ESTABLISHING TRUST NETWORKS BASED ON DATA QUALITY CRITERIA FOR SELECTING DATA SUPPLIERS

Ricardo P. del Castillo, Ismael Caballero, Ignacio García-Rodríguez, Macario Polo, Mario Piattini
Alarcos Research Group, UCLM-Indra Research & Development Institute, University of Castilla-La Mancha
P^o de la Universidad n. 4 13071, Ciudad Real, Spain

Eugenio Verbo
Indra Software Labs, Ronda de Toledo s/n 13003, Ciudad Real, Spain

Keywords: Trust Network, Data Quality, Data Network, Data Provenance.

Abstract: Nowadays, organizations may have Web portals tailoring several websites where a wide variety of information is integrated. These portals are typically composed of a set of Web applications and services that interchange data among them. In this setting, there is no way to find out how the quality of the interchanged data is going to evolve successively. A framework is proposed for establishing trust networks based on the Data Quality (DQ) levels of the interchanged data. We shall consider two kinds of DQ: inherent DQ and pragmatic DQ. Making a decision about the selection of the most suitable data supplier will be based on the estimation of the best expected pragmatic DQ levels. In addition, an example is presented to ilustrate framework operation.

1 INTRODUCTION

Currently, companies usually have several interrelated Web portals. These Web portals integrate different Web applications. Indeed, there may be external links to Websites of other organizations. Used information may not be stored in a centralized manner in order to be shared by all applications, but each application typically manages its own data (Yin et al., 2007). There is a data flow among these Web applications. Each application, site or service in the Web portal (named *node* in this paper) can act as a supplier or consumer of data in any given moment. The set of participating nodes is called *data networks* in (Cai and Shankaranarayanan, 2007). In these networks, a business process in a node may have defined several data source nodes that are not mutually exclusive. Thus, a certain node for a certain business process is entitled to collect data from its supplier nodes. However, the node only collects required data from one of the nodes at any given moment.

A problem of Data Quality (DQ) can appear in the scenario described above: If a node of the network needs to acquire pieces of data from another node, it might not meet the quality of incoming data (Cai and Shankaranarayanan, 2007) and thus, it may use data with inadequate levels of DQ. In other words, a Web application can only understand the quality of incoming data; the so-called *'inherent DQ'*. This DQ is the degree to which data accurately reflects the real-world object that the data represents (English, 1999). In spite of the node knows its *'inherent DQ'*, it does not understand how much quality the incoming data has until it is interchanged and used; this DQ is called *'pragmatic DQ'*. This DQ is the degree of node customer satisfaction derived by the use that it is made of pieces of data (English, 1999). Impossibility to meet the *pragmatic DQ* in this scenario is due to two main reasons. (1) Even in an hypothetical case of a node knowing the inherent DQ of the provided data, the DQ could be different after the acquisition, since *pragmatic DQ* is dependent on the context (Strong et al., 1997). (2) In the case of having different suppliers for the same information need (Wu and Marian, 2007), they are expected to provide data with different expected *pragmatic DQ* levels.

Low levels of DQ affect the overall efficiency of the organization (Caballero et al., 2004). According to (Eppler and Helfert, 2004), the cost of preventing DQ problems is lower than the cost of detecting and repairing them. So in this scenario of Web portals interchanging data, it would be reasonable to prevent DQ problems before they appear. One way to achieve this prevention, or at least minimize its effect, can

P. del Castillo R., Caballero I., García-Rodríguez I., Polo M., Piattini M. and Verbo E. (2009).
ESTABLISHING TRUST NETWORKS BASED ON DATA QUALITY CRITERIA FOR SELECTING DATA SUPPLIERS.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Databases and Information Systems Integration*, pages 37-42
DOI: 10.5220/0001862300370042

consist of selecting the best data supplier for a task.

This paper proposes a framework based on trust networks, which can be used by a node of the network to estimate the expected pragmatic DQ. These Trust Networks allow taking into account the data provenance (Prat and Madnick, 2008), i.e. all processing history of data from its source. The goal is to select, in a heuristic manner, among all available nodes which is the one offering higher DQ levels. In each network, expected pragmatic DQ will be estimated between each pair of nodes creating different supply chains (Nicolaou and McKnight, 2006). Each of these supply chains will provide, in the end, a *DQ pragmatic* value that represents the data provenance of the chain. This will allow choosing the most suitable data supplier. The remainder of this paper is structured as follows: the second section reviews related work. The third section presents the proposed framework and illustrates its usage by means of an example. The final section presents the conclusions and future work.

