Out of Overinformation by Information Filtering and Information Weighting

Laure Berti-Equille

Out of Overinformation by Information Filtering and Information Quality Weighting

Laure Berti
GECT, Equipe Système d'Information Multi-Média
Université de Toulon et du Var
B.P. 132, F-83957 La Garde cedex, FRANCE
berti@univ-tln.fr

Abstract

This paper presents our approach concerning the management of contradictory information in multi-source information fusion process. We break down the process of data capture which alterates information quality and, as compensation, we propose a value-adding process for information. More semantics of information is caught via quality meta-data. Weighting the impact of data quality criteria on user's satisfaction allows an ordering of contradictory data according to their relative quality. The ultimate objective is to accomplish a personal, adaptative information recommendation process with reduction of information loss for Technological Watch applications.

1. Introduction

Quality certification has been receiving increased attention in the last couple of years as a result of international and intra-organizational willingness for normalization [ISO]. Besides system quality engineering, software systems quality metrics and torture-tests, a more system-introspective trend of quality focuses on data as an information product. Data Quality research is clearly emerging with an arsenal of methodologies, frameworks for conceptual specifications, techniques and tools to fight against the omnipresent and costly data non-quality problem in information systems [WaStFi] [BaTa] [Firth]. In parallel, the significant increase in independant distributed database servers directly providing on-line information retrieval services to end-users has led to greatly facilitate information manipulation. Flexible access to vast quantities of information swamps business executives with an impressive amount of multi-source data.

In a more and more competitive business environment, powerful tools are needed for exploiting the information sources efficiently to watch technological developments closely and it's a question of technological survival to find, extract and store relevant multi-source information. In a such information warefare and business intelligence context, an Information Service Provider, such as a Technological Watch Group (TWG), is basically in charge of :

- selecting Information Sources (IS) - technological patents, technical reports, proceedings, brochures... - diffused by competent Primary Producers. The selection of Primary Producers is based on non-exhaustive high-quality criteria (ISQ) : credibility, good reputation to keep, interest to disclose information, domain competence, substancial financial supports, dependancies between Information Sources [ClWi]...
- collecting dynamically presumed relevant information ; the selection of presumed relevant information is based on four aim Information Quality dimensions (IQ) : accuracy, reliability, timeliness and completeness,
- information cross-checking and IQ and ISQ estimating and stamping,
- providing certified relevant information in adequation with the Decision Maker's requirements.

In the final stage of information pre-processing, the Technological Watch Group (TWG) stamps the data with certain degrees of confidence about the quality criteria. But doing so, the TWG becomes a Secondary Producer of information and provides value-added information to the consumers.

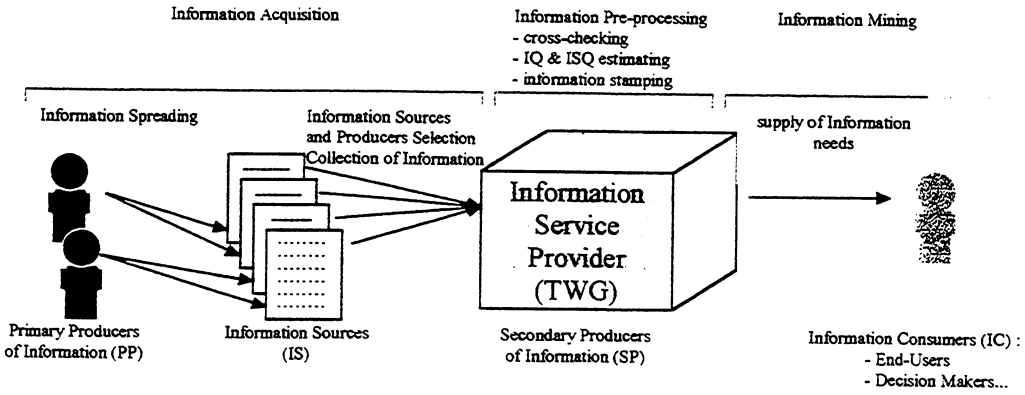


fig. 1. Information Processing

Our practical motivation in tagging Information Quality on data came first from the system, called VIGIWARE system we develop, which is intended to provide personal, adaptative information analysis and recommendation through dissemination of real-time news and Information Sources available electronically. An Information Consumer is generally not The End-User. Various intermediate information users, who implicitly become, in spite of themselves, Secondary Producers, alterate original information and consequently, information quality by replicating information pre-processings with different interpretation models and amplify information distortion.

