Measuring Information Quality in Collaborative Business Intelligence Networks

Jens Kaufmann
Department of Technology and Operations Management
University of Duisburg-Essen
Duisburg, Germany
jens.kaufmann@uni-duisburg-essen.de

Abstract—**Collaborative business intelligence in the meaning of cross-company data sharing and analysis can be conducted by the use of collaborative business intelligence networks and a peer-to-peer-approach. Despite the pure technological possibility, difficulties exist due to different data schemes and the necessary semantic mappings of them leading to information loss. We propose methods and measures to quantify the information quality of those networks and show first results of a prototypical simulation regarding local and global measures. We further outline research for future work.**

Keywords-Collaborative business intelligence; information quality; quality measures; peer-to-peer networks.

I. INTRODUCTION

Business intelligence (BI) has become a well-accepted and important part of business as of today. The main concept used in its context is the data warehouse (DW) [1]. This is often understood as a central point of structured, well-formatted data that is optimized for multi-dimensional analyses. It can be realized using a single database, but also may be scattered in different systems that all rely on the same scheme [2]. While those solutions are common in companies and their different departments, collaboration mechanisms only have gained attention over the past few years. The understanding of collaborative business intelligence (CBI) is still ambiguous. Some authors propose a definition that combines existing BI systems (i.e., systems for reporting, ad-hoc analysis, data mining, etc.) with collaboration techniques as seen in online social networks (sharing, 'liking', linking, rating, etc.) [3][4]. Others formulate an approach that involves different companies that share data for analyses or even work together on the analyses themselves [5][6].

We understand CBI in the latter way and take a look at the networks used for data sharing and combining. With the assumption that there does not exist a single scheme that is used by all companies involved, rules for matching the data of one company to at least one of the other companies have to be defined. A 'match' in this context is a successful mapping of information about an object in on company's view to a corresponding object in another company's view. It is very likely that in a situation like this no perfect match can be achieved, meaning some data can either not be transferred or received in the way it is supposed to or cannot be matched to the other companies' schemes at all [7]. While different approaches have been discussed to overcome the difficulty of creating matching tables for bigger data structures, the aspect of measuring how effective or well data can be shared, has not been a major topic of research so far in the field of BI.

In [8], the authors propose a peer-to-peer (P2P) network approach to build CBI networks among different companies. An example of practical use is given by a net of universities, exchanging information about research funding. The authors argue that P2P networks provide a maximum of autonomy for every participating partner and that matching tables between partners do not have to be built for every possible connection. Furthermore, the lack of a central scheme reduces dependencies of unanimous verdicts on how to share and organize data. They do not describe how those P2P networks should be organized and do not take into account the different strategies companies could pursue to minimize personal effort regardless of the overall quality of information in the network. To develop global strategies or basic principles that describe, how companies could (or should) choose their matching partners to maximize their and the overall information quality, means must exist to quantify information quality first. Two main problems are therefore identified and dealt with in this paper:

(1) How can information quality in P2P CBI networks be measured?

(2) How do different P2P CBI network structures affect the information quality, regarding the measures mentioned?

Section II will give a brief description of the state of the art in CBI nets. Considerations of quality measurements in CBI nets are discussed in section III, while section IV deals with the possibilities of influencing quality during the CBI net generation. We give a brief overview of first results with a prototypical simulation of P2P CBI nets and close with our plans for future research in Section V.

II. STATE OF THE ART AND PROBLEMS IN CBI NETWORKS

A comprehensive classification and state-of-the-art analysis on CBI has been given in [9]. It shows that most of the publications derive their understanding of CBI from a technical perspective and focus on additional collaborative functions or technologies in existing BI systems. Some approaches, however, give different views on inter-company

