Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2011 Proceedings

European Conference on Information Systems (ECIS)

Summer 10-6-2011

USING THE MUTUAL INFORMATION METRIC TO IMPROVE ACCESSIBILITY AND UNDERSTANDABILITY IN BUSINESS INTELLIGENCE TOOLS

Yohai Sabag

Adir Even

USING THE MUTUAL INFORMATION METRIC TO IMPROVE ACCESSIBILITY AND UNDERSTANDABILITY IN BUSINESS INTELLIGENCE TOOLS

Sabag, Yohai, Ben-Gurion University of the Negev,P.O. Box 653, Beer-Sheva, 84105, Israel, yohaisab@bgu.ac.il

Even, Adir, Ben-Gurion University of the Negev,P.O. Box 653, Beer-Sheva, 84105, Israel, adireven@bgu.ac.il

Abstract

The rapidly-growing organizational data resources introduce a growing difficulty to locate and understand the relevant data subsets within large datasets – what can be seen as a severe information quality issue in today's decision-support environments. The study proposes a quantitative methodology, based on the mutual-information metric, for assessing the relative importance of different data subsets within a large dataset. Such assessments can grant the end-user with faster access to relevant subsets within a large dataset, the ability to better understandits contents, and gain deeper insights from analyzing it – e.g., when such a dataset is being used for Business Intelligence (BI) applications. This manuscript provides the background and the motivation for integrating the proposed assessments of relative importance. It then defines the calculations behind the mutual-information metric, and demonstrates their applications using illustrative examples.

Keywords: Business Intelligence (BI), Data Warehouse, On-Line Analytical Processing (OLAP), Data Mining, Mutual Information.

1 Introduction and Background

This study addresses a key information-quality issue in today's decision-support environments – the growing difficulty to locate the relevant data subsets within the rapidly-growing organizational data resources, and understand them properly toward supporting managerial decisions. The study proposes a quantitative methodology, based on the mutual-information metric, for assessing the relative importance of different data subsets within a large data collection. Such assessments can be used to help the end-user understand the data, and generate online recommendations for better usage of large datasets – particularly, when those datasets are being used in Business Intelligence (BI) applications. Better usage is conceptualized here in terms of *accessibility* – the ability to reach the relevant data subsets within a reasonable time, and in terms of *understandability* **-** the ability to comprehend those relevant data subsets and gain important business insights from analyzing them. Both accessibility and understandability have been identified as important data/information quality dimensions – however, they have not been investigated much in the context of decision-making and BI.

Recent years have witnessed a major shift in the approach toward managerial decision making, as many successful organizations attribute their success to the adoption of decision-making culture that promotes intensive use of data resource and the derived predictive analytics, toward gaining some major competitive advantage (Davenport, 2006). This transition has led to broad adoption of BI platforms and tools, which permit rapid development and broad distribution of data-driven decision support utilities. BI involves the acquisition, interpretation and analysis of data, using advanced visualization and presentation. Today, the contribution of BI toward enhancing decision making, improving business operations and, as a result, increasing business profitability and competitiveness, is well recognized (Wixom et al., 2008). The growing adoption of BI tools is largely driven by the vast amounts of data collected by organizations. Relevant data for decision making and business analysis can be found both in a plethora of sources (March and Hevner, 2007). In order to realize the full potential of the data collected, the data from the different sources are typically collected into a Data Warehouse (DW) – an integrated, cleansed, and well-formatted data repository (Watson and Wixom, 2007). A successful DW must be as broad as possible and, at the same time, ensure high data quality, reliability and accesability (March and Hevner, 2007).

To benefit from the investment in BI environment, an organization must form a culture of data usage among its employees in all units and managerial levels (Wixom et al., 2008). The success of BI initiatives depends largely on a broad distribution of BI capabilities throughout the organization, as it increases the exposure of employees to data and information and provides them with an infrastructure for decision-making support (March and Hevner, 2007). Advanced BI capabilities require some knowledge in data analysis and statistics skills; however, a common issue with the promotion of data-driven decision-making culture in organizations is that decision makers often lack these knowledge and skills (Davenport, 2006). Further, taking advantage of BI requires some familiarity with the plethora of data resources provided by the DW. With the immense growth of DW's, gaining such familiarity is becoming practically impossible for the common end-user.