2 RELATED WORK

Many authors agree that data has quality if it fits the intended use for which it was created (Batini and Scannapieco, 2006; Strong et al., 1997). Inadequate levels of DQ in an organizational Information System will have a negative impact on the business performance (Caballero et al., 2004). Therefore, organizations should take into account DQ issues in order to improve their performance (Al-Hakim, 2007). Due to the existence of data networks (Cai and Shankaranarayanan, 2007), assessing the DQ of each Web node in the data network is not enough (Caro et al., 2008; Eppler et al., 2003). One of the most interesting strategies for tackling the study of DQ for data network context, is to break it down into 'minor qualities' known as DQ dimensions.

According to English (English, 1999), assessment of the inherent DQ, the DQ dimensions belonging to the intrinsic category given by (Strong et al., 1997), (Accuracy, objectivity, believability and reputation), may be used. On the other hand, the pragmatic DQ can be assessed through DQ dimensions of the contextual category (relevancy, added value, timeliness, completeness, amount of data) given by (Strong et al., 1997). For our proposal, we will be interested in measuring not only the inherent DQ of the pieces of data that it are interchanged between each pair of nodes, but we also hope to estimate how usable they will be for an application (Even and Shankaranarayanan, 2007). In order to estimate the *Pragmatic DQ*, the objective is to assist in the selection of the optimal data supplier, using DQ as a discriminator (Al-Hakim, 2007).

Moreover, the research in the DQ field suggests moving the focus from Information Systems to *Information Products* (IP) (Wang et al., 1998). This approach proposes considering pieces of information as products because standard techniques for managing DQ, like *Total Data Quality Management* (TDQM) (Wang, 1998), can be applied. IP-MAP graphical notation has emerged for depicting IPs (Shankaranarayanan et al., 2000). IP-MAP indicates how an IP is created during the manufacturing process. Moreover, an IP-XML file is used for representing IP-MAP meaning through metadata that can be interchanged (Cai and Shankaranarayanan, 2007).

In order to efficiently assess the quality of data, knowledge of where pieces of data have been provided from is necessary. Moreover, in this assessment, it is essential to know the historical transport of pieces of data. According to (Simmhan et al., 2005) data provenance is *"information that helps to determine the derivation history of a data product, starting from its original sources"*. This approach has been used in data sharing and data integration. For instance, provenance information is used to determine data updates, to explain relationships between source and target nodes that interchange data, and so on (Buneman and Tan, 2007).

Finally, the trust networks consist of a set of transitive relations of trust between people, organizations and information systems connected in a intercommunicated environment (Yin et al., 2007). In a specific semantic context, *trust* is transitive and may be derived from the network (Josang et al., 2007). Usefulness of these networks is in the ability to make trust-based decisions: these networks can infer trust in nodes that are not communicated directly (Josang et al., 2007). This is a key advantage of these networks, because an application or service on a Web site can choose the provider with a greater degree of trust. In this selection, the application or site will not be aware of all providers in the supply chain that are behind it (Josang et al., 2007). The Application or site knows only the nodes directly related to it.

3 PROPOSED FRAMEWORK

The selection of a data supplier could be made, taking as a basis, the observation of *inherent DQ* in each node acting as data supplier. However, the framework proposes to estimate the expected *pragmatic DQ* of the pieces of data supplied by each node in the data network (Tinglong and Xiangtong, 2007) as a criterion for selecting the best supplier node. Therefore, finding an approximate value that synthesizes the expected *pragmatic DQ* (English, 1999) along a supply route in the network is proposed.

The structure of the proposed framework is the following: the entire process for creating a trust network will be governed by a 'trust network creation' algorithm which uses three components that are also defined in the framework. (1) *'Matching method'* selects a subset of nodes involved in the *data network* which can be candidates belonging to the *trust network* of a given node. (2) *'Estimation of Expected Pragmatic DQ'* method which is responsible for estimating an approximated value of the expected *pragmatic DQ* along the supply chains in the trust network. (3) *'Function of data supplier selection'* allows selecting the most appropriate data supplier in terms of expected *pragmatic DQ*. The following paragraphs explain the details of each component.