The objective of this paper is to stress the multi-source aspect of information which can affect its Quality, in particular, if no contextual meta-information is captured and stored in the database. Section 2 describes the process of data capture and storage in terms of information loss and performance reduction. It proposes a compensatory process by creating value-added data. Section 3 presents an example in the context of Technological Watch applications. Conclusions and future work are given in Section 4.

2. Creation of Value-Added Information

The VIGIWARE system is a DBMS specifically meant for Technological Watch applications. The semantics of information (particularly meta-information about the contexts of information extraction and consumption) is crucial and should be captured with data. But, most of the time, this contextual meta-information is lost during the data capture and storage.

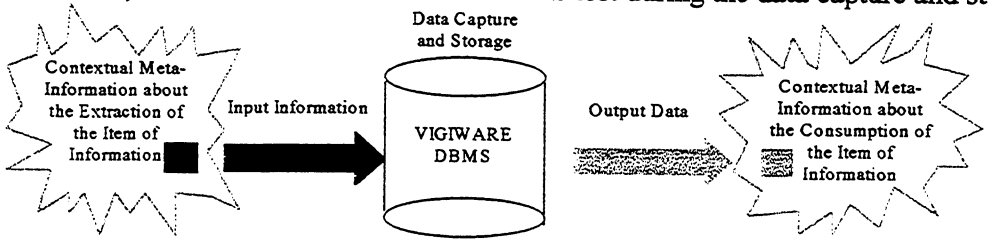


fig. 2. Loss of Contextual Meta-Information in Data Capture and Storage

We consider that an item of information is composed by an informational content, a form and an interpretation model. As we are particularly interested in subjective quality indicators

(as IQ and ISQ perceptions), the major problem is to explicitly represent the informal aspect of information (which orientates the interpretation model) and, more precisely, the notion of « information performance » ; that is, its capability to cause effects and reactions on the information receiver (e.g. he/she is septic or convinced)*. In proportion to the number of information pre-processings (which increase the loss of contextual meta-information), the information performance can be drasticly reduced.

2.1. Performance Reduction in Data Capture and Storage

The information performance reduction is engendered as soon as the item of information is captured and stored as one-source data in the VIGIWARE system. It loses a part of its semantical relief.

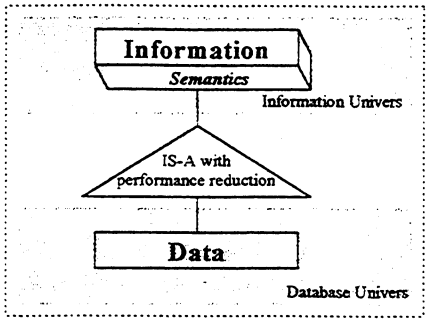


fig. 3. Information Performance Reduction in Data Capture and Storage

Semantics of information are atoms of meaning for the interpretation and the intended use. Actually, the data catches a part of the informational content from the original item of information and usually takes a different form. The information interpretation model is weither implicitly caught inside the content (or the form) or ignored to fit to the « data mould ». This primary process of information performance needs compensation effects ; so that, data become, at least, value-added data. So, we first develop, then implement a formal model using the step-by-step methodology proposed in [WaKoMa] for defining and documenting data quality criteria to be tagged to data items. But, it failed to address some issues that are particularly important to Technological Watch applications : the subjective and informal aspects of IQ and ISQ perceptions for the justified recommendation of multi-source contradictory information.

2.2. From Data To Value-Added Data

Basically, each Secondary Producer (each expert in the TWG), checks weither the item of information is valid or not and estimates IQ and ISQ. He defines his personal views on the quality of each item of information he collects. In the database, the representation of IQ, ISQ and contextual meta-information about extraction (Information Source ID, Secondary Producer ID, capture date,...) is actually made by meta-data which are tagged to the data. The Interpretation Models of each Secondary Producer are explicitly captured as justificative note pads explaining the relevant steps of his reasoning for the IQ and ISQ assessment.