 ISBN: 978-1-61208-351-3

collaboration and explicitly state that CBI is collaboration in the analysis process or parts of it rather than just communication over private analyses. While some publications only describe the idea of collaboration [6][10], others give more detail on possible implementations or architectures [11]. One of the most often cited publications is [8], where a "Business Intelligence Network" is defined and different architectural approaches are discussed, varying from a central data warehouse accessed by all partners to a completely loose-coupled P2P approach. The authors come to the conclusion that a P2P-based network is most effective for the specific use as a cross-company collaboration tool in BI. As BI systems usually keep most sensitive data about business developments, detailed revenues and other competition-relevant facts, most companies would like to share only parts of their data. The reason they do it at all is to (a) gain insight into the market at the cost of revealing a little bit of their own knowledge or (b) create alliances and/or supply chain partnerships where shared knowledge adds value to all companies' information base. Nevertheless, those business networks may work on a timely limited or project basis like, e.g., the automotive parts industry sometimes does [12]. Therefore, an easy entry into those networks has to be given as well as an opportunity to keep full autonomy of all shared data. In a BI context, data is often organized in a multidimensional cube, spanned by different dimensions that hierarchically structure attributes to describe data. Publications considering CBI networks or cross-company discussions about data of that type often assume that a common scheme (like a common 'cube') is created and used. Then, P2P-based networks can function without any translation schemes between the partners.

A more common and realistic version of dimensions in different systems is given with the example in Figure 1. It shows two versions of a geographical hierarchy. In this example, all attributes in the dimensions are organized in three levels, but that organization is company-dependent, so

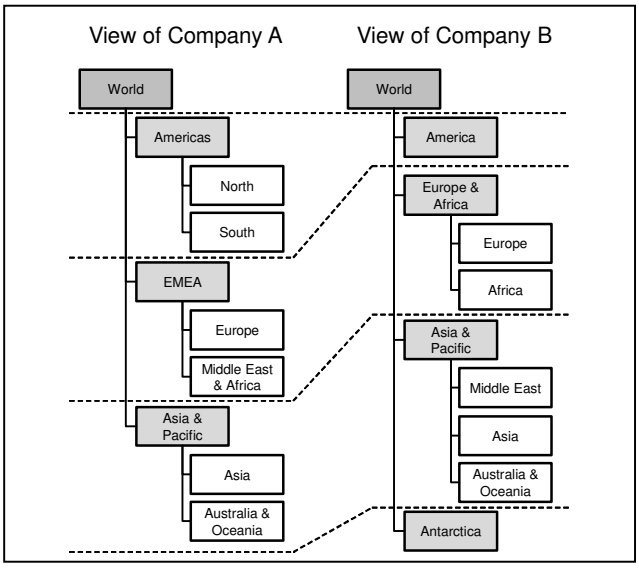

Figure 1. Matching problems between company schemes.

that different companies may use completely different 'structures of the world'. For a transfer of data from company A to B it can be seen that (a) the information about Americas loses granularity, (b) aggregations for EMEA and Asia&Pacific are not fully comparable, and (c) information about Antarctica cannot be transferred at all. Because this happens in nearly every DW integration project, different (semi-) automatic matching algorithms between dimensions have been proposed to create a global scheme or a translation table for different schemes [13][14][15]. Depending on the differences between the schemes, translations can be found more or less completely and information can be lost, when one partner keeps data at a higher level of aggregation than another partner.

III. POSSIBLE MEASURES FOR CBI NETWORKS

Matching heterogeneous data(base) schemes in general is a well-known problem. Matching algorithms for multi-dimensional data, however, are still under development and improvement; measures have been proposed sparsely as discussed in [15] and [16]. On a dimension level, three properties for a matching were proposed in [7]. The authors use the following terms to describe them: A 'level' is meant in a hierarchical way. So the top node of a hierarchy, unifying all underlying elements, is the first level. All of its descendants (or children) form the second level and so on. In the given example, level one is formed by 'World' and level two is (for company A) a view of world regions, consisting of 'Americas', 'EMEA' and 'Asia & Pacific'. If information of a lower level is aggregated in a higher level, it 'rolls up' to the higher level. In the example, the figures of all world regions roll up to 'World'. The properties for matching now can be described by:

Coherence: If in scheme A level l_1 rolls up to level l_2, then the matching levels to l_1 and l_2 of scheme B must roll up the same way.

Soundness: If there is a matching between levels in A and B, then all elements can be matched.

Consistency: The function defining the roll-up for all members in each level is the same for scheme A as for B.

A *perfect matching* is achieved, if all constraints apply.