3.1 *Trust Network Creation* Algorithm

To define the scope of a trust network our framework incorporates an algorithm that will define the limits of network on which *pragmatic DQ* is estimated. It starts from the node that requires pieces of data. The algorithm establishes the nodes within the trust network that it attempts to develop. The trust network is going to be built through transitive relations. These relationships are identified by a matching process. Through the algorithm (see Algorithm 1), the network is built starting from the *'node'* which tries to select the best data supplier for an *Information Product* (IP) manufacturing process (Wang, 1998). An XML-Based description of the IP-MAP diagram corresponding to the manufacturing process can be made by IP-XML (Cai and Shankaranarayanan, 2007). The IP-XML file, containing information about the data network, will be one of the arguments of the matching function. Each node will recursively ask its successive suppliers through the matching function *'getDirectSuppliers'*. The algorithm also accepts the argument *'threshold'* as a way to stop recursion (Josang et al., 2007). This limitation tries to minimize derived problems of cycles on the network. The threshold indicates the depth achieved by the algorithm during the node search (Tinglong and Xiangtong, 2007). Once the algorithm arrives at the deepest point of the different supply routes, the estimated values of expected *pragmatic DQ* (estimated trust) go backward within argument *'measures'*. When the algorithm reaches back to the consumer node, the node will be in disposition to select the most suitable data supplier by means of the function *'selectOptimal'*.

Algorithm 1: SelectSuplier.

```
input   :
          node: It is the consumer node where trust network will be built
          ipxml: It represents IP-MAP info associated whith node
          threshold: It is the maximum number of data interexchanges
output  :
          supplierNode: it is the optimal node to provide data to the node
begin
    if threshold = 0 then
        supplierNode ← node.getInherentDQ ()
    end
    else
        measures {} ← ∅
        suppliers {} ← node.getDirectSuppliers (ipxml)
        foreach sup ∈ suppliers do
                measures ← measures ∪ selectSupplier (sup,
                sup.ipxml, threshold-1)
        end
        supplierNode ← selectOptimal
        (measures.getExpectePragmaticDQ ())
        return supplierNode
    end
end
```

3.2 Matching Method

The matching method can determine the transitivity of trust in the network (Josang et al., 2007), i.e. the transitivity of *pragmatic DQ*. This method analyzes the IP-MAP diagram of each node and contrasts each IP-MAP in trying to find an *overlapping point* where offering fits demand (Cai and Shankaranarayanan, 2007). These *overlapping points* are determined through the comparison between *process* blocks in different IP-MAP diagrams. IP-MAP is a graphical notation to represent the elaboration process of Information Products (IP) (Shankaranarayanan et al., 2000; Wang, 1998). IP-MAP includes a set of construct blocks to depict the raw input/output data, processing, data storage, decisions and so on. For each process, the correspondence between the *raw input data* blocks and *raw output data* block in both IP-MAP diagrams is examined. This activity requires a mechanism that indicates the *semantics* of involved process in the data networks. Due to this *semantics*, the matching method will identify the overlapping points. In this paper, we propose to use IP-MAP (Cai and Shankaranarayanan, 2007). However, others mechanisms could be used for this task as *Business Process Modeling Notation* (BPMN) or *activities diagrams*. The algorithm (see Algorithm 1), through the matching method, determines the subset of trust network nodes among all data network nodes. At this moment, the algorithm is at the deepest point of recursion (see Algorithm 1), and has established the entire network of nodes involved in the assessment of trust (*pragmatic DQ*) through the matching method.

3.3 Estimating Expected Pragmatic DQ

At this stage, the framework should estimate the expected *pragmatic DQ* in each set of suppliers.

The *pragmatic DQ* will be spread backward until it reaches the basis node consumer, allowing it to select the best supplier (Eppler et al., 2003). This *pragmatic DQ* has to synthesize, somehow, the value of historic *pragmatic* and *inherent DQ* that there is behind each supplier in its supply chain (Al-Hakim, 2007). These supply chains represent the data provenance of each network node. Therefore, each node on the network has an associated *inherent DQ* value based on the DQ of supplied data for certain processes, and another estimated *pragmatic DQ* value. The i*nherent DQ* value will be measured under the following assumptions. (1) DQ dimensions must be established previously for measuring the *inherent DQ* (Eppler et al., 2003). These DQ dimensions are the same for each set of supplied data, and must be compatible with all network nodes. (2) It will use a synthesizing numerical value of *inherent DQ* for each node in the network. This value represents the degree of trust exhibited in the network (Yin et al., 2007). To obtain this unique value, a process of grouping values of the different dimensions has to be executed. It involves the following actions. (2a) Summarizing and grouping functions like averages, totals, maximums, and so on. (2b) For non-numerical dimensions, a set of linguistic labels and *soft-computing* techniques to obtain a numerical value. (2c) To normalize all DQ dimensions the same scale *'S'* is used which is defined by a minimum and maximum value.