In this study we suggest that gaining a broader adoption of BI requires a more fundamental change in the design of BI tools, and in the way that end-users interact with those tools. A possible approach for improving the usability of BI tools is the integration of recommender systems – textual and/or graphical utilities and visual cues that guide the end-user to consider using certain data subsets and/or analysis forms (Kolodner and Even, 2009). Recommender systems have been implemented successfully as an apparatus for coping with large amounts of information (Admoavicius and Tuzhilin, 2005). Recommender systems can be found today in many information-system contexts – particularly in e-commerce websites (Wang et al., 2007; Malinowski et al., 2008). However, despite their growing popularity, the application of recommender systems in BI environments is still rare.

In this study, we suggest that integrating a recommendation system, based on mutual-information metric, can improve the accessibility and understandability of BI tools – as such recommendations can help directing the end-user reach relevant data subsets, and analyze their content more precisely. Mutual-information metric, which are driven by entropy measurements, can identify data subsets in which the behavior of outcome variables is explained and differentiated better by a set of input variables. Such metrics are often used for data mining – however, they have not been examined in the contact of generating recommendations for better online data usage. According to the methodology proposed in this study, the mutual-information assessments are structured such that some of the calculations can be made in advance, during the preparation of the data for analysis, while others can be calculated on-the-fly, while the data is being used. Such a combination of front-end and back-end calculations can address the demand for short-enough response time, when the data is being analyzed online - e.g., within OLAP (On-Line Analytical Processing) BI utilities. In the reminder of this paper, we first provide the background and the motivation for the methodology developed towards improving information quality in BI environments. We then state the calculations behind the mutual-information metric, and illustrate how they can be used to generate recommendations in different OLAP-usage scenarios. We conclude by summarizing the study, and proposing directions for future research.

2 Conceptualization

2.1 Data Accessibility and Understandability in BI Environments

This study addresses a key issue with the usage and the adoption of BI – the growing difficulty to locate the relevant data within a large and complex DW, and the lack of data analysis and statistics skills needed to understand and gain value from the relevant data. We associate these issues with two data/information quality dimensions – *accessibility* and *understandability*, which have been recognized as being most important from the data consumer's perspective (Wang and Strong, 1996). Unlike other important data quality dimensions, they have not been studied extensively.

Accessibility reflects the extent to which the data consumer is able to access and retrieve the relevant data and information sources in a fast and convenient manner (Pipino et al., 2002). In the DW/BI context– we conceptualize accessibility as the ability of the end-users to reach the relevant subsets within a large data collection, such as a DW. We suggest that the issue of accessibility in a DW doesn't stem from technical difficulties – today's DW/BI technologies permit relatively-fast retrieval of subsets from a large dataset, and often provide easy-to-use utilities for forming the query that underlies such retrieval. However, when facing large and complex datasets, it is likely that end-users who are not familiar with the dataset contents would fail to recognize the relevant subsets. Moreover, accessibility – as we conceptualize it here - is highly context-dependent, as one subset may be relevant for certain decision-making scenarios, but irrelevant for others; hence, when attempting to aid the end-user with finding the relevant subset – the decision context must be understood.

Understandability (also referred to as clarity) reflects the extent to which data can be easily comprehended (Pipino et al., 2002). In the DW/BI context – we conceptualize understandability as the extent to which end-users can understand the data being provided, when being presented and visualized with a BI utility, and gain insights from it. As noted earlier, in-depth understanding of data often requires some advanced analysis and familiarity with statistical tools, beyond presentation of raw data, simple summary statistics, and basic visualization – skills that the common BI users do not possess. This lack of skills often hinders the ability to gain in-depth insights from data usage, and answer business questions at a high level of confidence. It is common the inexperienced BI user, in search of an answer to a business question, finds the BI tools too difficult to use for answering the question. Furthermore, BI users often find it difficult to even state correctly the right business question to ask, being unaware of the full range of DW/BI capabilities offered.

