# REGULATORS OF TIMELINESS DATA QUALITY DIMENSION FOR CHANGING DATA QUALITY IN INFORMATION MANUFACTURING SYSTEM (IMS)

Mohammad Shamsul Islam
Faculty of Engineering & Computing
Dublin City University
Dublin, Ireland
Email: mohammad.islam6@dcu.ie

## ABSTRACT

*The information manufacturing system (IMS) is the information system that manufactures information from raw data. This system works in both real time and non-real time environment in the real world organizations for providing information or decision support. Delivered information from the information manufacturing system could be poor quality for the timeliness and other objective data quality dimensions such as completeness and accuracy. Further, completeness and accuracy can be changed for the timeliness data quality dimension. This data quality dimension is regulated by some of the factors in general. Each of the general regulating factors of timeliness may not play the regulating role for changing data quality in IMS. Therefore, the purpose of this paper is to sort out those regulators of timeliness data quality dimension.*

## KEYWORDS

Age, Input Time, Delivery Time, Volatility, Timeliness

## 1. INTRODUCTION

The information manufacturing system (IMS) is the information system that manufactures information from raw data [1]. The most important component of IMS is the data storage system (DSS). It is integrated with the multiple sources of the system. Therefore, it contains inbound raw data come from multiple sources. Data comes from multiple sources are to be processed for the availability of data in the DSS of IMS. Available data in the DSS are then delivered for the information support. According to [2], [3], [4], refreshment and query function execute in the data storage system of the IMS to make data available and for information support respectively.

**Refreshment Function:** It is a complex process comprising the tasks, such as data loading, indexing and propagation of data for synchronizing the data in the information manufacturing system (IMS) [2], [3], [4].

.

**Query Function:** This function of data storage system in IMS is done by the query processing task. The requested query of the user is processed in the data storage system for delivering the information to the user.

Information manufacturing system works in both real time and non-real time environment in the real world organizations. Air traffic, road traffic, stock brokering, online telecommunication etc. work for delivering data in real time environment. Change frequency of data in this environment could be the fraction of seconds. Therefore, the data are to be updated as fast as possible for maintaining the quality of data. Hence, refreshment function is to work continuously for updating data in the DSS of IMS. As a result, there could have a trade-off between the availability and timeliness of data. Lack of available data can make the problem for the completeness of data. This completeness of data in IMS is the ratio between the number of data stored in DSS and the number of data that should be stored in DSS [5]. Furthermore, if data is not available at the right time in IMS for the processing period, it cannot produce complete information [6]. [7] define the accuracy as the extent to which data are correct. Data stored in the DSS of IMS could be incorrect for the obsoleteness. Therefore, current available data can make an impact on the accuracy data quality dimension. As a result, there is a role of timeliness for changing completeness and accuracy in IMS.

Sometimes, trade-offs need between timeliness and the other objective data quality dimension (accuracy, completeness) [8]. Basically, trade-offs between timeliness and other objective data quality dimension (accuracy, completeness) are to consider for time related data or real time data. In this case, if a query request is sent to the data storage system of IMS and the manipulated data in the data storage system are in refreshment state and there is an obligation to respond the query request within a certain time, the request will response an incomplete and inaccurate data set for the obligation of timeliness. Indeed, having timely data may cause lower accuracy (or completeness). Conversely, having accurate (or complete) data, timeliness is negatively affected [8]. Therefore, data quality is regulated by the timeliness. Hence, the purpose of this paper is to show the regulators of timeliness data quality dimension for which data quality could be changed (completeness, accuracy) in IMS.

## 2. RELATED RESEARCH

Data quality dimensions are correlated with each other. If one dimension is considered more important than the others for a specific application, then the choice of favouring it may imply negative consequences for the other ones. [8] describes the environment for the trade-off between the dimensions and the occurrence of trade-off between data quality dimensions. Basically, trade-off is done between the time- related dimension (timeliness) and the non- time related dimensions like completeness, accuracy and consistency.

