Quality-driven Source Selection using Data Envelopment Analysis

Felix Naumann[*†]
Humboldt-Universität zu Berlin
naumann@dbis.informatik.hu-berlin.de

Johann Christoph Freytag
Humboldt-Universität zu Berlin
freytag@dbis.informatik.hu-berlin.de

Myra Spiliopoulou
Humboldt-Universität zu Berlin
myra@wiwi.hu-berlin.de

Abstract

Due to the development of the Internet there is an increasing number of information sources available to users. This makes it necessary to query only the most appropriate sources. The information quality offered by these sources can and must be a criterion for source selection. However, information quality has many dimensions, both subjective and objective, and it is thus difficult to directly compare sources with one another or give a ranking of sources.

We propose to use Data Envelopment Analysis (DEA) to solve these shortcomings. This method does not directly compare sources, but focuses on individual sources, determining their efficiency in terms of information quality and cost. We adapt DEA to the task of source selection and discuss its advantages over other methods. Furthermore, we expand the model to include cost criteria and suggest a solution to additionally select sources in a query-dependent way.

1 Motivation

The Situation The development of the Internet and the World Wide Web during recent years has made it possible and useful to access many different information systems anywhere in the world to obtain information. For a certain piece of information a user can typically choose from many similar information sources. Due to time and cost considerations it must be decided which ones to query, either by the user or by an automated process.

[*] Contact author: Felix Naumann, Institut für Informatik dbis, Humboldt-Universität zu Berlin; Unter den Linden 6, 10099 Berlin, Germany; Phone: +49 30 2093-3025

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

Hence, we assume multiple information sources with partially overlapping content which can be uniformly queried by some system. The sources may be document collections or WWW sites, for instance with an address information service. The information sources show a varying quality of information and a varying quality of access, i.e. varying query cost.

The Problem Querying information from Internet sources is usually divided into the three tasks of (i) source selection, i.e. choosing the best possible information sources to evaluate a query, (ii) query evaluation at the sources, an (iii) merging the query results (Gravano, Chang, and Garcia-Molina 1997). 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 issue the query upon. This decision should be based on information quality, as without considering quality, a system might return an answer that is useless, harmful or at least unsatisfying.

Source selection is usually tackled in a straightforward manner: Some selector-component analyzes source capabilities and source contents. Matching the query against the capabilities of the sources determines which combinations of sources are *capable* of answering the query. Matching the query against the source contents determines the sources that will probably provide the most and the most *relevant* information. This technique relies on statistical information giving the total number of appearances of each distinct word, the document frequency for each word, and the total number of documents in source. With this information the appropriateness of each source for evaluating the query can be estimated. The sources with the highest estimates are then chosen to be queried. An information source is thus considered to be appropriate if the keywords of a query appear often and in many documents stored by the source.

We maintain that there is more to finding the best sources than counting the appearances of certain words. Consider a source containing not only text documents but also explanatory graphics on the subject. Should this source be valued higher than one without graphics? Consider a source that seems to match the query well but the documents are very short, hardly containing anything but the keywords of the query. Should these documents be retrieved? Finally, consider a source containing matching but outdated information. Is such information of interest? Quality of a source and the quality of the documents it contains must be measured with more than one criterion or quality dimension.

Multiple criteria typically have different units: Some criterion may be measured in a monetary unit, others are measured as subjective grades given by a user etc. Multiple criteria also typically have different scales: One criterion may have scores between 1 and 10, another from 0 to 1,000 etc. Some criteria may be rated positive (quality), others may be rated negative (cost). Comparing sources using multiple criteria typically poses the difficulty of finding a general weighting for all criteria, which levels these differences.

The Solution We propose to use Data Envelopment Analysis (DEA) for source selection. DEA solves these problems by not focusing on all information sources at once to find a general weighting, but rather finds an optimal weighting for each individual source. The result of DEA is not a total ranking of the sources, as a general weighting would return, but an efficiency-classification of the sources in terms of information quality. While this is not as discretionary as a ranking, the DEA method does make a powerful statement on which sources to prefer, as we will show in the following sections.