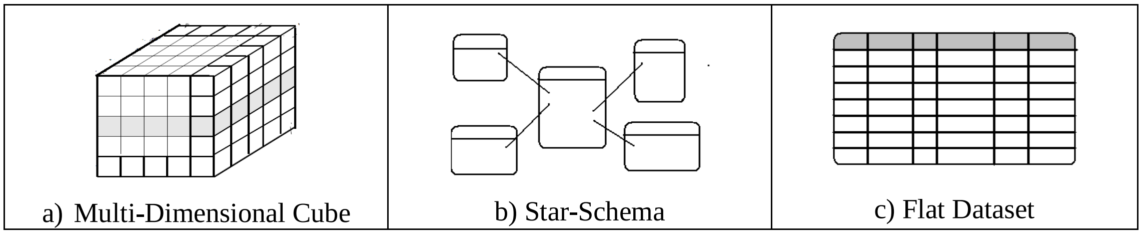
Higher data and information quality increases the usability of information resources and systems, and hence, the value gained (Wang and Strong, 1996). We therefore expect that improving the quality of BI tools – here, in terms of accessibility and usability – will improve their usability for decision making and the associated benefits, hence, promote greater adoption.

2.2 Data Representation and Usage Stylesin BI Environments

The design of data resources within a DW is directed by the need to answer BI queries (business questions, which can be defined via a BI tool), in a manner that can support managerial decision making. Such BI queries typically combine some basic "building blocks":

- ***Dimensions:*** important business entities, which are subject for analysis and comparison – e.g., "customers", "products", "locations", "employees", and "quarters" (or other time periods). Dimension data typically categorical, containing a finite set of distinct values (e.g., a list of the firm's "customers"). The data help for each dimension item typically contains a unique identifier and a few other descriptive dimension attributes (e.g., the age and gender of a "customer").
- ***Facts:*** metrics (mostly numerical) which can help assess and compare the performance of dimension items – e.g., compare "products" by the "items sold", "amount", and "production cost". BI queries typically present the facts in an aggregative form – e.g., "sum (items sold) per product", and "average (revenue} per store". It is also typical that some raw facts that are held in a DW (e.g., "items produced", "items sold"), will be used to derive higher-level calculated facts, often in a form of a ratio **-** e.g., "sale ratio = sum (items sold) /sum (items produced)". Derived facts in a DW are often associated Key Performance Indicators (KPI's) – metrics defined by the organization for assessing the success of business processes and units. It is therefore, common that the BI infrastructure is used for calculating and delivering information on relevant KPI's.
- ***Filters:*** conditions that limit the subset of data viewed and analyzed by the end-user. Filters can be set on dimensions, specifying a subset of distinct dimension items (e.g., show only "dairy products", in "regions A, B, and C"). Filters can also be set of facts, and be defined by a numeric range (e.g., show only "locations with total sale amount between $1M and $2M").
- ***Sorts:*** BI queries often require that the outcome will be sorted in a certain order, to emphasize the difference between the compared items. A typical sort will be defined as comparing the items within a certain dimension (or a combination of dimensions) along a certain fact – e.g., "sort the customers by a descending order of their total purchase amount".