[9] integrate multi-channel information system for storing data in single DSS to resist the redundant data. Multi channel integration could be effective for the quality of stable or long-term changing data. Conversely, this integration may not suitable for frequently changing or time related data. In the IMS, it is difficult to maintain the quality of data for both the timeliness dimension and other objective data quality dimension. [10] works on the time related data integration for data warehouses. Continuous data integration is one of the important requirements for time related data storage system (DSS) are discussed in this paper. According to [11], analyze the trade-off between the availability and timeliness data quality dimension for synchronizing or refreshment frequency of DSS in IMS. [12] works on the currency quality factors for data warehouse DSS. Data currency quality goal is expressed by currency constraint associated with every source relation in the definition of every input query. The upper bound in a currency constraint is set by the knowledge workers according to their needs. [4] showed in a survey that some organizations refresh data continuously, some organizations refresh data in every 5 minutes for updating data. Most organizations indicated that more frequent refresh during business hours would negatively impact on the system availability and the timeliness in IMS. [13] model a faster data warehouse to make available the current data as soon as possible.

## 3. TIMELINESS DATA QUALITY DIMENSION IN IMS

Timeliness is defined as the extent to which data are timely for use [7]. [15] define it as the property of information to arrive early or at the right time. Therefore, timeliness of data in IMS depends on whether data are available in time or not. According to [Batini et al, 2006] and [16], Timeliness can be defined as currency and volatility dimensions. More specifically it can be written,

$$\text{Max}\ (0, 1 - \text{currency}/\text{volatility})\ldots\ldots(1)$$

**Volatility of Data in IMS:** As the definition of volatility, we know that the length of time data remains valid is volatility [8]. Therefore, volatility of data depends on the expiry time of each individual data of DSS. Expiry time of data depends on the change frequency of data (i.e. frequently changing and long-term changing).

Expiry time of long-term changing data will be long. On the other hand, expiry time of frequently changing data is very short. For the shortness of expiry time, data could be obsolete if processing time or age of the data is longer than the expiry time of data. Therefore, the formula for the volatility of data in the single DSS in IMS is,

Volatility V (t) = Expiry Time – Start of Data Insertion Time……(2)

Start of data insertion time means starting of insertion time of data from sources to the DSS of the information manufacturing system. Start of data insertion from the sources can be represented as SIT. On the other side, Expiry time can be represented as $E_T$. Expiry time indicates the limit of the validity of data.

**Currency of Data in IMS:** According to [8], currency is defined as,

Currency = Age + (Delivery Time – Input Time)……(3)

Where Age measures how old the data unit is when received, Delivery Time is the time information product is delivered to the user and Input Time is the time data unit is obtained. Therefore, the currency dimension of data in the data storage system (DSS) depends on the age, delivery time and input time. In the database data storage system (DSS), these parameters can be recognized as below,

**Table 1.** Currency Parameter of IMS

| General Currency Parameter | DSS Currency Parameter | Notation |
|---|---|---|
| Age | Waiting Time + Refreshment Processing Time | W (t) + Rpro (t) |
| Delivery Time | Query Response Time | Q (t) |
| Input Time | Insertion Time of Data in DSS | I (t) |

**Age A (t):** It can be calculated in DSS by adding waiting time of data with the refreshment processing time of data. Waiting time means how long data is waiting in the source before the refreshment processing of data in IMS for the insertion of data in the DSS. Refreshment processing time is calculated by adding the following parameters.

**Table 2.** Refreshment Processing Period Parameters

| Refreshment Processing Time Parameters | Description | Notation |
|---|---|---|
| Loading Period | Time needs for loading data in the DSS | L (t) |
| Indexing Period | Time needs for indexing data in the DSS | Ix (t) |
| Propagation Delay | Time needs for propagating data from one DSS to another DSS. | P (t) |

Therefore, we can calculate the refreshment processing time of the data in the DSS in the following way,

Rpro (t) = L (t) + Ix (t) + P (t)……(4)

*Now, we can write the Age as,*

A (t) = W (t) + (L (t) + Ix (t) + P (t))……(5)

**Input Time:** To be available of data in the DSS, data have to be inserted in DSS of IMS. Data insertion will be completed if refreshment processing of data in the DSS is done. Therefore, the end of the refreshment processing time for each individual data will be the input time of individual data.

**Delivery Time:** It is defined in DSS by query response time. This query response time of DSS means, what time query request of a user query is responded in DSS.

Now, general currency formula from the IMS can be written as,

Currency C (t) = A (t) + (Q (t) – I (t))

= (W (t) + (L (t) + Ix (t) + P (t))) + (Q (t) – I(t))……(6)

## 4. TIMELINESS REGULATORS FOR CHANGING DATA QUALITY IN IMS

Four factors are included with timeliness data quality dimension. These are age, input time, volatility and delivery time. Except input time, others three factors are independent factors. It means that input time is dependent on any of the three independent factors. Therefore, age, delivery time and volatility play a regulating role in the timeliness of IMS.