Section 2 discusses related efforts to find the best sources for a query. In Section 3 we then introduce three quality criteria which will be considered in the source selection process described in Section 4. There the general DEA method is described and then applied to our problem. We then discuss shortcomings of the method and how they can be alleviated in the context of our problem. In Section 5 we introduce two extensions of the DEA method – cost criteria and query dependent criteria.

2 Related Work

Source Selection Several research projects have focussed on the problem of source selection. In the **GlOSS** system (Gravano, Garcia-Molina, and Tomasic 1994), the authors assume that each participating source provides information on the total number of documents in the source and for each word the number of documents it appears in. These values are used to calculate the estimated percentage of query-matching documents in a source. The source with the highest percentage is selected for querying.

In (Liu and Pu 1997) the authors propose a metadata approach to identify relevant and capable information sources. For each query the *query scope* and the *query capacity* are determined. The query scope describes synonyms for each part of the query; the query capacity describes the information source capability requirements for each part of the query. This metadata is matched with the source capability profiles of the information sources, which describe category, content and capabilities of a source. Levy et al. follow a similar method of describing source content and source capability (Levy, Rajaraman, and Ordille 1996).

Florescu et al. attempt to describe quantitatively the contents of information sources using probabilistic measures (Florescu, Koller, and Levy 1997). In their model two values are calculated: *Coverage* of information sources, determining the probability that a matching document is found in the source, and *overlap* between two information sources, determining the probability that an arbitrary document is found in both sources. These probabilities are calculated with the help of word-count statistics. This information is then used to decide which sources to query in which order.

To sum up, decisions on which sources to query are typically based on only one criterion: word counting. None of the above methods takes quality criteria into account when selecting sources.

Information Quality There is much research showing the importance of information quality for businesses and users (Wang and Strong 1996; Redman 1998) and many techniques have been proposed to improve and maintain quality of individual information sources (Wang 1998). To compare the measured quality of different sources we propose Data Envelopment Analysis.

To measure information quality metainformation is needed. In the **STARTS** proposal (Gravano, Chang, and Garcia-Molina 1997) a general list of required metadata fields describing information sources has been suggested. These include quality-relevant fields such as `Date/time-last-modified` and `DateExpires`. Additionally a `Source content summary` is required, containing statistical information such as the number of documents in the source, word-counts and the like.

By adding metadata tags to the database schema at different levels, Wang et al. model data quality by objective and subjective criteria in an ER-based way (Wang, Kon, and

Madnick 1993). A set of premises is established, concerning data quality requirements of users and data quality in applications, showing the need to incorporate quality information into the data itself.

3 Quality Criteria for Information Sources

There is no common definition or measure for information quality or for the quality of an information source, apart from general notions as "fitness for use" (Tayi and Ballou 1998). Rather, quality is conceived as some aggregated value of multiple subjective and objective criteria. These criteria usually have different units or no units at all and are often user dependent.

3.1 Three Quality Criteria

Wang and Strong have identified fifteen quality criteria and have classified these into the four categories "intrinsic quality", "accessibility, "contextual quality", and "representational quality" (Wang and Strong 1996). Out of these we choose three exemplary criteria, one from each of the three last categories: UNDERSTANDABILITY, EXTENT and AVAILABILITY. In the following paragraphs, we explain the usage of those criteria in the example context of information on telephone numbers and addresses of people and companies. For presentational reasons we deal with these three criteria/dimensions only. However, the model is neither restricted to this number, nor to only evaluating positive quality aspects as we will show in Section 5.

UNDERSTANDABILITY measures how well a source presents its information, so that the user is able to comprehend its semantic value. UNDERSTANDABILITY can for instance involve the spoken language used by the information source compared to the ability of the user in speaking that language. The existence of a help-function, a documentation explaining the information, or explanatory graphics also may increase quality. In an address information system the design of the input form into which users enter search terms is for instance of great importance for this criterion. User understandability is a subjective criterion. We measure it as a score between 1 and 10.

EXTENT of a given set of information is the average length of the single pieces of information. A piece of information can be a document, data about a person, a WWW link etc. Their EXTENT can be the average number of fields or the average number of words over the documents etc. It is an objective measure for an application domain.

