Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse

Yan Zhu Alejandro P. Buchmann

Department of Computer Science, Darmstadt University of Technology
64283 Darmstadt, Germany
{zhu, buchmann}@dvs1.informatik.tu-darmstadt.de

Abstract

A company's local data often is not sufficient to analyze market trends and make reasonable business plans. Decision making must also be based on information from suppliers, partners and competitors. Systematically integrating suitable external data from the Web into a data warehouse is a meaningful solution and will benefit the enterprise. However, the autonomy and the dynamics of the Web make the task of selecting relevant and qualified external data from the Web challenging.

In this paper, we develop a set of criteria for evaluating and selecting Web resources as external data sources of a data warehouse and discuss how to screen Web data sources using Multi-Criteria Decision Making (MCDM) methods. The final decision with respect to selecting Web sources is sensitive to critical factors, i.e., the criterion weight and the performance score of alternatives in terms of each criterion. In this paper, we analyzed the sensitivity of the final rank of alternatives in terms of the critical factors in oder to gain an insight into the stability of our final decision. The comparison of several MCDM approaches for Web source screening is also presented in the paper.

1 Introduction

The Web has become an independent platform for providing and accessing information of almost any type. At the same time, data warehousing emerged as a technique to support OLAP and decision making in an enterprise. As Web technologies develop, trading becomes faster and more complex, and the scope and type of business activities broaden. Data from an enterprise's internal data sources has already become insufficient for strategic business decisions, and external data has gained importance to complement business analysis and decision making. Because of the huge amount of information available on the Web, systematically integrating suitable external data from the Web with a company's internal data in a data warehouse is a promising approach [4, 19, 20, 22].

However, identifying relevant external data from the Web is like finding a needle in a haystack, and the situation becomes more complex because of the dynamics of the Web [4]. Therefore, the first task in setting up a Web warehousing system is to evaluate a set of relevant Web sources and to select high quality and compatible sources as the external information resources of a data warehouse. Several issues must be taken into consideration:

- **Web source stability**

  According to the statistics of Zooknic (`http://www.zooknic.com/Domains/counts.html`) on Feb. 17, 2002 , the total number of domains registered worldwide was 28,605,953 and the total number of .com's was 22,299,727. The owners of about 28 million Web sites may be government agencies, organizations, companies or individuals. This results in highly heterogeneous information and different design styles on the Web.

  Web sources are also very dynamic. Not only is Web-based data updated frequently, but new Web sources and new pages are made available on the Internet every day. It is estimated that 7 million or more new pages are being added daily. Already available sources may be changed drastically or even disappear.

- **Web data quality**

  The Web mechanism is so open and independent that Web masters can publish on the Web whatever information they like. A large amount of information on the Web is not as carefully examined, reviewed and filtered as traditional publications. Wrong information, incomplete data or vague facts exist on the Web, even correct data can become difficult to use due to poor presentation (e.g.: lack of units or time stamps). Therefore, data quality on the Web is irregular.

- **Application specifics**

  A common supported notation of data quality is that high quality data should conform to user requirements. Different

> consuming purposes of data can result in different requirements on data quality. We identify several requirements of a Web warehousing system on Web data sources: ***a.*** Relevance of external data to the business analysis; ***b.*** Easy extraction to needed data; ***c.*** Necessary meta data, such as data definition, data format, and derivation rules, can be provided with data together.
>
> Because of that the publishing intention of Web data is mostly for browsing rather than integration. These requirements may not be met.

Therefore, evaluating and selecting credible and compatible Web sources is an important task in a Web warehousing project. The designer of such a data warehouse must design a set of criteria to assess Web sources and apply decision-making criteria in selecting the most suitable sources. In this paper, we will investigate these issues and focus on:

- developing a set of criteria for selecting Web resources as external data sources of a data warehouse;
- evaluating Web sources with Multi-Criteria Decision Making (MCDM) methods;
- analyzing the sensitivity of the final decision with respect to critical measures;
- comparing several evaluation approaches.

The rest of the paper is organized as follows: In Section 2, we discuss the source quality problem and propose a set of criteria for web information evaluation. In Section 3, we introduce known MCDM methods into the area of Web source evaluation and selection and study how to use them to screen Web sources for warehousing. Section 4 analyzes the sensitivity of the final decision in terms of critical measures. The features of MCDM methods are compared in section 5. Finally, we discuss related work in Section 6 and present our conclusions in Section 7.

2 Web Source Evaluation Criteria

The problem of information quality (IQ) and data quality (DQ) has long attracted the attention of the research community. One example is the TDQM project at MIT. In this project, Wang et al. studied theoretically-grounded methodologies for data quality management and proposed four IQ categories and identified fifteen important attributes [18]. Wang's quality criteria are an excellent starting point. For the purposes of integrating a company's internal data with external Web sources, we found other aspects beyond data quality to be relevant, for example, the stability of Web sites and the availability of metadata. Therefore, we propose 12 evaluation criteria that are grouped into three categories: *stability of a Web source*, *quality of the Web data*, and *application specific or contextual issues*.

Stability of a Web source For the stability of a Web source we take into account *availability*, *accessibility*, *durability*, and *refresh rate*.

Availability describes the fact that a site is up and running, its response time, and whether the pages are reachable through the links.

Accessibility refers to whether the data is accessible without additional requirements, e.g., registration and password. This is particularly important when data is to be extracted automatically for transfer to the data warehouse.

Durability refers to the time a particular data item is kept at a Web site. Historical data may be kept at the Web site similar to a data warehouse or it may simply be overwritten or removed from the site. For data warehousing the implication is that volatile data must be extracted regularly and downloaded to the data warehouse to guarantee its availability.

Refresh rate covers two main aspects: for once it refers to the timeliness with which data is posted to the site, but it also means that volatile data that is overwritten at a fast refresh rate must be extracted at the same rate to avoid loosing data.

Quality of Web data The quality of Web data must be evaluated with respect to *origin*, *correctness*, *completeness*, *objectivity*, and *metadata*.

