Towards Semantic Data Analysis

Mohsen Rais-Ghasem, Robin Grosset, Martin Petitclerc, Qing Wei

IBM Canada Ltd.

Ottawa, Ontario

Abstract

Semantic-oriented data analysis techniques are rapidly gaining momentum in the information processing industry. In the Business Intelligence sphere, the products are moving away from traditional ETL paradigm that requires up-front preparation and modeling efforts, and towards more interactive and discovery-like analysis with the overall goal of achieving more insightful analyses faster and with fewer preparation steps.

In this paper we outline such an analysis framework in whose heart lies a semantic annotator system that employs business ontology to make sense of data. The outcome is then passed to an analysis expert system that consults a set of declarative analysis guidelines to automatically produce useful analyses.

1 Introduction

The recent advents in knowledge engineering technologies have inspired a generation of systems that leverage semantic in variety of applications to achieve smarter behavior. Similar trend is evident amongst Business Intelligence (BI) systems which try to offer more insightful and targeted analyses faster and without extensive data preparation and modeling (see next section for a short overview). In this paper we discuss a semantic data analysis framework we have been working on since 2012. The objectives of our work can be summarized as the followings:

- To significantly shorten the efforts and time required to analyze data by moving away from traditional reporting towards discovery and exploratory analysis.
- To build the analytic expertise into the system to make analysis available to a wider range of users beyond traditional report authors and data scientists.
- To provide easy and inexpensive means to extend and customize the system behavior so it can be used easily in different domains.

The smartness in our framework is grounded in a set of business and domain ontologies and set of analysis rules. Together, they provide means to understand the data and capture typical analysis patterns.

The paper breaks down as follows: Section 2 semantic reviews some of work in this area. Section 3 provides a more detailed view of the proposed semantic framework, where the ontologies used to annotate metadata, and the annotation process itself are discussed, along with the analysis expert sub-system, and a simple search-like interface that enables users to look for specific analyses. Section 4 contains concluding remarks and plans for future work.

2 Related Work

The use of semantic and specifically ontology among information processing systems is on the rise. Ontology in some BI systems is used primarily to capture knowledge about a specific domain. For example [4] outlines a credit risk analysis portal based on a set of financial ontologies. Similarly, [10] proposes an infrastructure that draws on series on ontologies to coordinate and

effectively manage information as it flows between various elements of a healthcare system. Such systems often incorporate a reasoning engine that is capable of making inferences following set of rules within the stated goals and restrictions of their specific domains (see [4]).

The focus of ontology in many BI systems is to facilitate data integration form various sources [3, 6, 11]. For example, [6] discusses a decision support system that uses ontology to digest online news from various news sources. Ontology in this system is primarily used to mine data and discover relations.

The solution outlined in [2] uses ontologies primarily to address data integration challenges that face modern BI systems. Specifically, a shared ontology is envisioned that acts as interlingua to map information from other information sources (e.g. web or data sources, application, user) each described by its own dedicated ontology. And [5] investigates the use of ontology to enable a dynamic and automatic orchestration of complex business services, such as decision making processes.

The work described in [1] also uses ontology to map information between different aspects of BI systems, namely physical, analytical and conceptual. In this work, these aspects are accompanied respectively by an enterprise information system ontology, a data warehouse ontology, and finally a business ontology.

Semantic information and ontologies are also used to facilitate transformation of transactional data into a multidimensional warehouse (see [7, 8, 9]). For example, [7] discusses an ETL process to automatically build OLAP dimensions that reflect the hierarchic structure of ontologies. The framework described in [8] and its more recent revision [9] exploits 3 sets of ontologies (data source, analytic and domain) to automatically construct an OLAP cube. The following snippet shows the definition of analysis in terms of dimensions, measures, filters and so on ([8], page 3).

```
(defclass
Analysis () (
(has_description :type string)
(has_measures :type Analysis_Measure)
(has_dimension :type Analysis_Dimension)
(has_filter :type Analysis_Filter)
(has_parameter :type Analysis_Parameter)
(has_creator :type User)
(has_allowed_user :type User)
(has_allowed_role :type Role)))
```

While our work shares many of the goals and design principles with these works, it differs with them on some key aspects:

- Our ontologies comprise business concepts; from common concepts such as year, quarter, location to more *ad hoc* such as Sales Channel and Target Sales.
- Analysis in our framework is not tied to any particular data model, most notably OLAP. Hence, our metadata model is expressed in terms of categories, metrics, and navigation paths which are less overloaded terms. Moreover, our system has been used with variety of storage systems from columnar databases, to OLAP cubes and relational datasets.
- In our framework, the semantic annotation of data is carried out automatically, unlike some works that presuppose tagged data.
- And finally, our framework incorporates an analysis expert system that is capable of recommending and ultimately producing fast, insightful out of the box analyses.

