Data Fusion and Data Quality

Felix Naumann[1]

Institut für Informatik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin

Summary

The recent development of the Internet has made an increasing number of information sources available to users. This makes it necessary to submit queries only to the most appropriate sources. When gathering and combining information from these sources the quality offered can and must be a criterion for source selection. However, information quality has many dimensions and it is thus difficult to directly compare sources with one another or give a ranking of sources. Selecting the best sources is thus a multiple attribute decision problem.
After introducing four techniques of multiple attribute decision making we apply them to the problem of quality-driven selection of sources: The Simple Additive Weighting method (SAW), the TOPSIS method, the Analytical Hierarchy Process method (AHP) and the Data Envelopment Analysis method (DEA). We analyze and compare these methods with respect to the assumptions they are based on, their discretionary value and user interaction.

1 Motivation

The development of the Internet and the World Wide Web during recent years has made it possible and useful to access many different information

[1]This research was supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316)

systems anywhere in the world to obtain information. For a certain piece of information a user can typically choose from many similar information sources with partially overlapping content and must decide which ones to query, due to time and cost constraints. These sources show a varying quality of information and a varying quality of access, i.e., varying query cost.

Given a query and a set of information sources, that are capable of answering the query to some extent, we address the problem of deciding which of the sources to query. Source selection is usually tackled in a straightforward manner: Some selector component determines the most relevant sources using statistical information. We maintain that there is more to finding the best sources than counting the appearances of certain words. Quality of a source and the quality of the documents it contains must be measured with more than one criterion or quality dimension. Source selection is thus a multi-attribute decision making problem.

2 The Data

We have chosen three distinct quality criteria from a collection empirically gathered by Wang and Strong (1996) to define information quality. To these quality criteria come two cost criteria, which we find to be the most important for WWW-information retrieval. The data is not given in a common unit nor in a common range of values. It will be the task of the decision making methods to deal with these difficulties. After a short description of the criteria an example data set is given which will be used to demonstrate the methods.

- UNDERSTANDABILITY measures how well a source presents its information, so that the user is able to comprehend its semantic value. User understandability is given as a score between 1 and 10.
- EXTENT of a given set of information is the average length of the single pieces of information, for instance the number of columns in a table.
- AVAILABILITY of an information source is the probability that a feasible query is correctly or at least satisfyingly answered in a given time range. This statistical value is given as a percentage.
- TIME measures the number of seconds that pass between submission of the query and the response of the information source.
- PRICE is the monetary cost of a query.

The index-set of the quality criteria is $J_Q = \{1, 2, 3\}$, that of the cost criteria is $J_C = \{4, 5\}$. We assume the existence of *linear utility functions*, i.e., the higher a quality score or the lower a cost score is, the higher this is valued by the user. Another important assumption is the *preference independence* of the criteria. The above criteria are preference independent but would

fail to be so if for instance the criterion RESPONSE SIZE was added. When criteria are not preference independent, the user should try to combine or drop criteria.

The measured or otherwise determined scores for each of these criteria and each source are given in a decision matrix, eg., matrix $D = (d_{ij})_{i,j=1,\dots,5}$ with fictious scores: Most decision making methods also require a weight-vector W, which reflects the importance of the individual criteria. We choose $W = (w_1, \dots, w_5)$ so that $\sum_{j=1}^{5} w_j = 1$. Using this input data we describe in brief four decision making methods.

$$D = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \begin{array}{c} \begin{array}{ccccc} \text{U} & \text{E} & \text{A} & \text{T} & \text{P} \end{array} \\ \begin{pmatrix} 5 & 22 & 20 & 5 & 0.5 \\ 3 & 18 & 99 & 180 & 10 \\ 10 & 10 & 50 & 10 & 0 \\ 3 & 12 & 55 & 3 & 1 \\ 10 & 10 & 35 & 10 & 0.1 \end{pmatrix} \end{array} \qquad W = \begin{pmatrix} \frac{18}{66} \\ \frac{9}{66} \\ \frac{6}{66} \\ \frac{22}{66} \\ \frac{11}{66} \end{pmatrix}$$

3 Four Decision Making Methods

3.1 Simple Additive Weighting (SAW)

The SAW method is one of the most simple but nevertheless good decision making methods, in that its results are usually very close to more sophisticated methods (Hwang and Yoon 1981). The method is comprised of three basic steps: Scale the scores to make them comparable, apply weights and sum up the values for each source.