$$scale(S) = S_{max} - S_{min} \tag{1}$$

Each node of the trust network offers data with an expected *pragmatic DQ* level (Q_P). The estimation of this Q_P value is carried out by means of the following heuristics. These are based on other similar studies as (Yin et al., 2007).

Heuristic 1. *Pragmatic DQ of a certain node depends on both Inherent DQ of this node and Pragmatic DQ of all nodes which interchange pieces of data whith the node.*

Heuristic 2. *The weighting of each Pragmatic DQ value, in each node that affect source node, is related to difference between Inherent DQ and Pragmatic DQ for each node.*

Therefore, Q_P value depends on its *inherent DQ* (Q_I) and on estimated *pragmatic DQ* of its set of suppliers. Both terms are given a node-dependent weight α and β (see (5) and (6)). For taking into account the *pragmatic DQ* values of the suppliers, it will make an average on every Q_P belong to set of suppliers ($\{suppliers\}$). The *heuristic 2* is used to obtain W_K: the weight associated with each term k belonging to $\{suppliers\}$ (W_K) will be proportional to how Q_P and Q_I differ in each node.

$$W_K = 1 - \frac{|Q_{P_K} - Q_{I_K}|}{scale(S)} \tag{2}$$

In (3) (using formula (2)), the suppliers' Q_P is summarized. This term is identified as σ_{P_K} which is based on provenance-based believability assessment presented in (Prat and Madnick, 2008):

$$\sigma_{P_K} = \frac{\sum_{k\in\{suppliers\}}(W_K \cdot Q_{P_K})}{|\{suppliers\}|} \tag{3}$$

Taking into account (2), (3) and also the *inherent DQ*, the estimated value of Q_P in the node $k+1$ is as :

$$Q_{P_{K+1}} = \alpha \cdot Q_{I_{K+1}} + \beta \cdot \sigma_{P_K} \tag{4}$$

This formula is a recurrent function which allows to propagating back Q_P values towards initial node. Moreover the framework establishes α and β weights in (5) and (6). For a specific node, if suppliers' Q_P varies greatly, it will give more weight to the Q_I of that node. In addition, there are two exceptional cases: on one hand, if the algorithm is at the network limits, and hence suppliers do not exist, it only considers Q_I, so $\alpha = 1$. And on the other hand, if there is only one supplier, and therefore cannot check the disparity of Q_P, then $\alpha = \frac{1}{2}$ for Q_I and σ_{P_K} have the same weight.

$$M = max(\{Q_{P_n} | n \in \{suppliers\}\})$$

$$m = min(\{Q_{P_n} | n \in \{suppliers\}\})$$

$$\alpha = \begin{cases} 1 & \text{if } |\{suppliers\}| = 0 \\ \frac{1}{2} & \text{if } |\{suppliers\}| = 1 \\ \frac{|M-m|}{scale(S)} & \text{if } |\{suppliers\}| > 1 \end{cases} \tag{5}$$

$$\beta = 1 - \alpha \tag{6}$$

3.4 Function of Data Supplier Selection

At this stage, the proposed algorithm has returned all *pragmatic DQ* values for each origin node's suppliers. At this point, the node will select the most suitable supplier according to the expected *pragmatic DQ* through a selection function (Al-Hakim, 2007; Tinglong and Xiangtong, 2007). The selection function must take into account the acquired knowledge of *data provenance*. This function aims to select the network node which will provide data. The selection function can implement criteria as simple as choosing the greatest Q_P value among all their supply nodes. However, the selection function could be more sophisticated, and consider for example: the Q_P evolution over time, combining several estimated measures, taking into account the *quality/cost* relationship and so on.