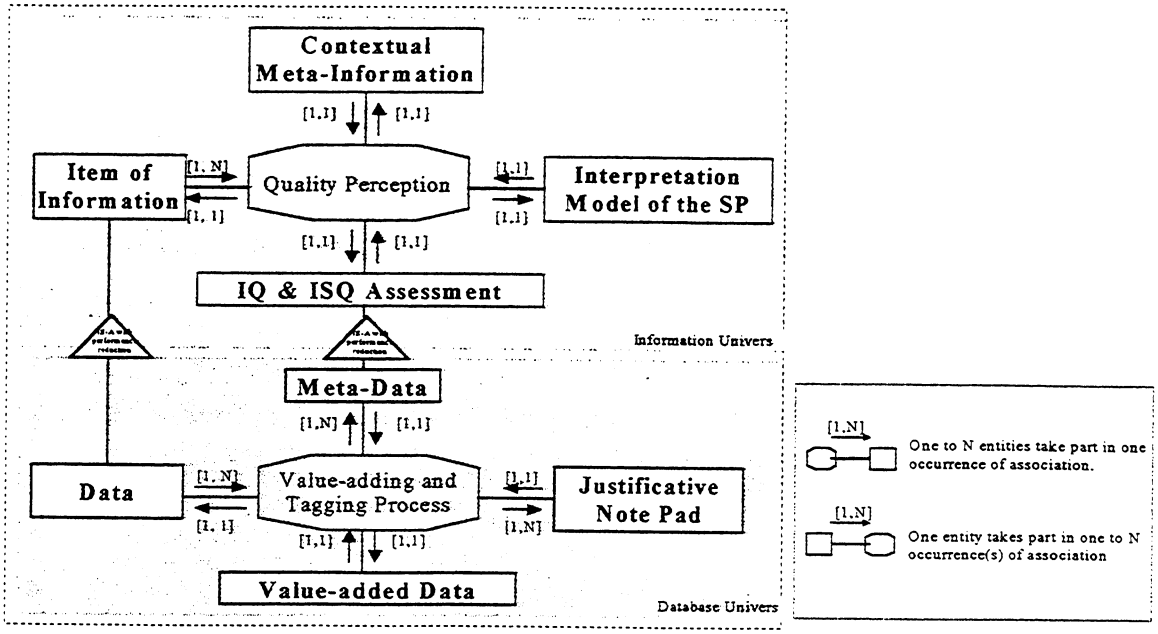


fig. 4. From Information and Meta-Information to Data and Meta-Data

2.3. Multi-Source Information Pre-Processing

The notion that more information leads to better inferences and decisions motivates the condideration of multiple Information Sources. So, in a multi-source context, information is considered as an « uncertainty reducer ». From this point of view, we develop an information fusion process. This approach is not to integrate or derive multi-source data but, to store a maximal number of data and build *multi-source-valueted attributes* and *maximal objects* which are *information agglomerates* at both syntactical and semantical levels. This step leads to overinformation. Overinformation is then considered as a distortion amplifier because multi-soure data are usually contradictory, overlapping, incomplete, or do not meet users' needs. The multi-source data fusion process has been formalized and implemented. It extends the above concepts and works out semantics of information according to users' points-of-view (the context of information extraction for the Producers (both Primary and Secondary Producers) and the context of information consultation for data Consumer.

3. A Typical Example In Technological Watch

The figure 5 presents an ideal example of information pre-processing with the VIGIWARE system. Let's consider a commercial product which is described by three Information Sources :

- IS1 is a commercial brochure,
- IS2 is a Technical Report,
- IS3 is the report of the design engineer, one of our best friends.

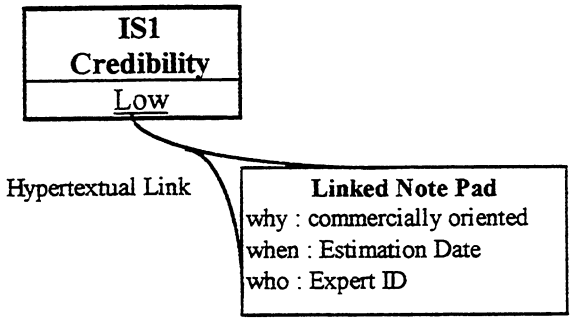
IS1 has been edited at T1 and it's the most recent Information Source, before IS3 and IS2 . The common characteristic described by IS1, IS2 and IS3, is the product satisfaction. For the simplicity of the example, this item of information is expressed with pourcentages :

- IS1 proposes around 100% of satisfaction,
- IS2 proposes 65.67% of satisfaction,
- IS3 proposes around 60% of satisfaction.