BI environment, and the underlying DW, would typically incorporate data-representation models that address the need to form queries along certain dimensions, facts, filters, and sorts. Three of the common DW models (Figure 1), can be supported by the methodology developed in this study:



a) Multi-Dimensional Cube

b) Star-Schema

c) Flat Dataset

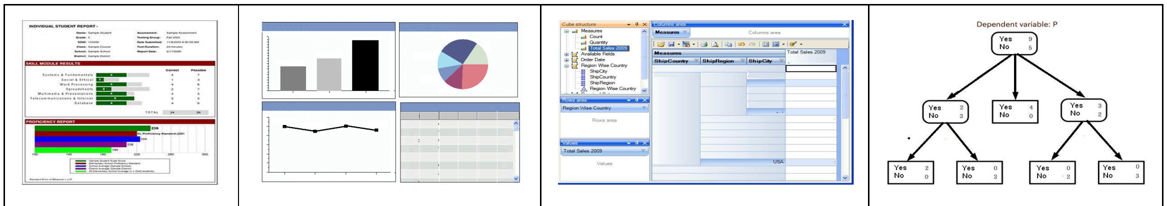
Figure 1. Data-Representation Models in a DW

- ***Multidimensional Cube:*** this model organizes the data in a multi-variable array, which conceptually resemble a cube structure. Each "side" of the cube reflects a dimension, and each cube unit reflects a specific combination of dimension items, and holds the set of facts associated with this combination (e.g., the "items sold" and the "total amount" for "product A" in "location B" in quarter C"). In real-world implementations of such cubes – some fact aggregations are typically pre-calculated, and stored within the cube to permit fast response-time to aggregative queries.
- ***Star-Schema:*** this model organizes the data in set of interlinked tables. The "center of the star" is a fact table, typically the largest within the schema, which contains the facts in a raw or aggregated form. The fact table is linked to a set of dimension tables – each containing the list of dimension items, and the associated attributes.

- ***Flat Dataset:*** this model organizes the data in a single "flat" table, where both the dimension attributes and the facts are columns within this table. Depending on the level of granularity, the flat dataset may contain detailed data (e.g., a single record per business transaction), or aggregated along a certain set of dimensions (e.g., the "items sold" and the "total amount" per each combination of "product", "location", and "quarter")

These three forms can be seen as equivalent datasets in terms of data contents and support for BI queries – all three permit aggregation of facts along a given set of dimensions, as well as filtering and sorting. Moreover, today's DW and ETL (Extraction, Transformation, and Loading) technologies permit relatively-easy conversion among these three forms. However, they may differ in their performance (The multidimensional cube, for example, is considered to be faster, in terms of data-retrieval speed), simplicity (the flat dataset is obviously the simplest representation among the three), convenience (the star-schema permit continent handling of dimensional data), and the database platform on which they can be implemented - the star schema and the flat dataset can be implemented with a standard RDBMS (Relational Data-Base Management System), while multidimensional cubes typically require dedicated MOLAP (Multidimensional OLAP) technologies.

The different models of data representation can support different forms of BI tools, and the software market offers a plethora of platforms that permit rapid development of such tools. *A report* (Figure 2a) for example, would run pre-defined query against a given data model, and present the output data in a pre-defined presentation format, which may combine different textual and/or graphical components. Similarly, a *digital dashboard* (Figure 2b) would combine several data visualization and presentation items that provide the manager with a snapshot picture of the organization's current state. Both reports and dashboards can be seen as relatively simple and static BI forms in terms of the required user intervention and analytical skills. Moreover, in both reports and dashboards, the business question addressed by the tool, the set of BI queries that translate it to data retrieval requests and the form in which the data is presented are largely predefined - and the tools offers only limited flexibility to change them. It is therefore common that BI tools that follow those forms are prepared in advanced by a unit of BI specialists, and then distributed among the end-users who are authorized to use them.



a) Report b) Digital Dashboard c) OLAP d) Data Mining

Figure 2. Business Intelligence Tools

In this study we focus more on two other common forms and BI tools, which can be seen as more advanced, in terms of the analytical skills required, and open-ended, in terms of the end-user's flexibility to define new business question and shape the presentation as desired:

- ***On-Line Analytical Processing (OLAP):*** OLAP (Figure 2c) is a common term for tools that permit interactive and dynamic investigation of data. When using OLAP, the user is granted with access to a certain dataset (e.g., a multidimensional cube, a star-schema, or a flat-dataset) and given the flexibility to navigate through the dataset, focus on specific subsets within it by applying certain filters, and view the facts within selected data subset aggregated along different dimensions. Generally, it can be said that OLAP permits the end-user to steer the analysis, and determine what data will be presented (Tremblay et al., 2007). OLAP tools have become an integral element of BI systems, and a plethora of studies have highlighted their benefits in terms of empowering the decision makers (e.g., Tremblay et al., 2007).