It is already seen that age is calculated by the refreshment processing period and the waiting period of data in the IMS. Refreshment period is calculated by subtracting end of refreshment processing time from the start of refreshment processing time. Similarly, waiting period is calculated by subtracting the start of refreshment processing time from the start of data involvement time with the system. Therefore, start of the refreshment processing time, end of the refreshment processing time and the start of data involvement time with the system are the properties of age. The variation of these properties affects on the age of data. Therefore, age of data ultimately regulates the timeliness of data.

Input time is the insertion time of data in DSS of IMS. Data is inserted in DSS with a refreshment process. Therefore, end of refreshment processing time is the input time of data. Each single unit of data is to refresh in IMS for making data available. Therefore, input time of each data unit will be the end of refreshment processing time of each data unit. This refreshment processing time is the property of age. As a result, input time does not have any regulating role in IMS.

There could have different changed frequency of data (frequently changing data and long-term changing data) in IMS. The expiry time of these changed frequencies of data is varied. Therefore, volatility of these multiple changed frequency data in IMS will not be the same. Expiry time of the frequently changing data is short, so, volatility period of frequently changing data will be short. On the other hand, volatility period of long term changing the data will be long for the long expiry time. As a result, timeliness will be varied for the variation of the volatility of the data.

The user can send a query request in IMS in any moment of time. Therefore, responding time will be different for the variation of the sending time of query request. Further, late query response could be done for the refreshment processing of inbound data to be available in the DSS of IMS or problem in the outbound data path or any problem in the application of presentation block. Therefore, timeliness is regulated by the query responding time.

### 4.1 Experimental Evaluation of Simulated IMS for Identifying Regulators of Timeliness

In the real world organization, data come from or collected from multiple sources are stored in the DSS of IMS. Expiry time of stored data could be same or different. Refreshment process makes these source data available in the IMS. The refreshment period of the refreshment process varies in the real world organization for the machine capacity, volume of data etc. Further, frequency of refreshment in the real world organization is continuous or periodic. Query request can come from the user in the IMS of real world organization in any moment of time.

Considering the scenario of the real world organization, a simulated information manufacturing system is implemented. Fixed volume of data is collected from the sources with the execution of refreshment function in this IMS for each experiment. This refreshment function executes continuously. For the variation of refreshment period, simulated IMS is implemented in the different capacity machine. The expiry time is included with the source data for measuring the volatility of the data. Further, query request was sent in simulated IMS in the different moment of time when the refreshment processing period was in progress. This query results are then assessed for measuring the data quality. Completeness and accuracy is assessed by the assessment function of [5]. Timeliness is measured by the function given in section 3.1. Three experiments were done. Results of experiment 1, experiment 2 and experiment 3 are shown in table 3, table 4 and table 5 respectively.

**Table 3.** Data Quality (DQ) Assessment in IMS

| User | Age | Input Time | Delivery Time | Currency | Volatility | Timeliness | Completeness | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | (%) | (%) |
| 1 | 0-249936<br>0 | 12:59:17 | 13:03:37 | 260530 | 836063 | 0.69 | 39.86 | 91.28 |
| 2 | 0-299500<br>0 | 12:59:17 | 13:04:21 | 304730 | 836063 | 0.64 | 59.80 | 74.00 |
| 3 | 0-346826<br>0 | 12:59:17 | 13:05:03 | 346966 | 836063 | 0.59 | 73.46 | 50.08 |
| 4 | 0-351796<br>0 | 12:59:17 | 13:05:20 | 363136 | 836063 | 0.57 | 79.73 | 35.93 |
| 5 | 0-402263<br>0 | 12:59:17 | 13:05:59 | 402140 | 836063 | 0.52 | 93.46 | 22.19 |