3 Semantic BI Framework

Figure 1 gives an overview of the proposed Semantic BI framework and its components.

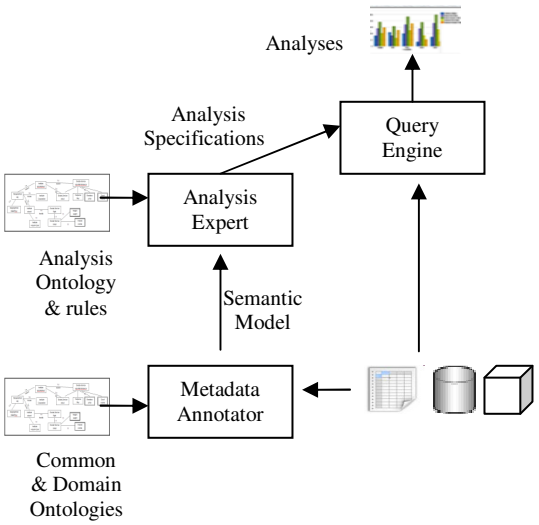

Figure 1: SBI Framework components

The framework relies on two declarative semantic sources. The first, common and domain

business ontologies, describes business concepts and their relationships; which is used by the metadata annotator to make sense of data, be it a semi-structured source, such as csv or xls dataset, or enterprise sources such as database or OLAP cube.

The second source, analysis ontology and rules, are used by an expert system that uses a semantic metadata model to recommend a set of analysis specifications, that is ultimately passed to a data-agnostic query engine that plan the individual analyses for their respective sources and return the results in some intuitive rendering, e.g. as visualizations. These components are further detailed in the following subsections.

3.1 Common, Domain Ontologies

Our common ontology focuses on business entities rather than any particular query paradigm. Our ontologies are represented as hierarchies of business concepts and their associations. The concepts in our ontology fall under one of three top classes: **category**, **attribute**, and **metric**[1].

Categories represent common or specific entities that businesses use to track or organize their information. Example of categories are products, geographies (city, country), or temporal (year, quarter). Attributes are associated with categories and describe different aspects of them (e.g. identifier, caption or longitude/latitude for a geographical location). Metrics are quantifiable indicators businesses use to measure performance along those categories.

Concepts can be related to each other either in a simple association (e.g. product has name, stock has an open, close prices) or in a whole-part relationship (e.g. country often consists of states/provinces, or year breaks down into weeks/months). Figure 2 illustrates a temporal hierarchy, where concept Month is defined as a specialization of concept Temporal, itself a derivation of root Category concept. Further Month contains Week and DayOfMonth.

[1] It is also possible to support concepts under multiple classifications which is needed for concepts such as age and temperature that could act both as metric or category in different analysis, as in average age for a department, or distribution of a department's employees by age.

Concepts in our ontology can further contain specific information such as default favorable trend for a metric (e.g. down for cost or up for revenue), common value ranges for age (e.g. [0-7], [12], [13-19] etc.).

The domain ontology provides more specific concepts for a targeted business domain (e.g. sales or human resources) or an analysis area (sales pipeline or finance planning). In addition to concepts and association, a domain specification can also include typical performance indicators (e.g. calculation for sales channel win rate) or common analysis patterns (e.g. overview of channel win rate for last n-quarter or by region). Specific goals can also be specified for a business domain (i.e. increase channel win rate by 10%). Such information is used by the analysis expert sub-system to recommend more targeted analyses and detect highlights.