4 USING THE FRAMEWORK

In this section, we present an example to illustrate the use of framework. The Figure 1 depicts the data network of an organization. The algorithm creates a trust network for a certain task in a certain node. In our example, the certain task is *'stock updating'* and the certain node is *sales Web application* (see Figure 1). The algorithm uses the IP-MAP diagrams during the process of matching. The sales application node obtains the IP-XML of those nodes with which it is logically interconnected (*production*, *intranet* and *corporative website* (see Figure 1)). The matching method has verified that two of the three, both the *intranet* and *production* nodes, can act as data suppliers for the IP in the consumer node. In this case, the matching method has contrasted that some data destinations in the IP-MAP of these nodes contain data sources in IP-MAP of the *sales Web application* node. The matching method is executed successively until all supply routes are established. The trust network based on DQ will be applied on the recently created network (see Figure 2).

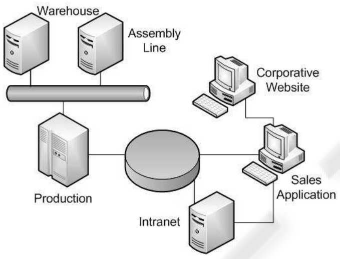

Figure 1: Network of an organization.

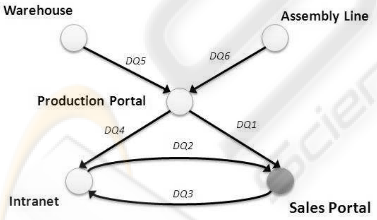

Figure 2: Created Trust Network.

For the sake of estimating the *pragmatic DQ*, each node of the trust network established previously for the case of *stock updating* in *sales Web application* should be borne in mind. In this stage, the algorithm will start estimations of expected *pragmatic DQ* in different network nodes. The network (see Figure 3) details *inherent DQ* values, offered initially by each network node. The scale of DQ values is between 1 and 10. In addition, the Figure 3 illustrates the first Q_P values (*Warehouse* and *Assembly Line* nodes). These are propagated within the network towards the origin node (*sales Web application*). In this case, the absence of suppliers makes $\alpha = 1$ which implies that $Q_P = Q_I$. Then, expected *pragmatic DQ* of the *production* node is calculated based on *Warehouse* and *Assembly Line* nodes (see Figure 4). The weights are $\alpha = 0.1$ and $\beta = 0.9$ because $Q_{P_{assemblyline}} = 5$ and $Q_{P_{warehouse}} = 4$, whose difference is 1. Therefore $Q_{P_{production}} = 0.1 \cdot 6 + 0.9 \cdot \left(\frac{(1-0)\cdot 4}{2} + \frac{(1-0)\cdot 5}{2} \right) = 4.65$. The estimated $Q_{P_{production}}$ value is offered to *intranet* and *sales application* nodes. Nevertheless, *sales Web application* node disposes of this value only, hence $Q_{P_{intranet}}$ must be also estimated (see Figure 4). Finally, expected *pragmatic DQ* of the *intranet* node is estimated (see Figure 5). The weights are $\alpha = 0.5$ and $\beta = 0.5$ because *intranet* node has a single supplier node; hence $Q_{P_{intranet}} = 0.5 \cdot 7 + 0.5 \cdot \left(\frac{(1-0.135)\cdot 4.65}{1} \right) = 5.51$. After all *pragmatic DQ* values have been estimated in the trust network, the optimal supply node can be selected. We must remember that in this case the selection function is as simple as selecting the greatest Q_P value. In the example (see Figure 5), the *sales Web application* will take data for updating the stock from the *intranet*, because the trust (Q_P) of this node with 5.51 is greater than the one of the *production* node whose value is 4.65

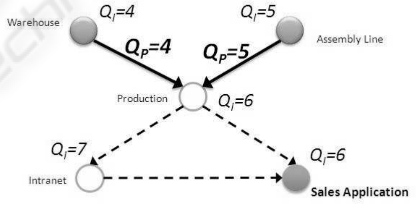

Figure 3: Trust calculations in the network (Step I).

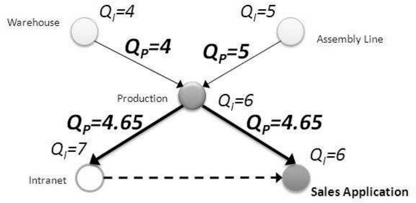

Figure 4: Trust calculations in the network (Step II).