Since we allow zero-values for our scores we apply the scaling factors

$$\begin{aligned} v_{ij} &= \frac{d_{ij} - d_j^{\min}}{d_j^{\max} - d_j^{\min}} \quad \text{for quality criteria and} \\ v_{ij} &= \frac{d_j^{\max} - d_{ij}}{d_j^{\max} - d_j^{\min}} \quad \text{for cost criteria} \end{aligned}$$

With this scaling all scores are in $[0, 1]$, the best score of any criterion obtains the value 1, and the worst score of any attribute obtains the value 0. This property assures comparability of scores. The final preference score for each source is $S_i = \sum_{j=1}^{5} w_j v_{ij}$, so the SAW-ranking is: 1. S_3: 0.7941; 2. S_5: 0.7751; 3. S_1: 0.7022; 4. S_4: 0.5463; 5. S_2: 0.1818.

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

The TOPSIS method was developed by Hwang and Yoon (1981). As with SAW the decision matrix is scaled and weighted. For the scale function the authors use

$$v_{ij} = \frac{d_{ij} \cdot w_j}{\sqrt{\sum_{i=1}^{5} d_{ij}^2}}$$

so that each criterion-vector is normalized. Unlike in SAW the authors now do not sum up the values, but rather calculate the relative Euclidean distance of the sources from a fictious ideal source. The source closest to the ideal source and furthest from the negative-ideal source is chosen best. The ideal and negative ideal solutions are

$$\begin{aligned} A^* &= (v_1^*, \ldots, v_5^*) := \{(\max_{i=1,\ldots,5} v_{ij} | j \in J_Q), (\min_{i=1,\ldots,5} v_{ij} | j \in J_C)\} \\ &= (0.175, 0.0884, 0.0691, 0.0055, 0) \\ A^- &= (v_1^-, \ldots, v_5^-) := \{(\min_{i=1,\ldots,5} v_{ij} | j \in J_Q), (\max_{i=1,\ldots,5} v_{ij} | j \in J_C)\} \\ &= (0.0525, 0.0402, 0.014, 0.3321, 0.1656) \end{aligned}$$

The Euclidean distance between each source and the ideal and negative ideal solution is defined as $S^{(*|-)}(S_i) := \sqrt{\sum_{j=1}^{5}(v_{ij} - v_j^{(*|-)})^2}$ and the relative closeness of a source to the ideal is defined as $C^*(S_i) := \frac{S^-(S_i)}{S^*(S_i)+S^-(S_i)}$. The results ranked by $C^*(S_i)$ are 1. S_3: 0.8641; 2. S_5: 0.8483; 3. S_1: 0.7782; 4. S_4: 0.7293; 5. S_2: 0.1417.

3.3 Analytical Hierarchy Process (AHP)

The AHP-method was developed by Thomas Saaty (1980). It is composed of four main steps: Development of a goal hierarchy, pairwise comparison of goals, consistency check of the comparisons and aggregation of the comparisons.

The goal hierarchy of source selection is depicted in Figure 1. The main goal of source selection is *user satisfaction*, split up into a *quality* and a *cost* goal. Each subgoal is made up of several criteria. The bottom level consists of the information sources.

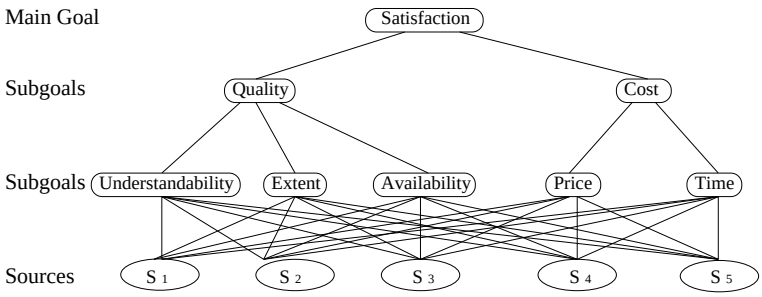

Figure 1: Goal Hierarchy of AHP

To represent the pairwise comparisons of goals, comparison matrices are defined for the main goal and for each subgoal of the hierarchy. The matrix entries for *satisfaction*, *quality* and *cost* reflect the weighting, the matrix

entries for the five criteria reflect the measured scores. Matrix entries are between 1 (same importance) and 9 (very much more important) or their complement values. For instance the 2×2 matrix for the subgoal *cost* shown below reflects that TIME is a bit more important than Price. The 5×5 matrix for TIME reflects for instance that source S_1 has a much better TIME-score than S_2.