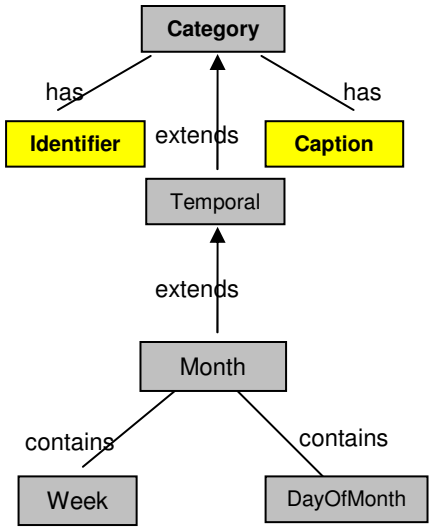

Figure 2: Definition of Month in the ontology

3.2 Semantic Annotator

The goal of this component is to make sense of a dataset in terms of categories, attributes, metrics and specific concepts in the common and domain ontology and build a meaningful semantic model. This is achieved by taking into consideration a variety of clues and hints, from data item labels, storage-specific information (e.g. whether an item participate in the key of a relational table), various

formatting and value decorators, and finally sample values.

Considering that labels and captions often provide the best insight into a what a data item represent, at the core of the semantic annotator lies a natural language processing (NLP) engine that utilizes a set of signifiers to map lexical clues to concepts in the ontology. The system has been trained to be sensitive to a variety of language-specific patterns, and nuances. For example, 'number' in English (but not in French or Japanese) is ambiguous, as in

a) Student Number
b) Number of Students

Our system in English detects (a) as 'identifier' item and (b) as 'count'.

Also, because one cannot assume that data item labels always have proper forms (consider column names in a table or headings in a spreadsheet), our system will recognize the following variations as well:

a*) Student_Num, Student No.
b*) No. Students, Num_Students

Earlier version of our system was very keyword-dependent, but over time the system has evolved to rely more on patterns. The following is a good example where a keyword-based approach would yield 3 metric candidates: expense, revenue, and ratio.

c) debt to income ratio

However, using an appropriate pattern, this label is correctly resolved to 'ratio'.

Following the initial lexical analysis, the candidates for each data item is corroborated using other clues such as data hints and sample values. For example, heuristically for a numeric item to represent a category identifier, certain data patterns are expected such as uniform values or non-negative numbers.

The final step is to organize the data items to represents the associations among the associated concepts in the underlying ontology. Simple statistical analyses (such as data correlation) are also used to discover and verify data dependencies. The result is a semantic model which is described in the following section.

3.3 Semantic Model

The semantic model consists of a series of nodes in a graph, where edges represent association (simple or whole-part). Each node also has a semantic designation in the form of a concept in one of the underlying ontology.

Consider the dataset partially shown in Figure 3 that contains information about cargo landed in various airports for years 2009-2010[2].

ST	Locid	Airport Name	City
TN	MEM	Memphis International	Memphis
AK	ANC	Ted Stevens Anchorage International	Anchorage
KY	SDF	Louisville International-Standiford Field	Louisville
FL	MIA	Miami International	Miami
IL	ORD	Chicago O'Hare International	Chicago
IN	IND	Indianapolis International	Indianapolis

Figure 3: Cargo Dataset (partial)

Parts of the semantic model built for this dataset is shown in Figure 4 in which we see a single airport category represented by two data items, 'Locid', and 'Airport Name'. Further, the semantic tagging in the model indicates that the first item represents the identifier for this category (concept *cIdentifier*) whereas the second item represents its title (*cCaption*).

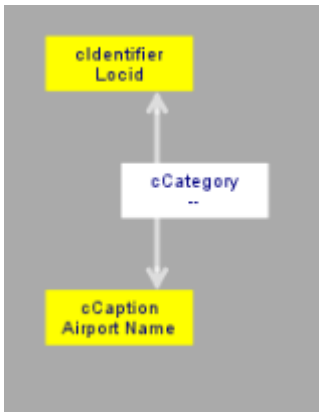

Figure 4: Airport Category

Such information, trivial as it may sound, is very useful. It greatly reduces the complexity of the

[2] See 'Qualifying Cargo Airports, Rank Order, and Percent Change from 2009" from http://www.faa.gov/airports

data and enables the metadata consumers to plan more efficient analysis (by using the identifier item) and better rendering (by using the caption).