- ***Data Mining (DM):*** Data mining (Figure 2d) is a common definition for utilities, typically based on advanced statistical techniques and machine-learning algorithms, which analyze large datasets (typically, in a flat format) automatically in a search for valuable information and useful insights. This search can discover unusual patterns, highlight hidden relationships and interdependencies, and generate rules to predict the correlations (Chen and Liu, 2004). Data mining techniques – such as decision trees, neural networks, association rules, Bayesian networks, and k-Nearest-Neighbors (kNN) classification - have become more common in recent years, and were shown to be helpful in many decision-making contexts, particularly in cases where when irregular and insightful patterns and correlations were difficult to detect manually due to a high dataset complexity (Fayyad et al., 1996; Han and Kamber, 2006; Wang & Wang, 2008).

While both OLAP and DM are considered advanced BI techniques, there is a significant distinction between them in terms of usage style, analysis capabilities, and potential outcomes and contributions. DM techniques has the advantage of being able to handle large and complex datasets without much time-consuming human intervention, and highlight valuable insights, which is difficult to find through manual data analysis. However, from the end-user's standpoint, DM can be seen as a "black box" – the end-user can define the input dataset and view the results but, else if he is an expert programmer, he has only limited control over the algorithm's operation. With this limitation, DM won't permit end-users to take advantage of their experience and in-depth understanding of the data, toward improving the analysis process and results. Another typical problem with DM applications is that the analysis results, even when being statistically significant, do not always make "business sense"; hence, they are often cannot be translated into decisions and actions (Wang and Wang, 2008).

On the other hand, in the OLAP environment the end-users are "steering the wheel", and therefore their skills and experience would have a crucial influence over the search process and the outcomes. Along with benefits – OLAP tools present some major challenges. With a large number of dimensions and quantitative facts - the OLAP tool creates an extremely large search space, in which there is a virtually infinite number of possible ways for slicing and displaying the data. This complexity might hinder end-users from taking full benefit of the OLAP tool's richness and flexibility. A phenomenon that might occur in the usage of OLAP tools is the "digging the same well" (Kolodner and Even, 2009) – an end-user, who is unfamiliar with the variety of search and presentation options, would repeat over and over the same search path, rather than attempting to explore other alternatives.

The solution that we explore in this study aims at maintaining the key advantage of OLAP tools – letting the end-users "steer the wheel", and taking advantage of their expertise in the business domain and their ability to interpret the data being analyzed in a business-oriented manner. However, the proposed integration of recommendations, may overcome some of the key issue in today OLAP tools by pointing out relevant data subsets, and helping the user reach and explore them faster. As further explained later, the recommendations are based on calculations, which ran in the back-end and attempt to detect valuable patterns automatically – an adoption and adaptation of a DM concept into OLAP.

2.3 OLAP Navigation

One of the main purposes of using OLAP software by the end-users is to get a better understanding of the data environment and to get an answer for business questions. The use of OLAP promotes self-exploration of datasets by the end-user. This exploration is supported by the navigation capabilities provided by the OLAP tool.The end-user navigates through the relevant data along certain *dimensions*. More specifically, by presenting *facts* values according to an aggregate function (such as sum, count, average etc), and slicing by chosen *dimension* attributes. Further, the user can decide to drill-down (zooming in to more detailed levels of hierarchies) by a selection of specific dimensions categories. Often, the objective of this action is to look for regions of anomalies. These anomalies may lead to identification of problem areas or new opportunities (Sarawagi et al., 1998).