In [16], the authors propose the concept of *strictness* to ensure usable mappings for BI systems. Strictness is acquired, if every member rolls up to at most one member of the parent level. This prevents double counting of elements which is crucial for, e.g., summing up revenues. To check for good matches, a *similarity score* based on the Similarity Flooding algorithm [17] was used and complemented by a *match factor* φ that is computed by taking matches of lower levels and elements into account, assuming that a chosen mapping is more likely to be a good match if the lower levels have a high match count, too. Similar ideas can be found when checking for duplicates in XML structures (which can be presented as hierarchical graphs) [18]. All of these approaches target on finding acceptable matches for automatic schema mapping, while only a few consider the measure of the fitting itself a main issue.

For these *local* dimension mappings, i.e. mappings without regarding other existing dimensions and/or partners,

 ISBN: 978-1-61208-351-3

some of the proposed matching factors or a quantified proportional fulfillment of the desired properties could be used as measures (e.g., 'How many dimensions are sound?' or 'How many percent of elements fulfill the consistency property?'). Taking into account that CBI networks do not work on a one-on-one base only, but do rely on multiple chained scheme translations, *global* measures have to be used to define, whether a CBI net is useful for all (or most of the) participants.

We use the term 'information quality' to describe the possible value of data exchange between partners as data from partners only becomes really useful if it can be matched to the schemes or structures used by a company itself, transferring it from data to information. We acknowledge, however, that the term 'information quality' does not have a single, undisputable definition and refer loosely to the ideas of [19], where information quality is defined by dimensions like ‚accesibility', ‚completeness', and relevancy'. To achieve high rankings on these dimensions, a CBI net must be designed in a useful way which leads back to the question on how to measure the quality of the net.

Figure 2 shows a small net of eight nodes and their connections, i.e., existing translation tables. We assume that a local measure (for simplicity: a function $\alpha_{XY} \rightarrow [0,1]$ with X,Y ∈ CBI net nodes) has already been defined. α_{XY} is a [0,1]-normalized quality measure, with α near to 1 if X can transfer data to Y with only a little information loss. Due to the use of aggregation functions it can easily be seen that most often $\alpha_{XY} \neq \alpha_{YX}$. α-values are provided for four exemplary nodes and their connections. When considering good routings for data, α-values are complex to handle. Unlike in, e.g., internet traffic routing, α-values cannot be simply multiplied or used to identify a bottleneck as it is not clear, which parts of information get lost at each node.

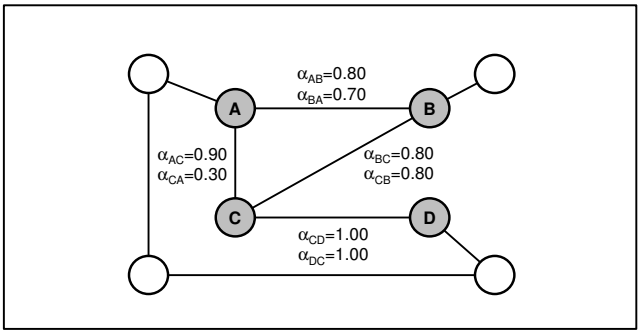

Figure 2. Exemplary CBI net with α-values.

Therefore it is not easily computable, if DCBA would be a 'better' way to send BI data from D to A than DCA (naively assuming that the low α_{CA} is a major problem of the net). To the best of our knowledge, neither detailed local measures for multidimensional data nor global measures for CBI networks have been developed – always considering a high information quality for multidimensional, hierarchical data. We are currently working on measures to overcome the presented issues and bring the following hypotheses up for discussion: (a) valued properties are the amount of directly assignable members and the degree of granularity kept up (because elements carry information and the more detailed they are, the more detailed the information can be presented), and (b) a global measure is crucial to determine a good structure of the whole net and to detect a reasonable relation between 'effort for creating mappings' and 'information quality for all partners' (because local measures only optimize direct connections instead of an information flow via different peers).

IV. INFLUENCING INFORMATION QUALITY IN CBI NETWORKS

To effectively influence quality, measures have to be identified. Otherwise, the effect of any means cannot be determined. Also, it can easily be seen that the simplest methods to ensure high quality may be impracticable: For example, if every partner defined a translation to every other, the effort needed to keep those translations running would outweigh the use of the net considerably. Another 'easy' solution is a 'star scheme' of the net, i.e., defining the partner with the most detailed scheme as the center of the net and (only) translating to this scheme. For one thing this would contradict to the autonomy aspect; for another thing it would crucially reduce the robustness of the net. If the center node fails or simply leaves the net, the net is not able to deliver any information. Building a useful net therefore has to take all aspects into account, i.e., quality of and effort for translations, robustness, and autonomy.