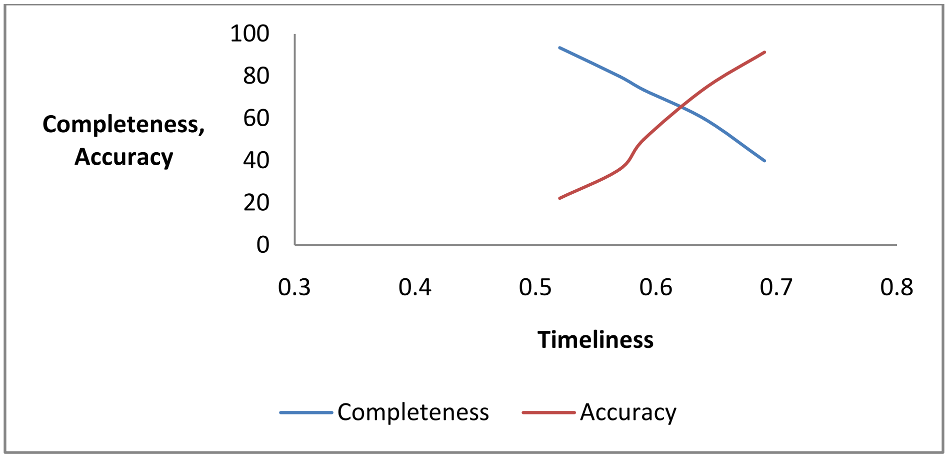

**Figure 1.** Data Quality (DQ) Assessment Graph of IMS

**Table 4.** Data Quality (DQ) Assessment in IMS

| User | Age | Input Time | Delivery Time | Currency | Volatility | Timeliness | Completeness | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | (%) | (%) |
| 1 | 0-13030<br>0 | 23:24:02 | 23:27:16 | 194535 | 633983 | 0.69 | 20.00 | 95.58 |
| 2 | 0-226340<br>0 | 23:24:02 | 23:28:04 | 240914 | 633983 | 0.62 | 39.86 | 75.65 |
| 3 | 0-290153<br>0 | 23:24:02 | 23:28:56 | 294972 | 633983 | 0.53 | 59.80 | 55.72 |
| 4 | 0-331200<br>0 | 23:24:02 | 23:29:28 | 329671 | 633983 | 0.48 | 70.00 | 45.52 |
| 5 | 0-386263<br>0 | 23:24:02 | 23:30:30 | 393069 | 633983 | 0.38 | 93.40 | 22.12 |

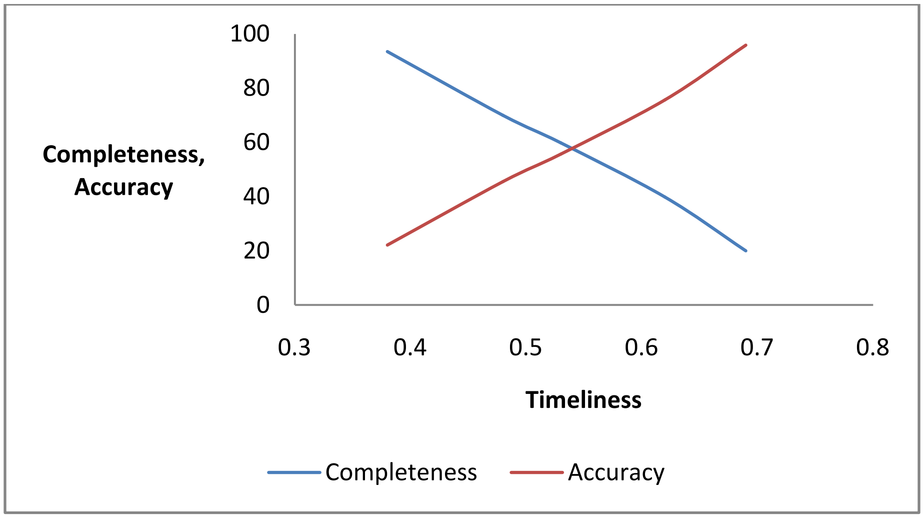

**Figure 2.** Data Quality (DQ) Assessment Graph of IMS

**Table 5.** Data Quality (DQ) Assessment in IMS

| User | Age | Input Time | Delivery Time | Currency | Volatility | Timeliness | Completeness (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0-11042<br>0 | 1:27:01 | 1:30:18 | 197042 | 633983 | 0.69 | 27.00 | 97.69 |
| 2 | 0-196265<br>0 | 1:27:01 | 1:30:53 | 232265 | 633983 | 0.63 | 48.65 | 83.45 |
| 3 | 0-260930<br>0 | 1:27:01 | 1:31:58 | 297972 | 633983 | 0.53 | 66.86 | 68.74 |
| 4 | 0-302022<br>0 | 1:27:01 | 1:32:37 | 336011 | 633983 | 0.47 | 79.20 | 53.38 |
| 5 | 0-360000<br>0 | 1:27:01 | 1:33:21 | 380390 | 633983 | 0.40 | 96.77 | 32.65 |