Origin has an effect on reliability of the data and the trust one can place in it. It is often referred to as data lineage.

Correctness denotes that data is free of errors.

Completeness describes the coverage of the data.

Objectivity refers to the lack of bias in the data.

Completeness and objectivity are not fully orthogonal criteria. For example, if a site presents benchmark results, these may be correct but incomplete because some results favoring a competitor are omitted. In this case the bias results from the lack of completeness.

Metadata refers to the availability of descriptive metadata that may range from units of measurement to calculation method and derivation rules for some data. This aspect is particularly important in the evaluation of a source for web warehousing because misinterpreted data may contaminate the data warehouse and produce wrong results.

Application specific or contextual issues Contextual issues are those issues that depend directly on the intended use of the collected data. To keep the complexity of the evaluation manageable, we limit the application specific issues to *relevance*, *presentation*, and *timeliness*.

Relevance often preempts all other parameters. For example, if the manager of an on-line bookstore wants to develop a marketing strategy, she will be forced to integrate pricing data and specials from the immediate

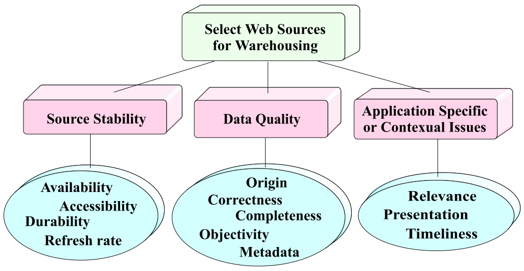

Figure 1: Hierarchy of criteria for Web sources selection

competitors. In this case, other quality issues loose importance. However, if we want to integrate reference data, such as exchange rates, quality issues such as origin and presentation become more important.

Presentation refers to the various formats the data can be presented in, ranging from HTML, XML, pdf, ps, doc or any other document format to pictorial or audio data. For automatic extraction and use as feed to a data warehouse, only structured or semi-structured data is useful.

Timeliness refers the promptness with which data is available. Clearly there is a trade-off between value and timeliness that is most clearly exemplified by the stock quotations which are free if provided with a time lag of 15 minutes but have a price if provided immediately. Applications have a varying sensibility to delays.

Summarizing, we propose three assessment dimensions for Web sources with a total of 12 criteria. These are summarized in Figure 1 and will be used as the basis for evaluation. Scores can be assigned to a source for each criterion. How the scores are assigned and used to produce a single, synthetic score will be discussed in the next section.

3 Evaluating and Selecting Web Sources Using MCDM Methods

Evaluating and selecting Web data sources can be classified as a Multi-Criteria Decision Making (MCDM) problem. There are two kinds of methods for solving a MCDM problem. One is compensatory and the other is non-compensatory [5]. The non-compensatory method does not permit tradeoffs among attributes. The MCDM techniques in this category are simple, but may not be very suitable for Web source evaluation. In contrast, the compensatory method allows tradeoffs among attributes. A slight decline in one attribute is acceptable if it is compensated by some enhancement in one or more other attributes. There are three subgroups in this category: scoring, compromising, and concordance methods.

We select four popular approaches from the compensatory method to assess Web sources, they are the SAW (Simple Additive Weighting) and AHP (Analytic Hierarchy Process) from the scoring methods, the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) from the compromising methods, and the DEA (Data Envelopment Analysis) from the concordance approaches.

In order to illustrate how to evaluate and select Web sources for data warehousing, we use online bookstores as an example:

An online computer book store manager uses a data warehouse to manage her e-commerce data. In addition, she wants to acquire the discount book information from other online computer book shops, integrate it with her company's data and materialize it in the data warehouse to better plan her own promotions. The manager will preselect some relevant Web sources as candidates for further evaluation by using a search engine, then evaluate them to find one or several Web sources which best meet quality criteria and data integration requirements. These sources will be determined as external data sources for the data warehouse system.

4 on-line computer book stores are identified as alternatives: A, B, C, D. Among the 12 criteria given in the last section, 2 criteria per dimension are selected in the assessment, in order to simplify the calculation.

3.1 The SAW method

In the SAW method, each criterion will be given a weight, the sum of all weights must be 1. Table 1 shows an example of weight values. Each alternative is rated with regard to every one of the 6 criteria selected for illustration. Since there is no standard for setting a rating scale in the SAW, thus it can be determined by decision makers. A scale from 1 (least desirable) to 9 (most desirable) was used in this example. Table 2 shows the performance score of alternatives in terms of each criterion.

Table 1: Weights of six quality criteria

Criteria	Weight
availability	0.15
accessibility	0.1
correctness	0.2
completeness	0.1
relevance	0.3
presentation	0.15

$$SAW_i = \sum_{j=1}^{N} a_{ij} w_j, \quad for \ i = 1, 2, 3, ..., M \qquad (3.1)$$

where, SAW_i is the SAW score of the i^{th} alternative; M and N are the number of alternatives and decision criteria separately; a_{ij} is the score of the i^{th} alternative in terms of the j^{th} criterion, and w_j is the weight of importance of the j^{th} criterion.

Table 2: Scores of alternatives in terms of each criterion

Web source	availability	accessibility	correctness	completeness	relevance	presentation
A	8	5	6	9	4	8
B	7	4	6	8	9	7
C	6	8	7	6	6	4
D	5	6	4	6	8	4

Applying the Formula 3.1 to Table 2, we obtain the final ranking scores as SAW_A = 6.20, SAW_B = 7.20, SAW_C = 6.10, SAW_D = 5.75. Source B is the best candidate for Web warehousing.

3.2 The AHP approach

The AHP method was developed by Thomas Saaty in 1980. It is composed of several previously existing but unassociated concepts and techniques, such as, hierarchical structuring of complexity, pairwise comparisons, redundant judgments, an eigenvector method for deriving weights, and consistency considerations [3, 7]. It can be carried out according to the following steps:

Step 1: Developing a goal hierarchy

In this step, the overall goal, criteria, and decision alternatives are built in a hierarchical structure, which is shown in Figure 2.

