIS = (0.0569 0.0568 0.0488 0.0534 0.0539 0.0590 0.0404 0.0401 0.0486 0.0355 0.0498 0.0732 0.0563 0.0548 0.0766 0.0534 0.0384 0.0506 0.0537 0.0530)

For Improvised TOPSIS

DPIS = (0.2494 0.2866 0.3117 0.2240 0.2426 0.1542 0.4128 0.3481 0.3605 0.4382 0.2816 0.1254 0.2278 0.2276 0.0770 0.2810 0.3235 0.2459 0.2452 0.1931)

DNIS = (0.0569 0.0568 0.0488 0.0534 0.0539 0.0590 0.0404 0.0401 0.0486 0.0355 0.0498 0.0732 0.0563 0.0548 0.0766 0.0534 0.0384 0.0506 0.0537 0.0530)

Step 5: relative proximity to the ideal solution

TOPSIS Proximity (P) =

(0.5018 0.5002 0.4589 0.5050 0.4957 0.5815 0.3588 0.3813 0.4231 0.3278 0.4564 0.6878 0.5144 0.5231 0.7160 0.4837 0.3864 0.4853 0.4847 0.5178)

Improvised TOPSIS Proximity (P) =

(-0.3009 -0.2401 -0.2078 -0.3855 -0.3445 -0.9026 -0.0992 -0.1690 -0.1386 -0.0970 -0.2624 -1.4183 -0.4166 -0.3994 12.3023 -0.2686 -0.1944 -0.3166 -0.3312 -0.5311)

Best Selection in TOPSIS = 15

Best Selection in Improvised TOPSIS = 15

7. Conclusion

In this article the TOPSIS approach is improvised by replacing the Euclidian Distance measure with Kullback Leibler Divergence to evaluate the web sources efficiently. Euclidean Distance measure only compute the linear distances between the alternative web sources (Alternative Solutions) and ideal web sources (Ideal Solutions) to compute the relative proximity[8,12]. Moreover the random and dynamic nature of web sources poses difficulty during computation of relative proximity using Euclidian distance measure. The better way is to use the probabilistic approach to deal for precise evaluation of Web sources. Kullback Leibler Divergence computes differences between two probability distribution. Kullback Leibler Divergence compute differences not only linear distance between them[9,10]. Consequently the evaluation is more relevant for web source selection.

8. References

1. Inmon WH. Building the Data Warehouse. John Wiley & Sons. 2005.
2. Ponniah P. Data Warehousing Fundamentals: A Comprehensive Guide for IT Processionals. John Wiley & Sons. 2001.
3. Pedersen TB, Jensen CS. Multidimensional Databases. The Industrial Information Technology Handbook. In: Zurawski R editor. CRC Press. 2005; 1 –13.
4. Perez JM, Berlanga R, Aramburu MJ, Pedersen TB. Integrating data warehouse with web data: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2008.
5. Xyleme L. A Dynamic Warehouse for XML Data of the Web. IEEE Data Eng. 2001; 24(2):40–7.
6. Tan X, Yen DC, Fang X. Web warehousing: Web technology meets data warehousing. Technology in Society. 2003; 25:131–48.
7. Zhu Y, Buchmann AP. Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse. Web Information Systems Engineering. 2002.
8. Parimala Devi R, Thigarasu V. A Semantic Deduplication of Temporal Dynamic Records from Multiple Web Databases. Indian Journal of Science and Technology. 2015 Dec; 8(34).
9. Carol I, Britto Ramesh Kumar S. Conflict Resolution and Duplicate Elimination in Heterogeneous Datasets using Unified Data Retrieval Techniques.Indian Journal of Science and Technology. 2015 Sep; 8(22).
10. Velasquez M, Hester PT. An Analysis of Multi-Criteria Decision Making Methods. International Journal of Operations Research. 2013; 10(2):56–66.
11. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951.
12. Johnson D, Sinanovic S. Symmetrizing the kullback-leibler distance. 2001.
13. Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information theory. 2003.
14. Triantaphyllou E, Shu B, Sanchez SN, Ray T. Multi-criteria decision making: an operations research approach. Encyclopedia of Electrical and Electronics Engineering. 1998.
15. Ullah A. Entropy, divergence and distance measure with econometric applications. Journal of Statistical Planning and Inference. 1996.
16. Cha S-H. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences. 2007.
17. Ross S. Introduction to Probability Models. Academic Press/Elsevier. 2012.
18. Johnson JL. Probability and Statistics for Computer Science. Wiley. 2008.

Appendix A

Data Set 1

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
WS1	7	2	9	2	2	7	5	2	7	7	1	6
WS2	5	6	6	3	2	6	7	1	4	4	6	6
WS3	2	1	5	1	6	5	2	3	7	6	1	2
WS4	6	7	8	4	9	2	4	7	3	2	6	1
WS5	2	8	5	3	9	7	2	6	1	1	8	3
WS6	7	9	5	3	2	3	3	5	5	9	3	8
WS7	6	9	7	4	5	4	4	4	3	3	5	9
WS8	1	5	4	2	4	9	5	3	9	6	6	7
WS9	1	3	3	5	5	8	1	5	9	9	1	7
WS10	2	1	6	7	3	4	5	7	4	2	5	3
WS11	1	5	8	3	1	3	3	7	1	2	6	6
WS12	4	6	4	8	4	6	3	6	7	4	2	8
WS13	8	7	2	1	2	9	3	7	8	9	8	4
WS14	6	1	4	4	1	9	2	6	9	4	9	9
WS15	5	6	3	1	9	6	9	2	1	3	8	1
WS16	8	5	4	2	4	3	9	5	5	2	5	3
WS17	1	2	8	1	9	8	8	4	6	4	3	5
WS18	9	9	4	2	7	4	7	1	7	4	7	1
WS19	1	6	4	2	1	9	2	2	7	2	3	7
WS20	5	4	4	3	7	1	4	2	6	4	9	6

Data Set 2

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
WS1	4	8	5	9	4	3	2	9	6	9	3	2
WS2	4	6	9	1	1	8	6	2	5	2	8	8
WS3	2	7	1	2	6	7	4	9	4	7	1	5
WS4	3	3	9	6	7	7	5	3	5	3	9	3
WS5	6	2	3	3	9	6	5	7	9	2	4	7
WS6	8	7	9	4	5	7	6	8	5	4	8	4
WS7	3	9	6	5	4	1	2	3	8	1	4	1
WS8	1	3	5	9	3	8	5	9	1	3	2	3
WS9	6	8	8	5	8	6	2	4	1	1	1	4
WS10	2	2	1	1	6	3	4	6	3	1	9	3
WS11	2	6	2	2	9	6	6	8	9	3	4	2
WS12	6	9	7	4	9	2	5	3	9	8	8	8
WS13	7	8	5	6	4	3	9	9	7	1	5	5
WS14	7	6	5	5	1	5	6	8	7	8	4	3
WS15	8	6	5	9	9	6	9	9	8	4	9	9
WS16	8	1	1	9	4	8	9	4	4	4	4	4
WS17	3	6	6	6	1	2	6	4	3	4	4	3
WS18	5	3	4	8	3	9	3	6	4	7	2	6
WS19	4	3	8	8	8	3	4	8	8	2	8	2
WS20	7	7	5	4	7	9	4	8	8	4	4	3