With this definition we do not measure the response size, nor the number of pieces of information in a response, nor the extent of a whole information source. The EXTENT of address and phone information would be measured as the number of fields or attributes for each person. For instance a source providing email addresses along with the usual address information would have a higher EXTENT than one without this additional field.

AVAILABILITY of an information source is the probability that a feasible query is correctly or at least satisfyingly answered in a given time range. For example, if an information source fails during query processing, an incomplete answer might still satisfy the user. AVAILABILITY can be measured with the help of statistics derived from previous queries to the information source and is thus an objective measure. Additionall knowledge of the technical equipment and software of the information source can help determine AVAILABILITY. It is usually given as the percentage of time that source a source is “up”.

3.2 Difficulties determining overall Quality

For the above and any other possible quality criteria we assume a high score to be better than a low score. This is clear for some criteria – a high AVAILABILITY is obviously better than a low AVAILABILITY – but is arguable for others such as the EXTENT. The assessment of the individual quality criteria using metadata or actually measuring the values is beyond the scope of this study. We assume to have computed the values for our three criteria for each of five different address information sources S_1 through S_5. Some fictive example values are shown in Table 1. They are drawn from the example of information sources providing address and telephone information.

To interpret these scores it would make no sense to simply sum up the results for each S_i or calculate the average, since each criterion is measured by a different unit. Another solution

	Underst.	Extent	Avail.
S_1	5	22 flds	20%
S_2	3	18 flds	99%
S_3	10	10 flds	50%
S_4	3	12 flds	55%
S_5	10	10 flds	35%

Table 1: Quality Scores

would be to normalize the results for each criterion, forcing the values into a range between 0 and 1, and to then find an average. This being straightforward for Understandability and Availability (divide by 10 and 100 respectively), it is not so obvious for the number of fields.

Even suppose one has found a method and has normalized each score, the next problem is to find a global weighting for each criterion, as one criterion may be of more importance than another. One source may provide a high Extent of information, but at low Availability, another source may have a high Availability but a small Extent. Which is to prefer? A solution to these problems is provided by the Data Envelopment Analysis method, introduced in the following section. With this method normalization and a global weighting are not necessary.

4 Data Envelopment Analysis

Before applying the Data Envelopment Analysis (DEA) method to the source selection problem in Section 4.2, we provide a brief overview of DEA and its common applications.

4.1 An Overview

DEA was first introduced by Charnes, Cooper and Rhodes in (Charnes, Cooper, and Rhodes 1978) as a general method to classify a population of observations. It was developed to compare the efficiency of several decision making units (DMUs) and was thus designed as a decision support tool for business management. It has since been applied to many fields, for instance comparing efficiency of hospitals, airlines or even baseball players. In the scope of our study, the DMUs are the individual information sources, as will be explained in the following section.

The DEA method differs from other comparison methods in that it does not focus on the complete set of data, but rather on individual DMUs. It determines the optimal weighting of the dimensions for each DMU, thus developing a discrete piecewise frontier of efficient DMUs, called "efficiency envelope". Those DMUs *on* the frontier are called "efficient", those *below* "inefficient", as shown in Figure 1.

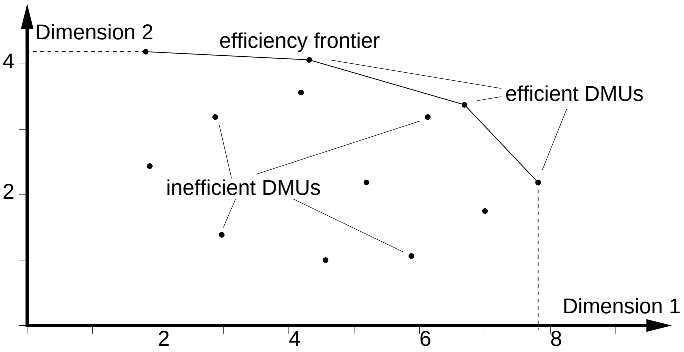

Figure 1: Efficiency frontier of DEA

To determine this efficiency envelope, a linear program (LP) is formulated and subsequently solved for each DMU. While the constraints of the LPs remain the same, the objective function is fitted to each DMU. A common way to solve LPs is the simplex method, developed by Dantzig (Dantzig 1963). The individual parts of the LP will be discussed in detail in the context of source selection in the next section.