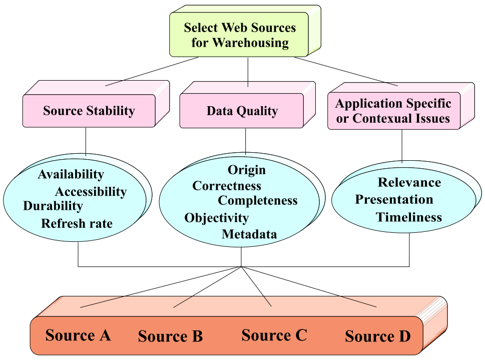

Figure 2: A goal hierarchy

After decomposing the problem into a hierarchy, alternatives at a given hierarchy level are compared in pairs to assess their relative preference with regard to each criterion at the higher level. A scale is needed to represent the varying degrees of preference. Saaty establishes a scale* (Table 3), where 9 is the upper limit and 1 the lower limit and a unit difference between successive scale values is used.

Step 2: Setting up a pairwise comparison matrix of criteria

A comparison is implemented among the elements that are on the same level of the goal hierarchy. In a comparing process, a value $\mathcal{V}$ from the scale is assigned to the comparison result of two criteria $\mathcal{P}$ and $\mathcal{Q}$ at first, then the value of comparison of $\mathcal{Q}$ and $\mathcal{P}$ is a reciprocal value of $\mathcal{V}$, i.e., $\frac{1}{\mathcal{V}}$. The value of the comparison of $\mathcal{P}$ and $\mathcal{P}$ is 1. Following these rules, a matrix is built. The weights of the matrix attributes are calculated through finding the eigenvector associated with the maximal eigenvalue of this matrix. A practically used algorithm is:

a. normalize each column by dividing each cell by the column total.

b. sum each of the rows into a new column.

c. normalize the new column by dividing each value by the sum of the column.

d. the normalized column represents the eigenvector, which contains the weight of each attribute .

The results are shown in Table 4 and Table 5†, separately.

Step 3: Ranking the relative importance between alternatives

In this step, the relative importance between each pair of alternatives in terms of a criterion will be assigned. As in step 2, all matrices are normalized and the weight of each alternative is also derived. Table 6 and Table 7 give an example of this calculation, the other computation results are omitted due to space limitations.

Step 4: Checking consistency of the comparisons

Since the data warehouse designer weighs all elements based on his own judgment, inconsistency is possible in building a weight matrix. For example, an element $\mathcal{P}$ could be weighted strongly more important than $\mathcal{Q}$, $\mathcal{Q}$ could be weighted more important than $\mathcal{R}$, and $\mathcal{R}$ could be slightly more important than $\mathcal{P}$, so that $\mathcal{P}$ is implied to be strongly more important than itself. A decision based on such an inconsistency is obviously meaningless.

An index of consistency ratio (CR) can be used to measure consistency of a n-order square decision matrix. In AHP, the threshold for CR is 0.1, when the value of a CR

*This scale is built based on psychological experiments, which have shown that individuals have difficulty to compare more than five to nine objects at one time.

†All elements from a pairwise comparison matrix are of the floating-point data type in a Java program. They are calculated and recorded in the corresponding normalized matrix by rounding off.

Table 3: Scale of degrees of preference

Importance	Definition
1	Equal importance
3	moderately preferred
5	strongly preferred
7	very strongly preferred
9	extremely preferred
2,4,6,8	intermediate values

$$S=\left\{\frac{1}{9},\frac{1}{8},\frac{1}{7},\frac{1}{6},\frac{1}{5},\frac{1}{4},\frac{1}{3},\frac{1}{2},1,2,3,4,5,6,7,8,9\right\}$$

Table 4: Pairwise comparison values of criteria

criterion	AVailability	ACcessibility	COrrectness	COMpleteness	RElevance	PResentation
AV	1	3	1/2	4	1/3	2
AC	1/3	1	1/5	2	1/6	1/2
CO	2	5	1	6	1/3	3
COM	1/4	1/2	1/6	1	1/7	1/3
RE	3	6	3	7	1	5
PR	1/2	2	1/3	3	1/5	1

Table 5: Normalized pairwise comparison value matrix of criteria

criterion	AVailability	ACcessibility	COrrectness	COMpleteness	RElevance	PResentation	Σ	Weights
AV	0.14	0.17	0.10	0.17	0.15	0.16	0.90	0.15
AC	0.05	0.06	0.04	0.09	0.08	0.04	0.35	0.06
CO	0.28	0.29	0.19	0.26	0.15	0.25	1.43	0.23
COM	0.04	0.03	0.03	0.04	0.06	0.03	0.23	0.04
RE	0.42	0.34	0.58	0.30	0.46	0.42	2.53	0.42
PR	0.07	0.11	0.06	0.13	0.09	0.08	0.56	0.10
Total	1	1	1	1	1	1	6	1.00

Table 6: Pairwise comparison matrix for relevance

Web source	A	B	C	D
A	1	1/6	2	1/4
B	6	1	8	3
C	1/2	1/8	1	1/5
D	4	1/3	5	1

Table 7: Normalized pairwise comparison matrix for relevance

Web source	A	B	C	D	Σ	Weight
A	0.09	0.10	0.13	0.06	0.37	0.09
B	0.52	0.61	0.50	0.67	2.31	0.58
C	0.04	0.08	0.06	0.04	0.23	0.06
D	0.35	0.20	0.31	0.22	1.09	0.27
total	1	1	1	1	4	1

is lower than 0.1, the decision matrix is accepted and can be applied to making decisions. Otherwise, the matrix must be reevaluated.