As shown in Figure 5, our system is also able to recognize data items 'ST' and 'City' as instances of the whole-part relationship between ST (State) and City as categories representing state/province (*cStateProvince*) and city (*cCity*) and '% Change' as a ratio metric (*cRatio*). Such semantic designations are very valuable by themselves since they can enable metadata consumers to calibrate their handling of this data, for example by offering a map-based visualization. Moreover, the system also knows that state/provinces consist of cities, which is reflected in the model via a whole-part association (the dark-thick arrow between the two categories). Such associations can be leveraged by analysis expert system to further refine the proposed analyses, such as recommending a time cycle analysis when there are two temporal categories which have a whole-part relationship.

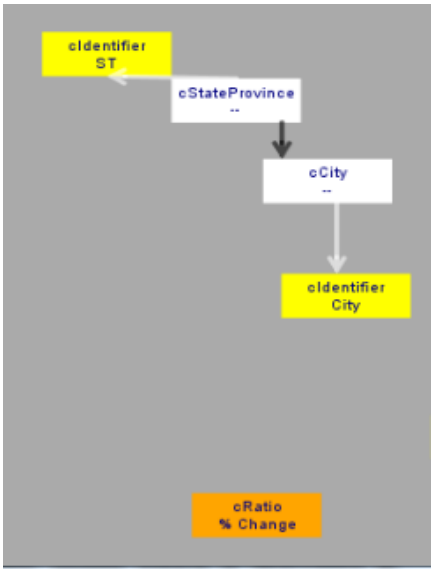

Figure 5: State-city categories, and Ratio metric

3.4 Analysis Expert System

The Analysis Expert System relies on declarative knowledge and rules to make recommendations on how best to analyze data corresponding to a semantic model.

Figure 6 illustrates a hierarchy of analysis types which the Analysis Expert System consults

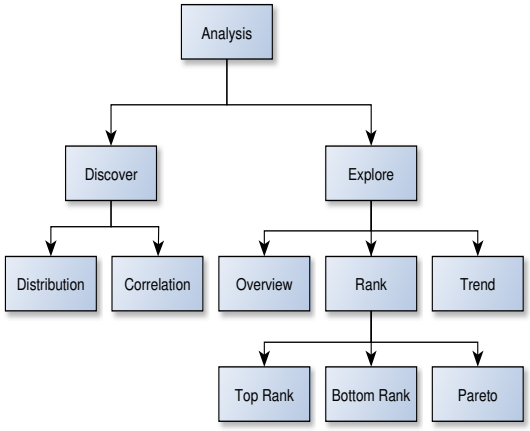

Figure 6: Analysis Types Hierarchy

.

It also uses a set of declarative rules that capture common or specific analysis methods or techniques. Each rule is specified along three sections:

- Data binding section which describes how to select eligible data items;
- Validation section that describes preconditions for an analysis to be applicable; and
- Scoring section that indicates how to measure the relevance and usefulness of an analysis, which is used when consolidating individual analyses in a single ordered list.

The rules used by the Analysis Expert System take into consideration a variety of factors, such as the specific concept of categories or metrics, the relationship between categories, the characteristics of the data behind the categories and metrics such as distinct value count. Figure 7 captures a few simplified steps followed to recommend a Top Rank Analysis.

The Analysis Expert System will first extract categories and metrics from the semantic model. To reduce the complexity, the categories are further filtered down using hierarchies from the model and user inputs. For the Top Rank analysis classification, it requires one metric and one category that have at least 20 distinct values (in order to rank the top 10). Once a valid set of categories and metrics for this analysis is identified, the system will then start generating different combinations.

The validation rules are then applied to each combination to remove the combination of data items that do not make sense for the current analysis. For example, if we have a comparison analysis between 2 categories, it doesn't make sense to compare sales between quarter and year. A validation rule will make sure that combination is invalidated.

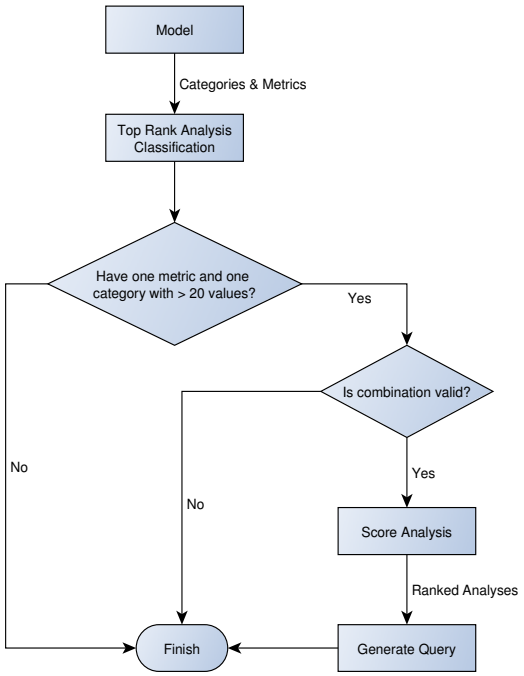

Figure 7: Top Rank Analysis Decision Tree

Once we have all the valid combinations for the analysis, the scoring rules will assign a compound final score for each combination based on data correlation, type of concept, data characteristics, and user inputs.