4.2 DEA for Source Selection

The decision making units of the DEA method for the problem of source selection are the information sources. We determine the scores for each quality- (and later on cost-) dimension and for each source, e.g. using the values of Table 1 on page 7. We define as efficiency of an information source the *weighted sum of its quality scores.* Then the DEA method determines the efficiency of each source S_{j_0} separately by solving the linear program LP-Quality shown below. In this LP, the variable w_i is the weight for criterion i; the coefficient q_{ij} is the measured quality score of criterion i for source j.

LP-Quality :

$$\begin{array}{lcccccll} \text{maximize} & w_1 \cdot q_{1j_0} & + & w_2 \cdot q_{2j_0} & + & w_3 \cdot q_{3j_0} & & \\ \text{subject to} & w_1 \cdot q_{1j} & + & w_2 \cdot q_{2j} & + & w_3 \cdot q_{3j} & \leq 1 & \text{for all sources } S_j \\ & & & & & w_1, w_2, w_3 & \geq \varepsilon > 0 & \end{array}$$

To better explain the semantics of the LPs, we expand LP-Quality into LP-QualityS1 for source S_1:

$$
\begin{array}{lrcrcrl}
\multicolumn{7}{l}{\textsf{\textbf{LP-QualityS1}}:} \\
\text{maximize} & w_1 \cdot 5 & + & w_2 \cdot 22 & + & w_3 \cdot 20 & \\
\text{subject to} & w_1 \cdot 5 & + & w_2 \cdot 22 & + & w_3 \cdot 20 & \leq 1 \\
& w_1 \cdot 3 & + & w_2 \cdot 18 & + & w_3 \cdot 99 & \leq 1 \\
& w_1 \cdot 10 & + & w_2 \cdot 10 & + & w_3 \cdot 50 & \leq 1 \\
& w_1 \cdot 3 & + & w_2 \cdot 12 & + & w_3 \cdot 55 & \leq 1 \\
& w_1 \cdot 10 & + & w_2 \cdot 10 & + & w_3 \cdot 35 & \leq 1 \\
& & & & & w_1, w_2, w_3 & \geq \varepsilon > 0
\end{array}
$$

In LP-QualityS1, the objective function attempts to maximize the efficiency of source S_1. Following intuition, it does so by adjusting the weights for the quality criteria, so that the criteria where S_1 scores well obtain high weights. The five following inequations state that the weights cannot obtain arbitrary values, though. In particular, the weights must be such that the efficiency *of each source* does not exceed 1. Since all values are positive, the efficiency range is $[0, 1]$. Note that there is one inequation constraint for each source.

The last constraint states that the weights cannot obtain values lower than a threshold ε. This prevents the objective function from nullifying a criterion for which the source S_1 does not score well.

The solution of LP-QualityS1 determines an efficiency of 1 for information source S_1, thus S_1 is efficient. The same LP is solved several times, once for each of the remaining four information sources, only changing the objective function but keeping the same constraints. The (rounded) solutions with efficiency and optimal weightings to all LPs are summarized in Table 2 (with $\varepsilon = 0.001$).

	Efficiency	UNDERST.	EXTENT	AVAIL.
		w_1	w_2	w_3
S_1	1	0.001	0.0443	0.001
S_2	1	0.001	0.001	0.0099
S_3	1	0.0551	0.001	0.001
S_4	0.6896	0.0551	0.0303	0.0029
S_5	0.9947	0.0653	0.0297	0.001

Table 2: Optimal solutions for each source

Interpreting the Results. Sources S_1, S_2, and S_3 each have an efficiency of 1, i.e., for each there exists a general weighting making them equal to or better than all other sources.

Consider for instance S_3: The optimal weighting found for S_3 applied to its quality scores gives an efficiency of 1 for S_3 and an efficiency ≤ 1 for all other sources. Sources S_4 and S_5 on the other hand have efficiency scores lower than 1, i.e., there is *no* weighting giving them an efficiency of 1 without violating the constraints.