The calculation of CR of a pairwise comparison matrix is implemented by taking the following solutions:

$$WtsAvg = MMULT(Comparison\ Matrix,\ Weights) \tag{3.2}$$

where, $Weights$ is a weight vector of the normalized comparison matrix, $WtsAvg$ is the return vector of the multiplication of the $Comparison\ Matrix$ and the $Weight\ Matrix$.

$$RatioAvg = avg(WtsAvg/Weights) \tag{3.3}$$

where, $RatioAvg$ is the average of the $Ratio$.

$$Consistency\ Index(CI) = (RatioAvg - n)/(n-1) \tag{3.4}$$

where, n is the order of a pairwise comparison square matrix.

$$Consistency\ Ratio(CR) = CI/RI \tag{3.5}$$

where, RI is the random inconsistency.

Step 5: Calculating AHP values

The AHP value is computed using the following formula:

$$AHP_i = \sum_{j=1}^{N} a_{ij} w_j,\ \ for\ \ i = 1,\ 2,\ 3, ...,\ M \tag{3.6}$$

where, M is the number of alternatives and N is the number of criteria; a_{ij} denotes the score of the i^{th} alternative related to the j^{th} criterion; w_j denotes the weight of the j^{th} criterion.

As the result: AHP_A = 0.229; AHP_B = 0.391; AHP_C = 0.208; AHP_D = 0.172;

Source B is the best alternative for warehousing, since it has the highest AHP score.

3.3 The TOPSIS method

The TOPSIS method was developed by Hwang and Yoon in 1981. Its basic approach is to find an alternative which is closest to the ideal solution and farthest to the negative-ideal solution in a multi-dimensional computing space. This multi-dimensional computing space is specified by a set of evaluation criteria as dimensions. The ideal solution represents a virtual alternative with a set of possibly best synthetic scores in terms of each criterion, while the negative-ideal solution is a virtual alternative with a set of worst scores. Physically, they are two points in the computing space with extreme values as dimensions.

In TOPSIS, four alternatives and six criteria in the running example result in a 4×6 matrix X, the value of each element in the matrix is the performance score of an alternative with regard to each criterion.

$$X = \begin{vmatrix} 8 & 5 & 6 & 9 & 4 & 8 \\ 7 & 4 & 6 & 8 & 9 & 7 \\ 6 & 8 & 7 & 6 & 6 & 4 \\ 5 & 6 & 4 & 6 & 8 & 4 \end{vmatrix}$$

To assess four alternatives, we must execute the following steps [17]:

1. **Normalizing the decision matrix**

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{M}\ x_{ij}^2}} \tag{3.7}$$

where, M is the number of the alternatives; x_{ij} denotes the performance score of the i^{th} alternative in terms of the j^{th} criterion.

2. **Building the weighted normalized decision matrix WY**

$$WY = w_j y_{ij} \tag{3.8}$$

where, w_j denotes the weight of the j^{th} criterion (refer to Table 1)

3. **Determining the ideal and the negative-ideal solutions**

The ideal solution S^+ and the negative-ideal Solution S^- are defined using (3.9):

$$S_j^+ = max(\ w_j y_{ij}\ ),\ \ S_j^- = min(\ w_j y_{ij}\ )$$
$$i = 1,\ 2, \cdots,\ M \tag{3.9}$$

In our running example, we have

$S^+ = (0.091, 0.067, 0.119, 0.061, 0.192, 0.099)$
$S^- = (0.057, 0.034, 0.068, 0.041, 0.085, 0.049)$

4. **Finding the Euclidean distances of each alternative**

In this step, the Euclidean distances of an alternative to S^+ and S^- will be calculated separately as:

$$D_i^+ = \sqrt{\sum_{j=1}^{N}(s_j^+ - w_j y_{ij})^2},\ \ i = 1,\ 2, \cdots,\ M \tag{3.10}$$

$$D_i^- = \sqrt{\sum_{j=1}^{N}(w_j y_{ij} - s_j^-)^2},\ \ i = 1,\ 2, \cdots,\ M \tag{3.11}$$

In the running example, we have:

$D^+ = (0.111,\ 0.042,\ 0.087,\ 0.086)$, $D^- = (0.073,\ 0.121,\ 0.075,\ 0.087)$

5. **Calculating the relative closeness to the ideal solution**

   The relative closeness of the i^{th} alternative to the ideal solution is defined as follows:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}, \ \ 0 \leq C_i \leq 1, \ \ i = 1, \ 2, \ 3, ..., \ M \tag{3.12}$$

   All alternatives are compared with the positive ideal solution and the negative ideal solution, if an alternative itself is the positive ideal solution, C = 1; if an alternative itself is the negative ideal solution, C = 0. The larger the relative closeness value, the closer to the ideal solution and the farther to the negative solution.

   In our example, the relative closeness values of 4 alternatives are: B : 0.743 D : 0.503 C : 0.466 A : 0.396, thus B is the most suitable Web source for warehousing.

3.4 The DEA approach

The Data Envelopement Analysis (DEA) is a Linear Programming (LP) based technique for performance measurement. It was first introduced by Charnes, Cooper and Rhodes in 1978 as a general method to evaluate the efficiency of a number of producers.

In our Web source screening problem, the DEA approach is employed to find one or more qualified Web sources with the highest score under a set of quality criteria. These sources are actually the optimal solutions of a set of linear programs, which can be expressed using the formulation of the basic DEA model:

$$\begin{aligned} maximize \quad & E_{S_i} = \sum_j O_{S_ij} v_j \qquad (3.13) \\ subject\ to \quad & E_{S_k} \leq 1, \quad for\ all\ alternative\ S_k, \\ & \sum_l I_{S_il} u_l = 1, \\ & v_j, u_l \geq 0 \end{aligned}$$

where, E_{S_i} is the efficiency score of alternative S_i; O_{S_ij} denotes the value of the j^{th} output criterion O for alternative S_i; I_{S_il} denotes the value of the l^{th} input criterion I for alternative S_i; v_j is the weight assigned to the alternative S_i for maximizing O_j; u_l is the weight assigned to alternative S_i for minimizing I_l.