The set of analyses is then passed to the Query Engine which generates source-specific queries (e.g. for a relational database or an OLAP cube) which are subsequently executed and the results are returned to user. The query Engine is not the focus of this work and is not discussed any further.

Figure 8 contains examples of Top Rank analyses generated for the cargo dataset. The depicted analyses are:

- Top 10: 2009 Landed Weight (lbs.) by ST
- Top 10: 2010 Landed Weight (lbs.) by ST

3.5 Analysis Search Interface

Our system also provides a type-in interface that allows user to search or guide the analysis process. For example the user can type in just the name of one or more data items or their values, and the system will recommend only analyses that involve at least one of those items and adjust the ranking based on the number of matching items.

Figure 8: Top Rank Analysis results

The user alternatively can type in a question-like text, such as 'sales by region over last 3 quarters' and the system will recommend a trend over quarter analysis, but also other trend-related analyses (such as sales by product over year).

The NLP technology behind this feature is very similar o the one discussed in 3.2 except that the outcome is a determination of the overall analysis type, such as trend, as well a set of data items in the underlying data sets that are deemed most relevant to the analysis.

This determination takes into consideration the strength of the match, i.e. whether the items was selected on a literal (perfect) match between the input word(s) and a data item labels, or semantic (e.g. data item Sales_Value for input 'income') or match based on data values (e.g. State for 'NY').

For example in the previous airport cargo sample dataset, when user types in "compare landed weight 2009 with 2010", the search interface will return the following as the gist of the given question:

- Analysis Type: Comparison, Correlation.
- Relevant Data Items: '2009 Landed Weight' and '2010 Landed Weight' columns.

The Analysis Expert System can then use this information to select a subset of data items in the data source that are deemed most relevant to the

search phrase and generate more targeted analysis recommendations, such as:

- Correlation: 2010 Landed Weight (lbs.) and 2009 Landed Weight (lbs.) by Hub
- Comparison: 2010 Landed Weight (lbs.) and 2009 Landed Weight (lbs.) by RO

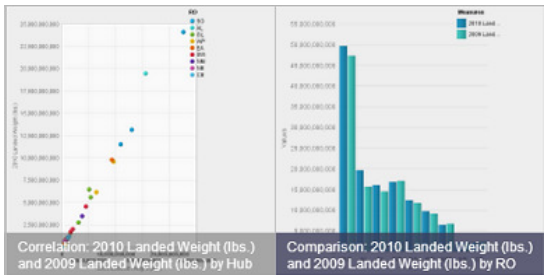

Figure 9: Comparison and Correlation Analysis

4 Performance & User Evaluation

A prototype of the system was developed over last year. The metadata semantic annotator alone was also incorporated in three separate products to build OLAP/ROLAP cubes and make visualization recommendations.

In our implementation, the common ontology comprised of about 60 concepts and 35 relationships, and the analysis expert system had roughly 20 rules.

The performance of semantic annotation step is a function of number of data items, and on a typical laptop computer (dual core, 8 GB RAM) varies from few milliseconds for a dozen data items to about 240 milliseconds for a dataset with over 200 data items. Similarly the analysis expert system performance, excluding data access, is in sub-seconds (e.g. 260 milliseconds for producing 16 analyses).

The prototype went through a number of user validations, using a variety of datasets from a wide range of domains such as sales/return data, crime statistics, insurance claims, and so on. Initially we had difficulty selecting a subset of insightful analyses from a multitude of possible analyses. The challenges were mitigated with a combination of refinements applied to the ranking of the generated analyses, and the invent of the search feature, which enabled users to guide and influence the analysis process.