Solving LP-Quality for each information source can be interpreted graphically, as seen in Figure 2. The three dimensions in the graph represent the three quality criteria. The five information sources are positioned within this space. By solving the five LPs the efficiency envelope around these sources is constructed – the shaded area. Those sources on the envelope are efficient, those beneath, such as S_4, are not efficient.

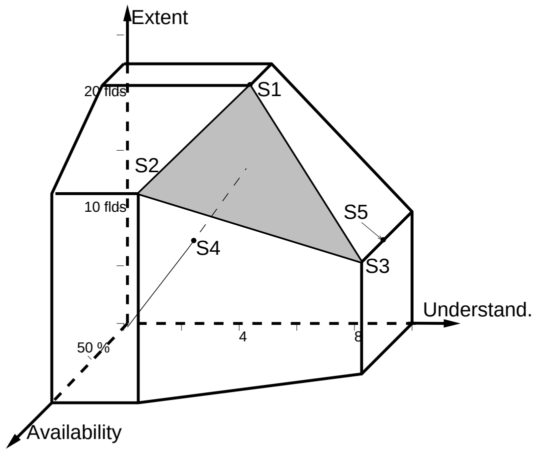

Figure 2: Comparing sources with DEA

4.3 Computational Aspects

A complete Data Envelopment Analysis involves solving n linear programs, where n is the number of sources.

The value of n is rather small: The number of sources accessed by a meta-crawler hardly ever exceeds 20 and is usually lower than 10. The archives with information relevant to an application domain mostly range in tenths, not in thousands. So, the computation of the LPs for source selection incurs a low overhead and can be performed at run time, i.e., upon each user query.

For applications issuing queries against a large number of sources, e.g., querying all computer science departments in Europe for a certain curriculum subject, the number n may increase prohibitively for run-time evaluation. In such rather unusual cases, DEA should be performed off-line. This is advisable also for the nature of the application: if queries against hundreds of information sources are issued often, the most appropriate sources for the typical queries should be identified prior to run-time.

The computational overhead of solving n LPs can be reduced in both cases by closely observing the interdependencies among the LPs. In particular, we show here that it is possible to (i) completely avoid computing some LPs and (ii) accelerate the computation of others.

Omitting Computations An LP does not have to be solved if it is possible to determine beforehand whether its corresponding information source is efficient or inefficient: As proven by Ali in (Ali 1993), sources that have the single best score for a certain criterion are always efficient. This effect can be seen in Table 1, where sources S_1 and S_2 both have the single best score in one of the quality criteria. An optimal solution to each LP will put as much weight as possible onto that one criterion, and as little weight as possible, i.e., the minimal weight ε onto the others.

Inefficient sources on the other hand, can easily be detected, if they are *dominated* by another source, i.e., if there is a source that is equal to or better than the dominated source in all criteria and strictly better in at least on criterion. In our example, S_4 is dominated by S_2 as can be verified in Table 1.

Accelerating Computations After the preprocessing described above, certain LPs are ruled omitted from computation and we have a set of efficient and a set of inefficient sources. These sets can be used to accelerate the simplex method for the remaining LPs in several ways, which are described thoroughly in (Ali 1993) and omitted here due to lack of space. Further improvement is possible due to the fact that for each LP only the objective function changes while the constraints remain the same.

4.4 Discretionary Power of DEA

DEA provides an elegant way to compare DMUs using more than one criterion and using criteria of different units. Since the focus lies on individual DMUs, it is not necessary to find any general weighting and apply it to the criteria for all DMUs. Rather, an optimal weighting is found for each individual DMU, under certain conditions.

A shortcoming of DEA is that it only determines a set of efficient sources without actually ranking them by efficiency. In our example we found three efficient sources for both LP-Quality and LP-QualityCost. DEA offers no way of comparing these sources to each other.