The basic DEA classifies all tested alternatives as efficient (E_S = 1) and non-efficient ($E_S < 1$), and does not always distinguish each alternative. Thus, Andersen and Petersen extended the basic DEA to a ranking model in 1993. The model allows the efficient score of an alternative to be greater than 1, so that the difference of efficient scores of all alternatives is identified. This model can be represented as:

$$\begin{aligned} maximize \quad & E_{S_0} = \sum_j O_{S_0j} v_j \qquad (3.14) \\ & subject\ to \\ E_{S_k} \leq 1, \quad & for\ all\ alternative\ S_k, and\ S_k \neq S_0, \\ & \sum_l I_{S_il} u_l = 1, \\ & v_j, u_l \geq 0 \end{aligned}$$

Now, we apply this model to the running example. The output value of each Web source in terms of each criterion is as shown in Table 2. Given $v_1, \ v_2, \ v_3, \ v_4, \ v_5, \ v_6$ as the weights of six criteria separately. 3 inequalities for an alternative must be resolved. The LP problem of Source A is for example expressed as:

$$\begin{aligned} maximize \quad & E_A = 8v_1 + 5v_2 + 6v_3 + 9v_4 + 4v_5 + 8v_6 \ (3.15) \\ subject\ to \quad & 7v_1 + 4v_2 + 6v_3 + 8v_4 + 9v_5 + 7v_6 \leq 1 \\ & 6v_1 + 8v_2 + 7v_3 + 6v_4 + 6v_5 + 4v_6 \leq 1 \\ & 5v_1 + 6v_2 + 4v_3 + 6v_4 + 8v_5 + 4v_6 \leq 1 \\ & \sum_l I_A u_l = 1 \\ & v_1, \ v_2, \ v_3, \ v_4, \ v_5, \ v_6 \geq 0 \end{aligned}$$

Apparently, a problem here is that the number of unknown variables is more than the number of conditions. In order to obtain a finite number of basic solutions, a common method for this case is to let the number of the unknowns be the same as the number of conditions through assuming the values of the rest of unknowns as 0, and remove those respective items from the constraints. That means, some output values (criteria) will not be taken into consideration. In our Web source evaluation example, each criterion represents one important index of a quality dimension. Our experiments show, omitting any one criterion may disturb final judgment on alternatives. In order to reasonably represent three quality dimensions, we synthesize the values of criteria of each quality dimension to one value per one dimension, rather than letting a part of unknowns be zero, so we obtain the new output values in terms of *source stability*, *data quality*, and *application requirement*, which are shown in Table 8.

Given x, y, z as the weights of 3 criteria, the LP problem of (3.15) is changed as follows:

$$maximize \quad E_A = 6.5x + 7.5y + 6.0z \tag{3.16}$$

Table 8: The new output value of each alternative

Web source	Criteria		
	source stability	data quality	application requirement
A	6.5	7.5	6.0
B	5.5	7.0	8.0
C	7.0	6.5	5.0
D	5.5	5.0	6.0

$$\begin{aligned} subject\ to\quad 5.5x + 7.0y + 8.0z &\leq 1 \\ 7.0x + 6.5y + 5.0z &\leq 1 \\ 5.5x + 5.0y + 6.0z &\leq 1 \\ \sum_l I_A u_l &= 1 \\ x, y, z &\geq 0 \end{aligned}$$

In order to resolve this kind of LP problem, Dantzig proposed the Simplex Solution in 1963 [10, 13]. Through solving LP problems of all alternatives following the Simplex Solution, we obtain the efficient score of each alternative as follows:

E_A = 1.1, E_B = 1.35, E_C = 1.07, E_D = 0.89

The Source B has the best efficiency in terms of the same set of criteria, thus it is the most qualified candidate for warehousing.

4 Sensitivity Analysis of MCDM approaches

4.1 Sensitivity analysis of the SAW and AHP

In the SAW and AHP approaches, the weight of a criterion and the performance score of an alternative with respect to each criterion are assigned subjectively, either directly as in the SAW or through pairwise comparison as in the AHP. A synthetic score for each alternative is calculated based on these measures to produce the final decision. Obviously, a change of these measures may influence the final synthetic score so that the decision must be made again. The sensitivity analysis will study the issues associated with such a circumstance, possible problems are:

How stable is the final rank of alternatives when critical factors (criterion weight, performance score) are changed ?

Which criterion or alternative is most sensitive ?

How much is the amount of a change on a measure to cause the final rank reversion?

Triantaphyllou and Sánchez classified the sensitivity analysis problem into Absolute Any (AA), Absolute Top (AT), Percent Any(PA), and Percent Top(PT) [16]. The *AA* problem is to find the smallest absolute change which causes any two alternatives to reverse their existing rank. The *AT* problem wants to determine the smallest absolute change which influences the rank of the best candidate. However, in many cases, an absolute change of a measure, say 0.01, has different meaning when the original measure value is 0.07 or 0.7. Therefore, the relative change can reflect the sensitive degree of measures more reasonably. The *PA* problem is to determine the smallest relative change which makes the ranking position of any two alternatives change, while the *PT* problem is to find the smallest relative change which influences the rank of the best alternative.

Applying the approach introduced in [16] on Table 1, we obtained the relative change of each weight value in Table 9. The positive value of a change means a decrease of the original measure, while the negative value means an increase of that measure, zero denotes that two alternatives have equal ranking scores. A relative change is infeasible when this value can not meet a restraint (marked as NF in the table).

As the result, the criterion *presentation* is determined as the *PA* critical criterion, because it has the smallest percentage change value in terms of the alternative A and C in the rank. The *PT* critical criterion is *relevance*.

The analysis is also supported by experiment results: In our previous SAW example, A has a higher rank than C. If the weight of criterion *presentation* decreases by more than 15.18%, say from 0.15 to 0.124, and the weight values of all criteria are renormalized, then C is superior to A with ranking score 6.156, while As ranking score is 6.152.