$$\begin{array}{c} \\ \mathrm{P} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{P} & \mathrm{T} \end{array} \\ \begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5 \end{array} \\ \begin{pmatrix} 1 & 9 & 2 & 1 & 2 \\ \frac{1}{9} & 1 & \frac{1}{8} & \frac{1}{9} & \frac{1}{8} \\ \frac{1}{2} & 8 & 1 & \frac{1}{3} & 1 \\ 1 & 9 & 3 & 1 & 3 \\ \frac{1}{2} & 8 & 1 & \frac{1}{3} & 1 \end{pmatrix} \end{array}$$

We omit a description of the next step, the consistency check, which basically gives hints on which comparisons are transitively inconsistent.

For the aggregation step a weight vector for each goal and subgoal is calculated. It is the normalized eigenvector of the maximum eigenvalue of the matrix. The weight-vector for TIME-matrix is $(0.299, 0.027, 0.157, 0.36, 0.157)$ to the maximum eigenvalue 5.118. Each arc of the hierarchy is then labeled with the value determined by the weight-vector. The final preference score of each source is calculated as the weighted sum along paths from that source to the main goal. For our example the AHP-ranking is 1. S_3: 0.2476; 2. S_5: 0.2215; 3. S_1: 0.2114; 4. S_4: 0.2011; 5. S_2: 0.1185.

3.4 Data Envelopment Analysis (DEA)

Unlike the previous methods, DEA does not deliver a total ranking of sources but rather suggests which sources are better than others. It was introduced by Charnes, Cooper, and Rhodes (1978) as a general method to classify a population of observations. The DEA method determines the efficiency of each source S_{j_0} separately by solving a linear program:

LP-DEA :

$$\begin{array}{lrl} \text{maximize} & w_1 \cdot d_{1j_0} + w_2 \cdot d_{2j_0} + w_3 \cdot d_{3j_0} - w_4 \cdot d_{4j_0} - w_5 \cdot d_{5j_0} & \\ \text{subject to} & w_1 \cdot d_{1j} + w_2 \cdot d_{2j} + w_3 \cdot d_{3j} - w_4 \cdot d_{4j} - w_5 \cdot d_{5j} & \leq 1 \ \forall \ S_j \\ & w_4 \cdot d_{4j_0} + w_5 \cdot d_{5j_0} & = 1 \\ & w_1, w_2, w_3, w_4, w_5 & \geq \epsilon > 0 \end{array}$$

With DEA the criterion weights are not specified by the user but rather are variables that are determined by the method. Therefore neither a scaling nor a user weighting are necessary. The optimal value of such an LP is either 1, i.e., the source is efficient, or a value between 0 and 1, determining the degree of inefficiency. For our example the optimal values are 1. S_1: 1; 2. S_3: 1; 3. S_4: 1; 4. S_5: 0.9849; 5. S_2: 0.947.

4 Comparison of Methods

In the following paragraphs we discuss selected properties of the decision making methods. We summarize these properties in Table 2.

Interaction User interaction, i.e., the necessity of the user to state preferences or compare alternatives, is undesirable because time consuming as long as the final response to a query is fully satisfying. User interaction should be confined to a one-time statement of preferences between criteria and possibly between sources.

This requirement is immediately met by the first two methods. The AHP-method additionally requires comparison matrices for each criterion. As these scores (usually) are cardinal this comparison can be done automatically without the user. The DEA-method even goes a step further by not even demanding (nor allowing) preferences between the criteria.

Weighting It is in the nature of quality and cost criteria that they have different importance to each user. Usually the weighting is given as a vector, which multiplied with the decision matrix gives the user-adjusted matrix. The AHP method does not require a vector but rather asks for pairwise comparison of the criteria. The DEA-method differs: No weighting is required and in fact any general weighting applied to the matrix would have no effect on the results of the method.

Dominance An information source dominates another, if it is equal to or better than the other source in all criteria and strictly better in at least one criterion. Source S_3 of our example dominates S_5. Obviously a dominating source is more preferable than the dominated source and any decision making method should be able to discover dominance and rank a dominating source above the dominated one with any given weighting. The four examined examples all discover dominance.