A typical OLAP tool provide visualized utility that lets the user focus on a certain subset within a large dataset, visualize its data contents, and use them to perform analyses. A data subset is defined by a selection of a certain set of dimensions and facts, and applying a filter that limits the subset further. We define a single act of navigation as a choice of dimensions, facts and filters that moves the end-user from one subset configuration to another. We differentiate between two classes of navigation acts. The recommendation mechanisms that we introduce later reflect these two different classes.

- *"Inside the box"* navigation act – a choice of configuration that would move the user from a given subset, to a smaller subset within. OLAP professionals often term this act as "drilling down" – focusing the attention on a smaller data domain, which possibly has higher importance from the end-users viewpoint (e.g., "show business activities only from the last two quarters").
- *"Outside the box"* navigation act **-** a choice of configuration that would move the user from a given subset, to another subset, not within it. This navigation can be done, for example, by replacing one or more dimension in the initial set (e.g., "split the total sales by regions instead of by product categories"), or by applying a different filter on a given dimensions (e.g., "Show customer who live in the north, instead of those living in the south").

Navigation acts in OLAP tools have conceptual resemblance to feature selection in machine learning and data mining algorithms. Feature selection aims to reduce the dimensionality of patterns for classificatory analysis by selecting the most informative variables, and ignoring the irrelevant and/or redundant ones. Feature selection become an essential preliminary step to supervised machine learning problems, specifically for classification tasks that use datasets with a large number of variables (Guyon and Elisseeff, 2003). The presence of irrelevant or redundant variables might significantly hinder classification algorithm in terms of speed and prediction accuracy (Akadi et al., 2008).

The objective of feature selection is to find the smallest subset of variables that maximizes the pattern recognition ability. Ideally, this can be achieved by examining all possible subsets and finding the one that satisfies the above criterion. This approach is known as exhaustive feature selection. Even with a moderate number of variables, the exhaustive selection is impractical due to its demanding computational requirements. Some heuristics were developed to reduce computational complexity by compromising performances (e.g., Guyon and Elisseeff, 2003; Akadi et al., 2008) – some of which will be discussed later, as they can be applied within the context of this study.

3 The Mutual-Information Metric

Objects within a dataset may vary significantly in their informative and value contribution; hence different dimensions may contribute differently within a given decision context. A customary way to estimate the degree of informative of a specific dimension lies in the assessment of its ability to differentiate the data in relation to a predefine *fact*. This ability can be estimate by well known measures such as variance and inequality. The bigger the value of these measures, the higher is the dimension's ability to differentiate the data and create meaningful insights.

The objective of feature selection techniques can be conceptualized as mapping the most informative dimensions in the data set. A few methods can support such mapping: Pearson correlation, General linear models, decision trees, information theory metrics, etc (Guyon and Elisseeff, 2003; Han and Kamber, 2006). Our research will focus on information theory metrics as main criteria for ranking the dimension's informative ability. One of the main advantages of those metrics is that they do not make any preliminary assumptions on data distribution.

3.1 Definition of Information theoryconcepts

Information theory provides tools that can quantify the uncertainty of random quantities or the sharing of information between them (Thomas and Cover, 2006). We consider in this study only finite discrete random variables (e.g., X, Y and Z), where $p(x)$ and $p(y)$ are the probability densities of X and Y, and $p(x, y)$is the joint density. In cases where the densities $p(x)$, $p(y)$and $p(x, y)$ are unknown, as it most often happen, they can be estimated by frequency counts (Guyon and Elisseeff, 2003).