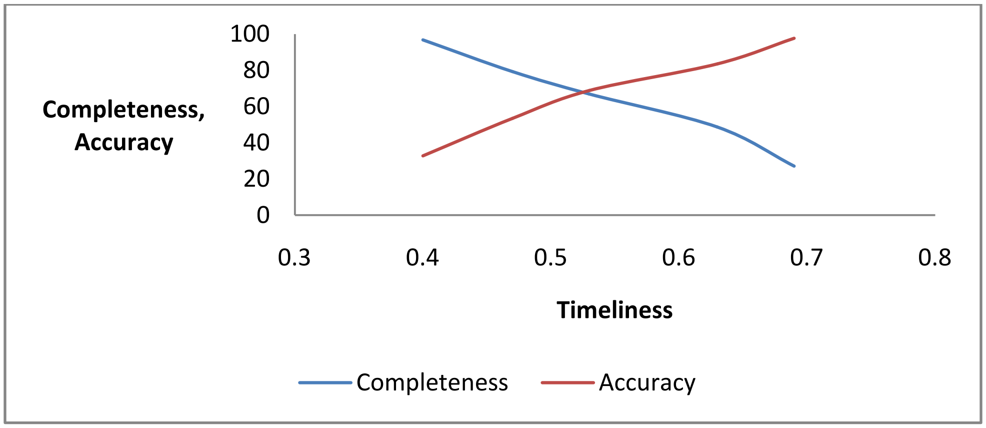

**Figure 3.** Data Quality (DQ) Assessment Graph of IMS

Age, delivery time and volatility played the regulating role for changing the data quality (completeness, accuracy) with timeliness. In the above experiments, it is seen that whenever the distance between delivery and input time increase, data quality varies with timeliness. Input time is fixed for being considered the first data insertion time as input time. Therefore, the distance between delivery time and input time is controlled by the delivery time. In experiment 2, it is seen that when delivery time is 23:27:16 timeliness is high (.69), but, completeness is low (20%) and accuracy is high (95.18%). Accuracy is high because it is the 95.18% of the complete outbound query result of 20%. On the other hand, if we considered the last delivery time (23:30:30) of this experiment, it is seen that timeliness of this experiment is low for this delivery time. Therefore, for this low timeliness (.38), we have seen that completeness is 93.40% but accuracy is 22.12%. This 22.12% accurate and 93.40% complete data means that in our outbound query result, we got 93.40% complete data but 22.12% of that 93.40% complete data are accurate. Now, it can be said that delivery time increase, data quality varies. This scenario is shown in all other experiments.

Now, if we look at the experiment 1 and experiment 2 of IMS, it is seen that volatility of data of experiment 1 is higher than the volatility of data of experiment 2. Therefore, the data quality scenario of experiment 1 is better than the experiment 2.

Age is also regulating the change of data quality with timeliness in IMS. If we analyze the each experiment 2 and experiment 3 of the simulated IMS, we will see that age is regulating data quality with timeliness. In these two experiments, volatility of data is indifferent. Therefore, in the experiment 2, the delivery time of the first query request is 23:27:16. On the other hand, the delivery time of the first query request is 01:30:18 in experiment 3. Now, if we observe delivery time for the query request of user 3 for both experiments, we get that query request was executed after 1 minute 40 second from the query request execution time of the first user. Age of the data of the query result of user 3 for experiment 2 and experiment 3 is 0-290153 and 0-260930 respectively. 0 is the lowest age of data and 290153, 260930 is the highest age of data of the query result. Now, if we look at the timeliness, completeness and accuracy of user 3 of experiment 2 and experiment 3, completeness and accuracy is 59.80, 59.72 and 66.86, 66.84 for timeliness .53 respectively. Therefore, it can be said that age is regulating the data quality in IMS.

## 5. CONCLUSION

Timeliness of data is measured by four parameters. These are age, delivery time, input time and volatility. Among them, age, delivery time and volatility were identified as independent parameters and input time as dependent parameter in the IMS. Therefore, age, delivery time and volatility played the regulating role for changing data quality in IMS. This paper will help us to work on the improvement of data quality in the IMS. Hence, improvement of data quality by considering the regulating factors in the IMS will be our future work.