The sensitivity analysis on alternatives is similar as that on criteria. The analysis results of Table 2 are shown in Table 10.

By analyzing this table, the most critical alternative in SAW can be determined as *C*, because it has the smallest percentage change value. If we increase the performance value of C in terms of the *relevance* from 6 to 6.36 (by 6%), then C is superior to A.

The sensitivity analysis method above is also suitable to the AHP approach. More detailed investigation about sensitivity analysis of SAW and AHP can be found in [**?**].

4.2 Sensitivity Analysis of the TOPSIS

In the TOPSIS case, we have a rank $B : 0.743$ $D : 0.503$ $C : 0.466$ $A : 0.396$. Different from the final rank in the SAW, the AHP, as well as the DEA, D is superior to A and C, and appears at the second position. This is because that the performance score of alternatives in terms of the most important criterion *relevance* is greatly increased by using the Euclidean distance approach in the TOPSIS, so that this value is dominant over the performance values in terms of other criteria. In addition, the performance measure of D in terms of the *relevance* is better than the performance measures of A and C related to the same criterion. This difference furthermore dominance the final ranking score. Therefore, three important factors – the difference on the performance measure in terms of the most

Table 9: Percent changes (δ') in weight values in the SAW

Pair of alternatives	availability	accessibility	correctness	completeness	relevance	presentation
B-A	-686.47	-910.89	Infinity	-1148.51	**71.29**	-607.26
B-C	NF	-250.49	-500.99	NF	NF	NF
B-D	NF	-660.39	NF	NF	NF	NF
A-C	34.32	-30.36	-45.54	38.283	-17.82	**15.18**
A-D	NF	-409.90	NF	NF	-40.10	68.32
C-D	NF	NF	53.14	Infinity	-62.38	Infinity

Table 10: Percentage changes of alternatives performance in terms of each criterion in the SAW

Pair of alternatives	availability	accessibility	correctness	completeness	relevance	presentation
B-A	98.07	NF	75.91	NF	39.60	86.75
...	...	...	...	...	...	...
C-B	-125.85	-125.25	-71.57	-210.56	-65.35	-166.99
C-A	-11.44	-11.39	-6.51	-19.14	**-5.94**	-15.18
...	...	...	...	...	...	...
D-C	-48.05	-53.14	-39.85	-66.99	-15.59	-53.14

important criterion and the Euclidean distance approach – make D to prevail over A and C.

To the best of our knowledge, there is no similar formula of sensitivity analysis for the TOPSIS approach as above for the SAW and AHP. To study the impact of the weight of the *relevance*, we give a weight-final score chart (Figure 3). In this chart, we can find four critical points, where the rank of alternatives has a change. For example, if the weight of *relevance* decreases from 0.3 to 0.263 (by 12.3%) and all weight values are renormalized, then the rank of D and C is reversed. If this weight is furthermore decreased to 0.236 (by 21.3%), then the rank of A and D is reversed.

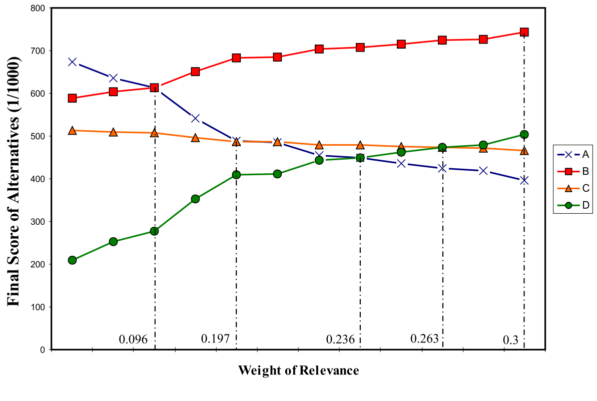

Figure 3: Weight-Final Score chart of relevance

By analyzing the sensitivity of the other criteria, the needed minimum relative changes are all greater than 12.3. For instance, if we try to increase the weight of *correctness*, the minimum relative change of this criterion is 20%, i.e., if its weight value is increased from 0.2 to 0.24 by more than 20%, the rank of D and C is reversed. Therefore, we can determine that the most sensitive criterion in the TOPSIS is *relevance*.

By analyzing the sensitivity of alternatives in terms of the most sensitive criterion *relevance* through numerical test, we can obtain the following results: the performance score of A is increased from 4 to 4.82 by 20.5%, the rank becomes BDAC; the performance score of B is decreased from 9 to 6.02 by 33.1%, the rank becomes DBCA; the performance score of C is increased from 6 to 6.39 by 6.5%; the rank becomes BCDA, the performance score of D decreases from 8 to 7.48 by 6.5%, the rank becomes BCDA. Considering the smallest relative change, C and D both are the most sensitive alternative in the TOPSIS approach.

4.3 Sensitivity analysis of the DEA

In the DEA approach, it is unnecessary to assign a weight value to each criterion as in the other three MCDM methods, the only subjective measure here is the performance score of each alternative with regard to each criterion.

In the running example in Section 3.4, the rank of alternatives is B, A, C, D. Assuming that B's measure in terms of *application requirement* is changed, the amount of the change is δ, then we have a new simplex tableau for solving LP problem of A as follows:

Applying the Simplex Solution to this table, the optimization process remains the same as in Section 3.4, and the optimal solution is not changed, because the change δ can only influence columns of z and s_1. E_A is still 1.09.