A fundamental concept in information theory is the entropy $\mathrm{H(Y)}$ of a random variable, which quantifies the uncertainty of Y. The entropy $\mathrm{H(Y)}$ of a discrete random *variable Y is defined by*

(1) $\mathrm{H(Y)} = -\sum_{\mathrm{y \in Y}} \mathrm{p(y)} \log_2 \mathrm{p(y)}$

The entropy is a functional of the distribution of Y. It does not depend on the actual values taken by Y, but only on their probabilities. Let X and Y be two random variables. If we can get knowledge on Y indirectly by knowing X, the resulting uncertainty on Y knowing X is given by its conditional entropy:

(2) $\mathrm{H(Y \mid X)} = -\sum_{\mathrm{x \in X}} \sum_{\mathrm{y \in Y}} \mathrm{p(x, y)} \log_2 \mathrm{(y \mid x)}$

The mutual information between X and Y measures the amount of knowledge on Y provided by X (or, vice versa, the amount of knowledge on X provided by Y). Therefore, it can be defined as:

(3) $\mathrm{I(X; Y)} = \mathrm{H(Y)} - \mathrm{H(Y \mid X)}$

This expression is the reduction of the uncertainty of Y when X is known. If Y is the dependent variable in a prediction context, the mutual information is thus particularly suited to measure the pertinence of X in a model for Y. In that case, we can rewrite the definition of mutual information as

(4) $\mathrm{I(X; Y)} = \sum_{\mathrm{x \in X}} \sum_{\mathrm{y \in Y}} \mathrm{p(x, y)} \log_2 \frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}}$

The relationship between the equations above is expressed in a Venn diagram (Figure 3). Notably, the mutual information $\mathrm{I(Y; X)}$ corresponds to the intersection of the information in X with the information in Y (Thomas and Cover, 2006).

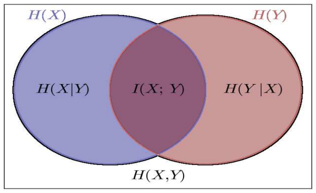


Figure 3. Relationship between entropy and mutual information

Another well known measure in information theory, which is derived from the measures above, is the conditional mutual information (CMI):

(5) $\mathrm{I(Y;\ X \mid Z)} = \mathrm{H(Y \mid Z)} - \mathrm{H(Y \mid X, Z)}$

This value is an estimate of the quantity of information shared between Y and X when Z is known. It can also be seen as the difference between the average remaining uncertainty of Y when Z is known and the same uncertainty when both Z and X are known (Fleuret, 2004).

3.2 OLAP Navigation – a Formal Definition

Going forward, we will use the flat dataset (Figure 4) as the baseline for our definitions, arguing that similar recommendation can also be developed for a multi-dimensional cube or a star-schema. We denote the *dimension* attributes in this dataset as $X_1, X_2, \ldots, X_n$ (perceived as independent variables) and consider one *fact attributed* denoted as Y (perceived as the dependant variable). Each of the *dimension* attributes has an associated value domain – e.g., a value domain of $\{x_{1,1}, x_{1,2}, \ldots, , x_{1,m}\}$ fordimension X_1.The values within those domains can be used by the end-user as *filter* categories. When navigating, the end-user can slice the data along specific *dimension* attributes and for getting a more focused view, s/he can *filter* those attributes by their value domains (an act which is similar to 'WHERE' operation in the SQL language).To present our algorithms further, we use as an example a dataset with 8 *dimension* attributes $X_1, X_2, \ldots, X_8$ which can be related to customer attributes such as gender, income level, education, etc. The dataset has one *fact*, Y, reflecting the customer's value.

	X_1	X_2	⋯	Y
Customer 1	$x_{1,2}$	⋯	⋯	⋯
Customer 2	⋯	⋯	⋯	⋯
Customer 3	⋯	$x_{2,2}$	⋯	⋯
⋯	⋯	⋯	⋯	⋯

Figure 4. A Sample Flat Dataset

To illustrate the OLAP navigation scheme presented above, and the connection to the information theory metrics, we use Figure 5 as reflecting a simple OLAP-navigation scenario. The left-hand side shows current data subset, generated by a filter operation. The end-user has selected to zoom in on category $x_{1,1}$ of attribute X_1 and, after that, has to decided whether to slice the data by X_2 or X_3. Estimating the mutual information for this two choices yields $\hat{\text{I}}(X_2; Y) = 1$ (the right most chart) and $\hat{\text{I}}(X_3; Y) = 0$ (the middle chart). The implication is that the user should clearly slice the subset by X_2, as it produced a much more informative view than X_3.