Scaling Measured scores of different criteria are scaled to make them comparable. There are basically two ways to scale the scores – scale transformation (SAW) and normalization (TOPSIS, AHP). The transformation used in SAW leads to a good comparability, however the transformation is not linear. Normalization has the advantage of scaling proportionally but the disadvantage of not considering the maximum or minimum values of a criterion. Thus the maximum score in one criterion might have a lower normalized value than the maximum of another criterion.

Result Type We distinguish two result types – a total ranking of sources and a classification of sources. Decision making methods that result in a ranking determine some overall score for each source and rank the sources

by that score. Querying the top-ranked source should result in the best response. The rank-scores suggest an discretionary power that in general does not correspond with reality. The DEA method delivers a more general classification of the sources.

The rankings or classifications produced by the individual methods are summarized in Table 1. In our example the first three methods all deliver the same ranking, whereas the DEA-method delivers a slightly different classification, due to the lack of any weighting of the criteria.

	SAW	TOPSIS	AHP	DEA
1.	S_3	S_3	S_3	S_1, S_3, S_4
2.	S_5	S_5	S_5	
3.	S_1	S_1	S_1	
4.	S_4	S_4	S_4	S_5
5.	S_2	S_2	S_2	S_2

Table 1: Rankings of Decision Making Methods

	SAW	TOPSIS	AHP	DEA
Interaction	weighting	weighting	decisions	%
Criterion weighting	vector	vector	scaled comparison	%
Dominance detection	$\surd$	$\surd$	$(\surd)$	$\surd$
Scaling function	$\frac{d_{ij}-d_j^{\min}}{d_j^{\max}-d_j^{\min}}$	$\frac{d_{ij} \cdot w_j}{\sqrt{\sum_{i=1}^{5} d_{ij}^2}}$	norm. eigenvector	%
Result type	ranking	ranking	ranking	classification

Table 2: Selected Properties of Decision Making Methods

5 Quality of Fusion

After querying multiple information sources the separate results are merged to a single response to the user. The quality and cost of the merged result will in general differ from the quality and cost of the individual sources. The following table gives a fusion function for each introduced criterion through which the criterion score for the merged result can be calculated. Using these functions a source selection method can be applied to all combinations of sources to find the best combination to query. These fusion functions are basic guidelines for new criteria and also should be revised for different application domains.

- Fused UNDERSTANDABILITY is calculated as the average of the scores weighted by EXTENT: $\sum_{i=1}^{n} \text{UND.}(S_i) \cdot \text{EXT.}(S_i) / \sum_{i=1}^{n} \text{EXT.}(S_i)$
- Fused EXTENT is the extent of the union of the individual sources: $\text{EXT.}(S_1 \cup \ldots \cup S_n)$. The union of the sources is determined by the way the information is merged.
- Fused AVAILABILITY is the probability that all queried sources are available: $\prod_{i=1}^{n} \text{AVAIL.}(S_i)$
- Fused TIME is the maximum of the individual scores: $\max_i [\text{TIME}(S_i)]$
- Fused PRICE is the sum of the individual prices: $\sum_{i=1}^{n} \text{PRICE}(S_i)$

6 Conclusion and Outlook

Each of the reviewed decision making methods has advantageous and disadvantageous properties. The SAW and the TOPSIS method should be chosen if only a few criteria are taken into account and only a small number of information sources are queried. If more criteria are reviewed, the TOPSIS method should be chosen as it facilitates the user weighting process. If a large number of sources is available or if new sources are added DEA is a good method to automatically preselect the best sources which can then be further analyzed by one of the other methods.

Further research will add more criteria to the presented catalog and develop methods to determine the individual criteria scores. For data fusion only the simple merging of information of different sources has been considered. We plan to extend the model to also allow the response of one source be input into the query of another. This implies data dependencies which will effect quality and cost scores.

References

Charnes, A., W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research 2*, 429–444.

Hwang, C.-L. and K. Yoon (1981). *Multiple Attribute Decision Making*. Number 186 in Lecture Notes in Economics and Mathematical Systems. Berlin/Heidelberg/New York: Springer.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill, Inc.

Wang, R. Y. and D. M. Strong (1996). Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems 12, 4*, 5–34.