The information pre-processing in the VIGIWARE system is composed by four steps :

Step 1) Information Fusion
The Expert E1 of the TWG stores the 3 items of information in the VIGIWARE system (and if necessary, resolves the existing scale or naming conflicts, domain mismatch...). Product Satisfaction is now a 3-valueted attribute in the DBMS, composed by the different values, Information Source IDs and IS Edition Dates.

Step 2) IQ & ISQ Assessments
Each value is estimated according preponderant quality criteria (IQ and ISQ) defined by the Expert E1. The weighting of assessment is (Low, Medium, High) but it could be refined.
Each estimation depends on the expert but it is justified in a linked Note Pad (Text type). For example, credibility of IS1 is Low according to the Expert E1 because the objective of IS1 (as a commercial brochure) is to sell and it may exaggerate information about perfomances of the product ; it's a commercial argument.



Step3) Input of Information Customer Preferences
Before querying the VIGIWARE system, each customer defines and weights the most relevant quality criteria according to its needs. For example, IC1 gives priority to Timeliness. The Consumer IC2 focuses on Credibility (Medium), Technical Competence (High) and Accuracy (High). The Consumer IC3 focuses on Credibility (High), Timeliness (Low), Technical Competence (High) and Accuracy (Low).

Step 4) Personal Recommendation of Value-Added Data
Finally, relevant value-added data meant for a specific consumer's profile are recommended. This process is enabled by the matching between estimated Information Quality criteria by the Expert E1 (as Secondary Producer's perceptions) and specified Information Quality criteria by the Information Consumers IC1, IC2 and IC3 (as customer's requirements). The matchings are represented with the following colours :


for IC1
for IC2
for IC3

A particular data is recommended for each Consumer when it satisfies a maximal number of weighted Preferences he defined. Each data is proposed with its quality meta-data and each meta-data is hypertextually linked with the corresponding note pads for justifications or further explainations.