A solution to this problem is suggested by Dyson et al. in (Dyson, Thanassoulis, and Boussofiane 1990), who add individual constraints on the weights to the LP. We have already introduced the constraints $w_i \geq \varepsilon > 0$ to ensure that no criterion can be totally nullified. ε can be individually modified for each criterion. On the other hand, to avoid the effect of one criterion being overrated and the others chosen as little as possible, an upper limit can be introduced individually for each criterion. Choosing these upper and lower limits however is a difficult task, since the tighter these constraints are, the closer we are to a general weighting which we originally wanted to avoid by using DEA. Further discretionary power can be added by using DEA as a preselection method after which a ranking is obtained using another criterion, as suggested in Section 5.2.

5 Extensions to the Model

5.1 Including Cost Criteria

In the previous sections, only quality aspects were used to determine the efficiency of an information source. Cost aspects such as Response Time and Price, which may be just as important to users, have been omitted. These two exemplary criteria will now be incorporated into the model.

First, we redefine efficiency according to the ratio form model as the ratio of *weighted sum quality* and *weighted sum cost*. We then extend LP-Quality by the cost criteria and we obtain the *fractional linear program* FLP-QualityCost shown below. We denote by q_{ij} the measured quality scores, as before, and by c_{ij} the measured cost scores.

FLP-QualityCost :

$$\begin{array}{ll} \text{maximize} & \frac{w_1 \cdot q_{1j_0} + w_2 \cdot q_{2j_0} + w_3 \cdot q_{3j_0}}{w_4 \cdot c_{1j_0} + w_5 \cdot c_{2j_0}} \\ \text{subject to} & \frac{w_1 \cdot q_{1j} + w_2 \cdot q_{2j} + w_3 \cdot q_{3j}}{w_4 \cdot c_{1j} + w_5 \cdot c_{2j}} \leq 1 \quad \text{for all sources } S_j \\ & w_1, w_2, w_3, w_4, w_5 \geq \varepsilon > 0 \end{array}$$

Any problem of this form has an infinite number of optimal solutions obtained by scaling the weights in numerator and denominator. We therefore transform FLP-QualityCost into the equivalent linear program LP-QualityCost using a method proposed by Charnes and Cooper in (Charnes and Cooper 1962). The LP-QualityCost maximizes in effect quality at constant cost:

LP-QualityCost :

$$
\begin{array}{lrrl}
\text{maximize} & w_1 \cdot q_{1j_0} + w_2 \cdot q_{2j_0} + w_3 \cdot q_{3j_0} & -w_4 \cdot c_{1j} - w_5 \cdot c_{2j} & \\
\text{subject to} & w_1 \cdot q_{1j} + w_2 \cdot q_{2j} + w_3 \cdot q_{3j} & -w_4 \cdot c_{1j} - w_5 \cdot c_{2j} & \leq 1 \text{ for all } S_j \\
 & & w_4 \cdot c_{1j_0} + w_5 \cdot c_{2j_0} & = 1 \\
 & & w_1, w_2, w_3, w_4, w_5 & \geq \varepsilon > 0
\end{array}
$$

The fractional objective function of FLP-QualityCost is linearized by fixing the denominator to an arbitrary constant and only maximizing the numerator. When maximizing a fraction, not the individual value of the fraction, but the *relative* magnitude is of importance. The other conditions are simply linearized by subtracting instead of dividing by the denominator. Thus we have enhanced the quality model to a quality and cost model.

Example We will continue the previous example by adding exemplary cost information (Response Time and Price) to the quality scores and recomputing efficiency with LP-QualityCost. The example cost scores and the new results for each source are shown in Table 3.

	Response Time	Price	Recomputed Efficiency
S_1	5 sec.	0.50 $	1
S_2	180 sec.	10.00 $	0.947
S_3	10 sec.	0.00 $	1
S_4	3 sec.	1.00 $	1
S_5	10 sec.	0.10 $	0.9849

Table 3: Cost Scores and Recomputed Efficiency

With the new analysis, source S_2 is no longer efficient. Obviously, its weakness lays in its response time and price which cannot be made good by the high quality scores. On the other hand, source S_4 is now efficient because (i) its response time is low and (ii) its formerly dominating competitor S_2 is now inefficient.