To get a first impression on how choosing neighbors in a net influences the overall quality, we created a simplified simulation of the evolution of a CBI network. The settings are as follows: The number of nodes n is set to 10, 20, and 30. The number of new connections each new node makes is varying from 1 to 4, but the same for every node. There exists a value $\beta_{XY} \in [0,1]$ defining the 'completeness' of a mapping, a value $\gamma_{XY} \in [0,1]$, defining the granularity kept (β, γ ∈ [0.4,1]), and the assumption that α=(β+γ)/2 is somewhat simple, but sufficient for a first simulation of the whole net. In further work we plan to create comprehensive 'master' dimensions in all nodes and a full simulation of the effect of reduced dimensions with automatic mapping. For simplicity, this time we assume that on a path through the net, γ can be treated as a 'bottleneck'-variable (meaning the lowest γ-value counts for the path, as the loss of levels cannot be repaired) and β only takes a 50%-effect at a query on each node it passes through the net. (An example: If $\beta_{AB} = 0.8$ and $\beta_{BC} = 0.4$, then the calculated $\beta'_{AC} = 0.8*(0.4+(0.6)/2) = 0.8*0.7 = 0.56$, as the 0.6 information loss between B and C only affects 50% of the relevant query data.) Of course, assumptions and values are disputable and more thorough studies will be conducted. Finally, $\delta_{AB} = \max(\alpha_{AB1, .., } \alpha_{ABm})$ with 1..m describing all possible connections between A and B and the overall quality is $\Delta = \sum\delta/(n*(n-1))$. With this setting, we evaluated three scenarios for a linear build-up of the net. First, random translations were built, i.e. random β- and γ-values were created. Second, every new node connected to the most connected nodes in the net (on parity to the ones with the lowest index), creating a star scheme. Third, every new node A connected to the best fitting other node(s) B_1, B_2, … regarding α_{BA} (i.e., data reception is

 ISBN: 978-1-61208-351-3

valued higher than data delivery). Our findings are presented in Table 1.

TABLE I. RESULTS OF A PROTOTYPICAL SIMULATION FOR CBI NETS (Δ-VALUES FOR DIFFERENT PARAMTER COMBINATIONS)

n	Sce-nario	Number of connections			
		1	*2*	*3*	*4*
10	**1**	0.5423	0.6823	0.7471	0.7794
	2	0.5999	0.7481	0.7552	0.7667
	3	0.5569	0.7292	0.7839	0.8020
20	**1**	0.4969	0.6502	0.7185	0.7506
	2	0.6072	0.7407	0.7510	0.7739
	3	0.5195	0.7262	0.7861	0.8235
30	**1**	0.4678	0.6346	0.7044	0.7386
	2	0.5987	0.7244	0.7318	0.7541
	3	0.4947	0.7205	0.7820	0.8101

They show that higher connection counts lead to better results, which naively seems to be natural. The changes from bad to good quality are quite similar for every net size. When the number of connections exceeds two, scenario three (best-fitting nodes) leads to better results than a random or 'star' approach. Considering that not the overall Δ was optimized, but a greedy approach was taken, this is not obvious and provides an interesting basis for further research.

V. CONCLUSION AND FUTURE WORK

We showed that P2P-based CBI networks can provide useful information for autonomous companies in supply chains or strategic alliances. Measuring the quality of translations between partners and defining the overall quality of the CBI net is most important to ensure a reasonable structure of the net. Only a few measures for dimension mappings exist and those cannot be directly transferred to CBI nets. Concerning our research topic (1), we therefore evaluated basic principles for more sophisticated measures. In respect to (2) we showed with a simple prototype that, when entering a net, building 'easy' translations does not always lead to an efficient CBI net from a global perspective. Further research will be directed to a comprehensive definition of information quality measurement in CBI nets and recommendations on how to choose directly connected partners wisely.

REFERENCES

[1] W. H. Inmon, Building the Data Warehouse, 4th edn., Wiley, Indianapolis, 2005.

[2] R. Kimball and M. Ross, The data warehouse toolkit: The complete guide to dimensional modeling, 2nd edn., Wiley, New York, NY [u.a.], 2002.