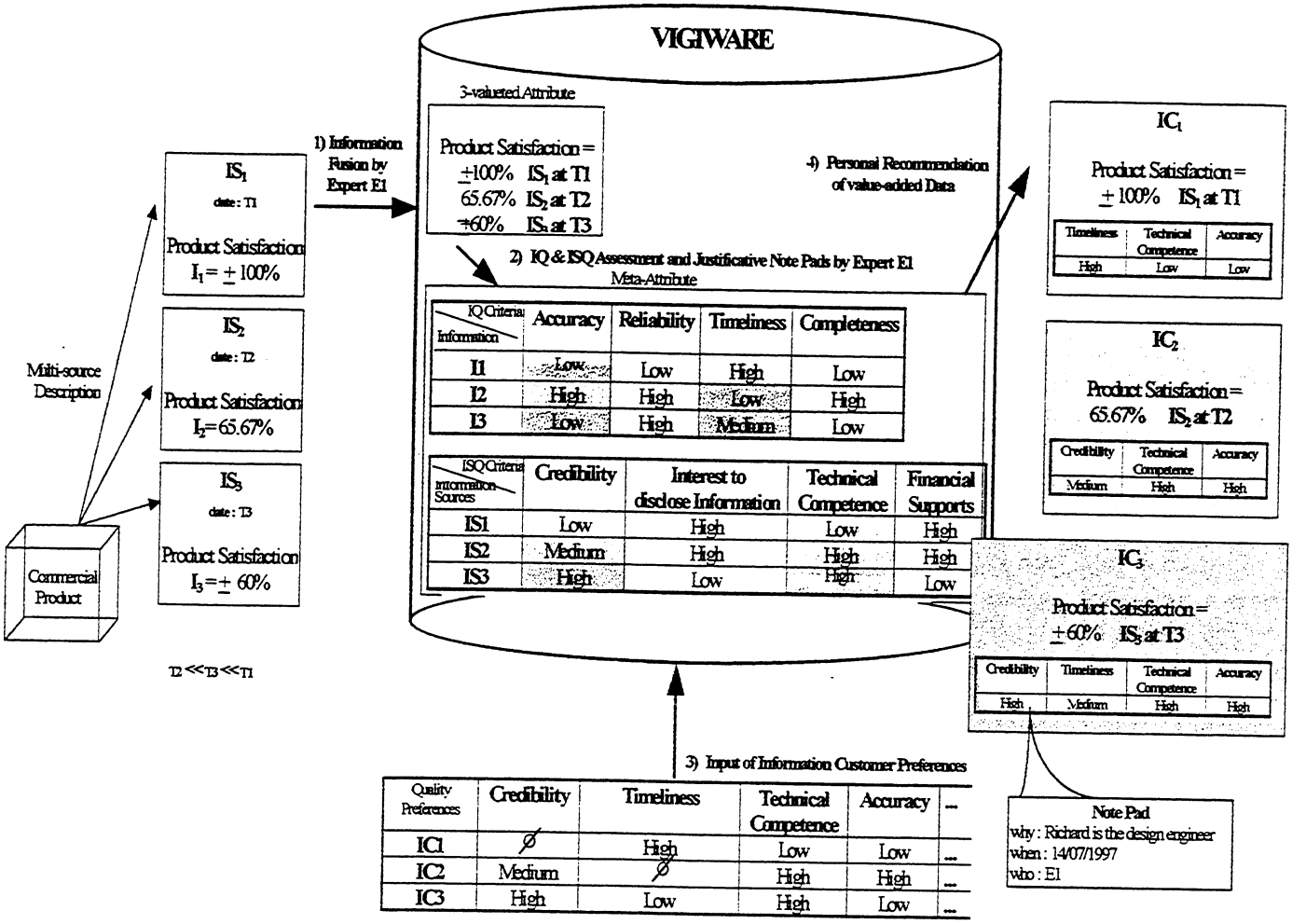


fig. 5. Information Pre-Processing in the VIGIWARE System

4 Conclusions And Perspectives

This paper presented our practical approach for extending DBMS functionalities for Technological Watch applications. This necessarily includes the consideration of the Quality of Information relative to specific Information Producers and Consumers' profiles and the multi-source and contextual aspects of information. Our ongoing work adresses the issues of information context (see [SaLiGa]) and these related to non-intrusive user profile acquisition and « user-oriented » information filtering and retrieval (see [RaBe] and [HuSt]). The future concerns of our work will focus on extending our formal model, implementing and validating it in a new version of the existing VIGIWARE system.

This work has been supported by AERO (France) under contract PA/2594.0107.

5. References

[BaTa] Ballou D. P., Tayi, K. G., "Methodology for Allocating Resources for Data Quality Enhancement", Communications of ACM, 32, 3, (1989), pp. 320-329.

[ClWi] Clemen, R. T., Winkler R. L., "Limits for the Precision and Value of Information from Dependent Sources", Operations Research, 33, 2, (March-April 1985), pp. 427-442.

[Firth] Firth, C., http://sunflower.singnet.com.sg/~cfirth/dataquality

[HuSt] Huffman, S. B., Steier D., "Meta-Information for Knowledge Navigation and Retrieval : What's in there", Working Notes of the 1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, (1995).

[ISO] ISO9000 Int'l Standards for Quality Management, Int'l Organisation for Standards, (1992).

[RaBe] Raskutti, B. and Beitz A. "Acquiring User Preferences for Information Filtering in Interactive Multi-Media Services", Proceedings of 4th Pacific Rim International Conference on Artificial Intelligence, Lecture Notes in Computer Science, 1114, Springer-Verlag, (1996), pp. 47-58.

[SaLiGa] Satur, R., Liu, Z.Q., Gahegan, M., "Multi-Layered FCMs Applied to Context Dependent Learning", Proceedings of the Intl. Workshop on Fuzzy Database System and Information Retrieval, Yokohama, Japan, IEEE Transactions on Knowledge and Data Engineering, (july 1995), pp. 561-568.

[WaKoMa] Wang, R. Y., Kon, H. B., and Madnick, S. E., "Data Quality Requirements Analysis and Modeling", Proceedings of the Ninth International Conference on Data Engineering, (1993), pp. 670-677.

[WaStFi] Wang, R. Y., Storey, V. C., and Firth, C. P., "A Framework for Analysis of Data Quality Research", IEEE Transactions on Knowledge and Data Engineering, 7, 4, (august 1995), pp. 623-640.

---

* These can be compared to the perlocutionary effects identified by J. L. Austin in *How To Do Things With Words*, J.O. Urmson (Ed.) Oxford University Press (1962).