Now, we take a look at the new simplex tableau of B (Table 12):

Comparing the attributes of variable z and y in the objective function, if $8.0 - \delta > 7.0$, i.e. $0 < \delta <$

Table 11: The simplex tableau of A for sensitivity analysis

Objective function	6.5	7.5	6.0	0	0	0		
	x	y	z	s_1	s_2	s_3	Index	
	5.5	7.0	8.0-δ	1	0	0	1	s_1
	7.0	6.5	5.0	0	1	0	1	s_2
	5.5	5.0	6.0	0	0	1	1	s_3

Table 12: The simplex tableau of B for sensitivity analysis

Objective function	5.5	7.0	8.0-δ	0	0	0		
	x	y	z	s_1	s_2	s_3	Index	
	6.5	7.5	6.0	1	0	0	1	s_1
	7.0	6.5	5.0	0	1	0	1	s_2
	5.5	5.0	6.0	0	0	1	1	s_3

1, then the optimization process remains the same as in Section 3.4, because 8.0-δ makes z still having the largest positive Simplex Criteria, and the column of z is still the pivot column. The next calculation is as the same as in the previous example of the DEA approach. As the result we have optimal solution E_B = 0.071y + 0.107z = 1.35-0.107δ, the value of E_B can vary from 1.24 to 1.35. The result of new numerical calculations shows that the rank of alternatives is still B, A, C, D within such a varying scope of δ.

If $8.0 - \delta < 7.0$, i.e. $1 < \delta < 8$, then variable y will have the largest positive SC in the first iteration rather than z, this results in a new optimization process. The rank of alternative B will be changed. This analysis is confirmed by experimental results, the rank of alternatives is A, C, D, B within such a varying scope of δ.

Above is a simple analysis on the sensitivity in the DEA approach. Since the DEA employs the LP technique, the sensitivity analysis strategies in the LP area can be used in the DEA method, the involved issues include:

- changing problem coefficients for a variable
- changing the right hand side of a constraint
- adding new variables or constraints

These issues have been extensively studied in [23].

5 Comparison of Several MCDM Methods

We have applied four MCDM methods to evaluate and select Web sources for warehousing. These methods are popular in decision making activities, and have different features, such as, scoring in SAW and AHP, the weighted distances to the positive ideal solution and negative ideal solution in TOPSIS, and the Linear Programming in DEA. There is no unique best method for a MCDM problem, each approach has its strengths and limitations.

The SAW method is easy to understand and widely used. It has a simple mathematial principle and can synthetically calculate the impact of performance values of an alternative in terms of all evaluation criteria. One drawback is that the weight of criteria and performance scores of alternatives must be assigned subjectively.

The AHP is one of the most popular MCDM methods. It has solid theoretical foundation and objectivity to some degree. AHP is based on three principles: decomposition, comparative judgments, and the synthesis of priorities, and can help decision makers to develop systematic approaches for a variety of problems. However, it has several shortcomings, such as, man-made inconsistency in pairwise comparisons, rank reversal when new options or elements are introduced or important elements are omitted. Besides, $M \cdot (M-1) \cdot N$ pairwise comparisons are time consuming, if there are M alternatives and N criteria in a MCDM problem.

The TOPSIS uses the available information in a decision matrix to develop a compromise solution by explicitly defining each alternatives' best and worst characteristics. This approach provides another way for quality assessment, which is different from the SAW and the AHP. However, from the sensitivity analysis we can find that the performance score of alternatives in terms of the most important criterion has pretty great influence on the final rank. If the performance score of an alternative is dominant over the score of the other alternatives in terms of the most important criterion, this alternative is also privior to the other alternatives in the final rank, although the performance of this alternative involved with the other criteria may be worse than the other alternatives.

The DEA method is a linear programming based technique for measuring the relative performance of organizational units where the presence of multiple inputs and outputs makes comparisons difficult.

Sarkis [14] and Stewart [15] have compared the traditional goals of DEA and MCDM. The basic DEA arises from the situation where the goal is to determine the productive efficiency of a system of DMUs by comparing how well these units convert inputs into outputs, while MCDM models arise from problems of ranking and selecting from a set of alternatives that in general have conflicting criteria. A methodological connection between DEA and MCDM is to define maximizing criteria (benefits) as outputs and minimizing criteria (costs) as inputs. Identifying whether a criterion is minimizing or maximizing aids in determining whether the criterion could be considered as an input or output in the DEA model.

As discussed in section 3.4, the basic DEA approach is better viewed as a classifying tool, because it identifies alternatives as two classes: efficient and non-efficient. The extended DEA approach allows an efficiency score to be greater than 1, so that it can draw a distinction among a set of alternatives. Thus this model can be used for ranking Web sources. The DEA is viewed as an "objective" approach for evaluating different alternatives in the sense that it does not need to assign a weight to each criterion. However, it suffers of having to adapt the number of variables to the number of available constraints by assuming a zero value for some variables.

To sum up, the SAW is the simplest and the most robust method among all four approaches, and the DEA and the AHP have more objectivity than other.

6 Related Work

The problem of Web sources evaluation and selection is related to several research areas.

One direction involved is ranking Web sources. Web ranking is a method used in the Web search engines to locate relevant information on high quality Web documents. A few years ago, the occurrences of words queried in a Web page was the single main heuristic in ranking Web pages. Recently, a link analysis approach [1, 8] was introduced. The main idea of link-based approaches is that links generally signify approval of the linked document and its relevance to the topic of the linking document. Some aggregated approval ranks can be mechanically computed using some flow model when a certain kind of approval units flows along the links of the considered subset of the Web graph [9]. Besides, several other heuristics have been added, including anchor-text analysis, page structure analysis, the use of keyword listings and the URL text itself [2]. These approaches devote themselves to promptly obtaining highly precise documents over rapidly growing Web sources and enrich the technique of Web source ranking.

From the view of multi-criteria Web source selection, Web source ranking is the prelude of our work. That is, we use Web search engines to preselect several Web sources that are highly ranked and are most relevant to the subjects of a data warehouse. Then we systematically evaluate these sources using MCDM methods and select the most qualified sources as the external information sources of the data warehouse.