## 6. REFERENCES

1. Wang, R.Y., Ziad, M. and Lee Y.W.: *Data Quality*. Publisher: Kluwer Academic (2001).

2. Santos, R.J. and Bernardino, J.: *Real-Time Data Warehouse Loading Methodology*. In: Proceedings of the International Database Engineering & Applications Symposium, pp. 49-58 (2008).

3. Bouzeghoub, M., Fabret, F. and Matulovic-Broqué, M.: *Modeling Data Warehouse Refreshment Process as a Workflow Application. In:* Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99).pp. 6.1-6.12 (1999).

4. Mannino, M.V. and Walter Z.: *A framework for data warehouse refresh policies.* Decision Support Systems, vol. 42, pp. 121-143 (2006).

5. Cappiello, C., Francalanci, C. and Pernici B.: *A Self-monitoring System to Satisfy Data Quality Requirements*. Vol. 3761, pp. 1535-1552. Springer, Verlag (2005).

6. Vrbsky, S.V.: A data model for approximate query processing of real-time databases, ACM Data & Knowledge Engineering, vol. 21, pp.79-102 (1996).

7. Wang, R.Y., and Strong, D.M.: Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, vol. 12, pp. 5-33 (1996).

8. Batini, C. and Scannapieco, M.: *Data Quality: Concepts, Methodologies and Techniques.* Publisher: Springer, Berlin, Germany (2006).

9. Cappiello, C., Francalanci, C. and Pernici, B.: *Time-Related Factors of Data Quality in Multichannel Information Systems*. Journal of Management Information Systems, vol. 20, pp. 71-92 (2003).

10. Bruckner, R.M., List, B., Schiefer, J. and Tjoa, A.M.: *Modeling Temporal Consistency in Data Warehouses***.** In: Proceedings of the 12th International Workshop on Database and Expert Systems Applications.pp. 901-905 (2001).

11. Cappiello, C. and Helfert M.: *Analyzing Data Quality Trade-Offs in Data-Redundant Systems.* Interdisciplinary Aspects of Information Systems Studies, pp. 199-205 Physica-Verlag HD (2008).

12. Theodoratos, D. and Bouzeghoub, M.: *Data Currency Quality Factors in Data Warehouse Design.* In: Proceedings of the International Workshop on Design and Management of Data Warehouses. pp. 15.1-15.16 (1999).

13. Hanson, J. H., Willshire M.J.**:** Modeling a Faster Data Warehouse, IEEE, pp. 260-265 (1997).

14. Wang, R. Y., Kon, H. B. and Madnick, S. E.: Data quality requirements analysis and modeling. In: Ninth International Conference on Data Engineering, pp. 670–677 (1993).

15. Ballou, D.P., Wang, R.Y., Pazer, H.L. and Tayi, G.K.: *Modeling Information Manufacturing System to Determine Information Product Quality*. Management Science. vol. 44, pp. 462-484 (1998).

16. Wang R. Y., Reddy M. and Gupta A.: *An object-oriented implementation of quality data products.* In: Proceedings of Third Workshop on Information Technology and Systems, pp. 670–677 (1993).

17. Chaudhuri, S., Dayal, U.: *An Overview of Data Warehousing and OLAP Technology*. Vol. 26, pp. 65-74, ACM SIGMOD Record, (1997).

18. Dong, C., Sampaio, M., and Sampaio, F.: *Expressing and Processing Timeliness Quality Aware Queries: The* $DQ^2L$ *Approach*. Vol. 4231, pp. 382-391 Springer Verlag, (2006).

19. Hu, Y., Sundara, S. and Srinivasan, J.: *Supporting Time-Constrained SQL Queries in Oracle.* In: Proceedings of the 33rd international conference on Very large data bases, pp. 1207-1218 (2007).

20. Sampaio, M., Dong, C. and Sampaio F.: *Incorporating the Timeliness Quality Dimension in Internet Query Systems.* LNCS vol. 3807, pp. 53-62 Springer-Verlag, (2005).

21. Vassiliadis, P., Bouzeghoub, M., Quix, C.: *Towards Quality-Oriented Data Warehouse Usage and Evolution*, Journal of Information System, vol. 25, pp. 89-115 (2000).