5.2 Including Query-dependent Criteria

The DEA method compares and evaluates the general efficiency of information sources. Having done this once, the obtained efficiency scores can be used to rank the sources for queries. However, some sources may be better for certain queries than others. To take this into account, the efficiency scores must be calculated in a query-dependent way.

One way to do this is to add another dimension/criterion to the model, which determines how appropriate a source is in answering the query at hand. We will call this criterion RELEVANCE, as in "How relevant is this source to my query?". Several methods of calculating RELEVANCE have been introduced in literature. The GlOSS model by Gravano et al. (Gravano, Garcia-Molina, and Tomasic 1994) calculates the probability that a source provides at least one document containing all the keywords searched for. The model by Florescu et al. determines the *coverage* of an information source of a given topic (Florescu, Koller, and Levy 1997).

As RELEVANCE seems to be a very important criterion, another method to emphasize it further is to calculate efficiency of all sources as described above and then rank all efficient sources by RELEVANCE in a second phase. Thus, all inefficient sources are no longer considered while the efficient sources are differentiated further, overcoming the problem of little discretionary power discussed earlier.

6 Summary and Outlook

In the scope of this paper we have shown that taking quality measures into account when searching the Internet is of great importance, but also of great difficulty. We have applied the Data Envelopment Analysis method to source selection and have shown that it overcomes these difficulties and provides a "fair" way to compare sources. In particular, we have quantitatively modelled quality and cost criteria and demonstrated how the comparison of sources according to those criteria can be expressed as a number of linear programs. We have further taken query-dependent quality measures into account, for which sources are compared according to DEA in a query basis.

We plan to extend the DEA method to not only select individual sources in a query dependent way, but to also provide support in deciding for the best query execution plan to answer a given query: Whenever responses from more than one source can be merged into one reply, a system must decide which information sources to query in which order, and must decide how to combine the answers. This results in a multidimensional comparison of different combinations of sources. Thus, DEA may provide a helpful solution.

We found DEA to be a promising approach to the problem of making comparisons according to multiple criteria, as is necessary for comparisons on information quality. Further research will deal with open issues such as fine-tuning constraints upon the weights and incorporation of user preferences in the source selection process. Further, we plan to compare DEA to other decision making methods.

References

Ali, A. I. (1993). Streamlined computation for data envelopment analysis. *European Journal of Operational Research 64*, 61–67.

Charnes, A. and W. Cooper (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly 9*, 181–185.

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

Dantzig, G. (1963). *Linear Programming and Extensions.* Princeton, NJ: Princeton University Press.

Dyson, R., E. Thanassoulis, and A. Boussofiane (1990). *Tutorial Papers in Operational Research*, Chapter Data Envelopement Analysis. Operational Research Society.

Florescu, D., D. Koller, and A. Levy (1997). Using probabilistic information in data integration. In *Proc. of the 23rd VLDB Conference*, Athens, Greece.

Gravano, L., C.-C. K. Chang, and H. Garcia-Molina (1997). STARTS: Stanford proposal for internet meta-searching. In *Proc. of the ACM SIGMOD Conference.*

Gravano, L., H. Garcia-Molina, and A. Tomasic (1994). The effectiveness of GlOSS for the text database recovery problem. In *Proc. of the ACM SIGMOD Conference.*

Levy, A. Y., A. Rajaraman, and J. J. Ordille (1996). Querying heterogeneous information sources using source descriptions. In *Proc. of the 22nd VLDB Conference*, Bombay, India.

Liu, L. and C. Pu (1997). A metadata based approach to improving query responsiveness. In *Proc. of the 2nd IEEE Metadata Conference.*

Redman, T. C. (1998). The impact of poor data quality in the typical enterprise. *Communications of the ACM 41(2)*, 79–82.

Tayi, G. K. and D. P. Ballou (1998). Examining data quality. *Communications of the ACM 41(2)*, 54–57.

Wang, R. Y. (1998). A product perspective on Total Data Quality Management. *Communications of the ACM 41(2)*, 58–65.

Wang, R. Y., H. B. Kon, and S. E. Madnick (1993). Data quality requirements analysis and modeling. In *Proceedings of the ICDE*, pp. 670–677.

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.*