In the source selection area, the work of Naumann et al. [12, 13] is closely related to ours. They use the basic DEA model for quality-driven source selection. In their work, *ease of understanding, reputation, reliability and timeliness* are proposed as evaluation criteria. The advantage of their approach is the DEA method needs relatively little additional information from the *decision maker* and avoids assigning a weight to each criterion. But the basic DEA focuses only on classifying sources to good sources and non-good sources, it can not exactly rank the alternatives and give the differences among these alternatives. Different from their work, we focus on applying and comparing various kinds of MCDM approach for strict source screening. The extended DEA model discussed in our paper is more discriminating than the basic DEA model. In addition, considering the requirements of integrating Web data into a data warehouse on designing criteria can help us to develop quality measures that are source-specific as well as target-specific.

The research of Mihaila et al. [11] is also involved with source selection and ranking. They focused on maintaining metadata about source content and data quality and providing ranked data sources which meet the specified source content and quality conditions a user proposed. Four quality criteria (*completeness, recency, frequency of updates, granularity*) are adopted in their work. When a query arrived, they used SQL-like language to select those sources satisfying the conditions. In their approach, evaluation of one source is independent of the evaluation of the other sources, furthermore, all quality parameters are treated without difference. Due to the range rules in the queries, sources are only classified. In contrast, approaches discussed in our work can more systematically and more comprehensively evaluate and rank sources, thus sources can be strictly screened.

7 Conclusion

The utilization of Web sources has increasingly attracted attention. Evaluating and selecting sources of high quality is necessary for any further usage of data from the Web.

In this paper, we investigated this important aspect in a Web data warehousing environment. We analyzed *source stability*, *data quality*, and *application-specific or contextual requirements*. A set of criteria was developed for describing these dimensions. Based on these criteria, four MCDM methods are applied to evaluate and screen Web sources. The MCDM approaches discussed are highly systematic and comprehensive on assessing and selecting qualified Web sources. The limitation of most of them is that decision makers must subjectively assign a weight to each criterion or make a subjective comparison among alternatives to develop a performance score for each alternative with respect to each criterion. In view of this, we carried out the sensitivity analysis of the final rank in terms of critical measures to each approach, in order to gain an insight of the stability of the final decision.

Comparing with the other two methods, the SAW and the AHP are suitable for our Web sources evaluation and selection. Further research is needed to determine if methods that do not require the subjective assignment of weight but require other simplifying assumptions, e.g. the DEA model, can produce adequate results in the selection of Web sources. One another focus of our future work is to analyze the sensitivity of the selected qualified Web sources when several critical factors are changed jointly.

References

[1] S. Brin and L. Page: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, WWW7/Computer Networks, 30(1-7): 107-117, 1998

[2] C. Dwork, R. Kumar, M. Naor and D. Sivakumar: *Rank Aggregation Methods for the Web*, WWW10, China, 2001

[3] M. A. Selly and E. H. Forman: *Decision by Objectives*, World Scientific Publishing Co., Dec. 2001, also available in `http://mdm.gwu.edu/forman/`

[4] R. D. Hackathorn: *Web Farming for the Data Warehouse*, Morgen Kaufmann Publishers, Inc., 1999

[5] C. L. Hwang and K. Yoon: *Multiple Attribute Decision Making*, Lecture notes in economics and mathematical systems 186. Berlin: Springer-verlag, 1981

[6] H. V. Jackson JR.: *A Structured Approach for Classifying and Prioritizing Product Requirements*, doctoral thesis, North Carolina State University, 1999

[7] Z. Jandric and B. Srdjevic: *Analytic Hierarchy Process in Selecting Best Groundwater Pond*, 31^{st} International Geological Congress, Brazil, 2000

[8] J. Kleinberg: *Authoritative Sources in a Hyperlinked Environment*, SODA98, USA, 1998

[9] M. Lifantsev: *Voting Model for Ranking Web Pages*, Proceedings of the International Conference on Internet Computing, USA, 2000

[10] C. McMillan, Jr.: *Mathematical Programming: An Introduction to the Design and Application of Optimal Decision Machines*, John Wiley & Sons, Inc., 1970

[11] G. A. Mihaila, L. Raschid, M.-E. Vidal: *Using Quality of Data Metadata for Source Selection and Ranking*, WebDB00, USA, 2000

[12] F. Naumann, J. C. Freytag, and M. Spiliopoulou: *Quality-driven Source Selection Using Data Envelopment Analysis*, IQ98, USA, 1998

[13] F. Naumann, U. Leser and J. C. Freytag: *Quality-driven Integration of Heterogeneous Information Systems*, VLDB99, Scotland, 1999

[14] J. Sarkis: *A Comparative Analysis of DEA as a Discrete Alternative Criteria Decision Tool*, European Journal of Operational Research, 123(2000): 543-557

[15] T. J. Stewart: *Relationships between Data Envelopment Analysis and Multicriteria Decision Analysis*, Journal of the Operational Research Society, 47(5): 654-665, 1996

[16] E. Triantaphyllou and A. Sánchez: *A Sensitivity Analysis Approach for some Deterministic Multi-Criteria Decision Making Methods*, Decision Sciences, 28(1): 151-194, 1997

[17] E. Triantaphyllou, B. Shu, N. Sanchez, and T. Ray: *Multi-Criteria Decision Making: An Operations Research Approach*, Encyclopedia of Electrical and Electronics Engineering, 15: 175-186, 1998

[18] R. Wang: *A Product Perspective on Total Data Quality Management*, Communications of the ACM, 41(2): 58-65, 1998

[19] Y. Zhu, C. Bornhövd and A. P. Buchmann: *Data Transformation for Warehousing Web Data*, WECWIS01, USA, 2001

[20] Y. Zhu, C. Bornhövd, D. Sautner, and A. P. Buchmann: *Materializing Web Data for OLAP and DSS*, WAIM00, China, 2000

[21] Y. Zhu: *Integrating External Data from Web Sources into a Data Warehouse for OLAP and Decision Making*, doctoral thesis, 2002

[22] Y. Zhu: *A Framework for Warehousing the Web Contents*, ICSC99, China, 1999

[23] S. Zionts: *Linear and Integer Programming*, Prentice-Hall, Inc., 1974