[3] A. Bitterer, "Hype Cycle for Business Intelligence, 2012", Gartner RAS Core Research Note G00227572, 2012.

[4] M. Muntean, "Business Intelligence Approaches", Mathematical Models & Methods in Applied Sciences, Vol. I, 2012, pp. 192–196.

[5] M. Golfarelli, F. Mandreoli, W. Penzo, S. Rizzi, and E. Turricchia, "BIN: Business intelligence networks", in Business Intelligence Applications and the Web, M. E. Zorrilla, J. N. Mazón, Ó. Ferrández, I. Garrigós, F. Daniel, and J. Trujillo, Editors. 2011. IGI Global.

[6] L. Liu and H. Daniels, "Towards a Value Model for Collaborative, Business Intelligence-supported Risk Assessment", in Proceedings of the 6th International Workshop on Value Modeling and Business Ontology (VMBO 2012). 2012, pp. 1-5.

[7] R. Torlone, "Two approaches to the integration of heterogeneous data warehouses", Distributed and Parallel Databases, 23(1), 2008, pp. 69-97.

[8] S. Rizzi, "Collaborative Business Intelligence", in Business Intelligence, M.-A. Aufaure and E. Zimányi, Editors. 2012. Springer Berlin Heidelberg.

[9] J. Kaufmann and P. Chamoni, "Structuring Collaborative Business Intelligence: A Literature Review", in Hawai'i International Conference on System Sciences (HICSS-47), Waikoloa, Hawaii. 2014, pp. 3738-3747.

[10] T. Mettler and D. Raber, "Developing a collaborative business intelligence system for improving delivery reliability in business networks", in 17th International Conference on Concurrent Enterprising (ICE). 2011, pp. 1-20.

[11] V. A. Martins, J. P. L. daCosta, and R. T. deSousa, "Architecture of a Collaborative Business Intelligence Environment based on an Ontology Repository and Distributed Data Services", in Proc. 4th International Conference on Knowledge Management and Information Sharing, Barcelona, Spain. 2012, pp. 99-106.

[12] C. Scholta, Success factors of inter-company cooperation using the example of the medium-sized automotive supply industry in Saxony (Erfolgsfaktoren unternehmens-übergreifender Kooperation am Beispiel der mittelständischen Automobilzulieferindustrie in Sachsen), Chemnitz, 2005.

[13] A. A. Vaisman, M. Minuto Espil, and M. Paradela, "P2P OLAP: Data model, implementation and case study", Information Systems, 34(2), 2009, pp. 231–257.

[14] M. Banek, B. Vrdoljak, A. M. Tjoa, and Z. Skocir, "Automating the Schema Matching Process for Heterogeneous Data Warehouses", in Data Warehousing and Knowledge Discovery, 9th International Conference, DaWaK 2007, Regensburg, Germany, September 3-7, 2007, Proceedings, I.Y. Song, J. Eder, and T.M. Nguyen, Editors. 2007, pp. 45-54.

[15] S. Bergamaschi, M. O. Olaru, S. Sorrentino, and M. Vincini, "Semi-automatic Discovery of Mappings Between Heterogeneous Data Warehouse Dimensions", International Journal on Information Technology, 1(3), 2011, p. 9.

[16] D. Riazati, J. A. Thom, and X. Zhang, "Enforcing strictness in integration of dimensions: beyond instance matching", in Proceedings of DOLAP 2011, ACM 14th International Workshop on Data Warehousing and OLAP, I. Y. Song, A. Cuzzocrea, and K. C. Davis, Editors, Glasgow, United Kingdom, 08.10.2011. 2011, pp. 428-438.

[17] S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity flooding: a versatile graph matching algorithm and its application to schema matching", in 18th International Conference on Data Engineering, San Jose, CA, USA, 26 Feb.-1 March 2002, pp. 117-128.

[18] P. Calado, M. Herschel, and L. Leitão, "An Overview of XML Duplicate Detection Algorithms", in Soft Computing in XML Data Management, J. Kacprzyk, Z. Ma, and L. Yan, Editors. 2010. Springer Berlin Heidelberg: Berlin, Heidelberg.

[19] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information Quality Benchmarks: Product and Service Performance", Commun. ACM, 45(4), 2002, pp. 184–192.

 ISBN: 978-1-61208-351-3