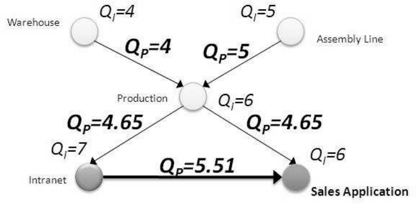

Figure 5: Trust calculations in the network (Step III).

5 CONCLUSIONS

This paper has proposed a framework based on trust networks applied to data networks. The framework estimates an expected value at each node in the supply chain, taking into account the remaining nodes that supply data to it. The presented framework is able to determine which data supplier offers the most suitable expected *pragmatic DQ* in each provenance scenario. The proposed framework uses, undoubtedly, an approximated measurement, therefore there is no guarantee of finding the optimal provider in all situations. In the future, we will work on two key aspects. (1) It will be validate in empirical manner as well as by means of simulation or analytical evaluation. (2) We will provide several selection functions which take into account other factors as quality/cost relationship or historical data in order to increase support to decision-making in these networks.

ACKNOWLEDGEMENTS

This research is part of the projects ESFINGE (TIN2006-15175-C05-05/), DQNet (TIN2008-04951-E) and HERMES (TSI-020100-2008-155) supported by the *Spanish Ministerio of Educación y Ciencia*; and project IVISCUS (PAC08-0024-5991) supported by the *Consejería de Educación y Ciencia of Junta de Comunidades de Castilla - La Mancha*.

REFERENCES

Al-Hakim (2007). The effects of information quality on supply chain performance: New evidence from malaysia. In *Information Quality Management: Theory and Applications*. Igi Global, 1 edition edition.

Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications*. Springer-Verlag Berlin Heidelberg, Berlin.

Buneman, P. and Tan, W.-C. (2007). Provenance in databases. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173, New York, NY, USA. ACM.

Caballero, I., Gomez, O., and Piattini, M. (2004). Getting better information quality by assessing and improving information quality management. In *ICIQ 2004: 9th International Conference on Information Quality*, Cambridge, Boston , USA.

Cai, Y. and Shankaranarayanan, G. (2007). Managing data quality in inter-organisational data networks. *International Journal of Information Quality*, 1(3):254 – 271.

Caro, A., Calero, C., Caballero, I., and Piattini, M. (2008). A proposal for a set of attributes relevant for web portal data quality. *Software Quality Journal. Springer Science*.

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc.

Eppler, M., Algesheimer, R., and Dimpfel, M. (2003). Quality criteria of content-driven websites and their influence on customer satisfaction and loyalty: An empirical test of an information quality framework.

Eppler, M. J. and Helfert, M. (2004). A framework for the classification of data quality costs and an analysis of their progression. In *MIT Conference on Information Quality*.

Even, A. and Shankaranarayanan, G. (2007). Utility-driven assessment of data quality http://doi.acm.org/10.1145/1240616.1240623. *SIGMIS Database*, 38(2):75–93.

Josang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2):618–644.

Nicolaou, A. I. and McKnight, D. H. (2006). Perceived information quality in data exchanges: Effects on risk, trust, and intention to use. *Info. Sys. Research*, 17(4):332–351.

Prat, N. and Madnick, S. (2008). Measuring data believability: A provenance approach. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 393, Washington, DC, USA. IEEE Computer Society.

Shankaranarayanan, G., Wang, R. Y., and Ziad, M. (2000). Ip-map: Representing the manufacture of an information product. In *Proceedings of the 2000 Conference on Information Quality*.

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36.

Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997). 10 potholes in the road to information quality. *Computer*, 30(8):38–46.

Tinglong, D. and Xiangtong, Q. (2007). An acquisition policy for a multi-supplier system with a finite-time horizon. *Comput. Oper. Res.*, 34(9):2758–2773.

Wang, R., Lee, Y., Pipino, L., and Strong, D. (1998). Manage your information as a product. *Sloan Management Review*, 39(4):95–105.

Wang, R. Y. (1998). A product perspective on total data quality management. *Commun. ACM*, 41(2):58–65.

Wu, M. and Marian, A. (2007). Corroborating answers from multiple web sources. In *WebDB 2007: Proceedings of the 10th International Workshop on Web and Databases*, Beijing, China.

Yin, X., Han, J., and Yu, P. S. (2007). Truth discovery with multiple conflicting information providers on the web. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1048–1052, New York, NY, USA. ACM.