The prototype was also utilized in specific domains by extending the business ontology with typical concepts and analysis templates in those domains. One such extension was in the area of social media analytics where our system was successfully deployed to replace a large set of carefully authored analyses with a small set of concepts and analyses templates.

5 Discussion and Conclusion

In this paper we described an analysis framework which is capable of greatly reducing the time-to-value in data analysis use cases. Our system is capable of providing a user with useful analyses out-of-the-box without requiring complex and time consuming modeling and ETL efforts.

Moreover, our framework makes analysis available and accessible to users that are not necessarily qualified as data scientists or professional data analysts.

Also, because it is largely driven by external, declarative rules and knowledge, the system behavior can be easily adopted for different domains. For example, the system is currently being customized for a consumer analytic application.

This work could be potentially extended to utilize various descriptive and predictive statistics to further assists users to discover interesting aspects in their data. For example, data mining algorithms could be used to discover what common factors drive the drop in the sales numbers; or automatically make predictions (and subsequently warn user) that given the expenses over the last 3 quarters it is very unlikely the cost reduction goals to be met.

About the Authors

Mohsen Rais-Ghasem is a senior software developer at IBM Canada. He received his PhD degree from Carleton University in 1988. He can be reached at mohsen.rais-ghasem@ca.ibm.com.

Robin Grosset is a IBM Distinguished Engineer working in Business Analytics segment at IBM Canada. He can be reached at robin.grosset@ca.ibm.com.

Martin Petitclerc is a software architect at IBM Canada. He has been working on various business analytics products and technologies, including OLAP databases, reporting tools and data mining. He can be reached at martin.petitclerc@ca.ibm.com.

Qing Wei is a software developer at IBM Canada. He has been working on various business analytics products at IBM. He can be reached at qing.wei@ca.ibm.com.

References

[1] Longbing Cao, Chao Luo, Dan Luo, C. Zhang. Integration of Business Intelligence Based on Three-Level Ontology Services. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pages 17-23, 2004

[2] Zhan Cui, E. Damiani, M. Leida. Benefits of Ontologies in Real Time Data Access. In *Digital EcoSystems and Technologies Conference, 2007. DEST '07. Inaugural IEEE-IES*, pages 392-397, 2007

[3] A.A. Fernandes, L.C. Amaro, J.P.C.L. Da Costa, A.M.R. Serrano, V.A. Martins, T. De-Sousa. Construction of Ontologies by using Concept Maps - a Study Case of Business Intelligence for the Federal Property Department. In *Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on*, pages 84-88, 2012

[4] S.B. Kotsiantis, D. Kanellopoulos, V. Karioti, V. Tampakas. An ontology-based portal for credit risk analysis. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 165-169, 2009

[5] A. Krupaviciute, J. Fayn. Ontology Driven Approach Enhancing Business Services Orchestration. In *Computational Intelligence, Communication Systems and Networks (CICSyN), 2010 Second International Conference on*, pages 344-348, 2010

[6] A. Martin, D. Maladhy, V.P. Venkastesan. A Framework for Business Intelligence Application using Ontological Classification. In *International Journal of Engineering Science and Technology (IJEST)*, pages 1213-1221, ISSN : 0975-5462, Vol. 3 No. 2 Feb 2011

[7] F. Sciarrone, P. Starace, T. Federici. A Business Intelligence Process to support Information Retrieval in an Ontology-Based Environment. In *Ninth International Conference on Intelligent Systems Design and Applications*, pages 896-901, 2009

[8] D. Sell, L. Cabral, E. Motta, J. Domingue, R. Pacheco. Adding Semantics to Business Intelligence. In *Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on*, pages 543-547, 2005

[9] D. Sell, D.C. da Silva, F.B. Ghisi, M. Napoli, J.L. Todesco. Adding Semantics to Business Intelligence: Towards a Smarter Generation of Analytical Tools. In *Business Intelligence - Solution for Business Development*, Chapter 3, ISBN 978-953-51-0019-5, February 2012

[10] X.S. Wang, L. Nayda, R.Dettinger. Infrastructure for a clinical decision-intelligence system. In *IBM Systems Journal*, Vol. 46, No 1, 2007

[11] Xu Xi ; Xu Hongfeng. Developing a Framework for Business Intelligence Systems Integration Based on Ontology. In *Networking and Digital Society, 2009. ICNDS '09. International Conference on*, pages 288-291, 2009