X1	X2	X3	⋯	Xn	Y
$x_{1,1}$	$x_{2,1}$	$x_{3,1}$			100
$x_{1,1}$	$x_{2,1}$	$x_{3,2}$			100
$x_{1,1}$	$x_{2,2}$	$x_{3,2}$			50
$x_{1,1}$	$x_{2,2}$	$x_{3,1}$			50

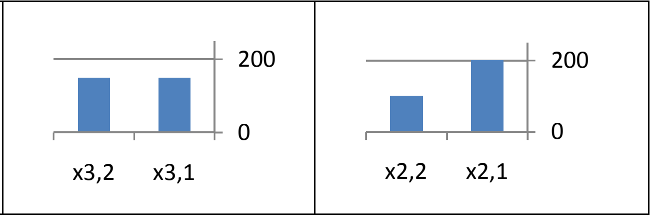


Figure 5. An OLAP Navigation example

The OLAP environment consists of dimensions and facts. Dimensions are usually discrete, while facts are often continues numeric attributes, associated with aggregation functions (sum, count, average, etc). The equations of information theory presented so far relate to discrete variables only. It is possible to calculate each of those measurements with continuous variables (by switching the sum to integral); however, this action is impractical due to its computational load (Guyon and Elisseeff, 2003). A common method to handle continues variables, is discretizing them or approximates their densities with a non-parametric method such as Parzen windows (Guyon and Elisseeff, 2003). For simplification, we suggest, as a preliminary step, to create another variable with finite set of categories that will include discrete values of the relevant fact. This process can be driven by business or statistical rationale, or by using an unsupervised algorithm such as k-means (Han and Kamber, 2006).

3.3 Inside-the-Box Recommendations with theMutual-Information Metric

Earlier we define “Inside the Box” recommendation as a navigation act that moves the user from a given subset, to a smaller subset within. In the algorithm for "Inside the box recommendations" presenting here, all the operations can be performed online, as the user uses the OLAP tool:

- Estimate the mutual information for each *dimension* attributes with the relevant fact

$$\hat{I}(X_1; Y), \hat{I}(X_2; Y), \cdots, \hat{I}(X_n; Y)$$

- Rank the *dimension* attributes by their mutual information value and present the ranking to the end-user (e.g., by color-coding as shown in Figure 6a).

$$\hat{I}\left(X_{v(1)}; Y\right) > \hat{I}\left(X_{v(2)}; Y\right) > \cdots > \hat{I}\left(X_{v(3)}; Y\right)$$

(a)

Dimension attributes	Mutual information
X8	0.1265
X7	0.0794
X5	0.0738
X2	0.0493
X3	0.0314
X6	0.0256
X4	0.0077
X1	0.004

(b)

Value domains	Element value
$X_{2,1}$	0.101
$X_{2,2}$	0.021
$X_{2,3}$	0.0045

Figure 6. Dimensions ranking (left) and value domains ranking (right).

- The ranking above makes a recommendation for the best next steps – choosing the *dimension* for which the mutual information is the highest (X_8in this example).
- Once the end-user has made a decision (has chosen a dimension for slicing), suppose X_i,the mechanisms will present him the degrees of contribution to criterion value of each value within the dimension's value domain:

  $$\hat{I}(X_i;Y) = \sum_{y \in Y} \hat{p}(x_{i,1},y) \log_2 \frac{\hat{p}(x_{i,1},y)}{\hat{p}(x_{i,1})\hat{p}(y)} + \cdots + \sum_{y \in Y} \hat{p}(x_{i,n},y) \log_2 \frac{\hat{p}(x_{i,n},y)}{\hat{p}(x_{i,n})\hat{p}(y)}$$

  so that:

  $$\underset{(j \mid x_{i,j} \in X_i)}{\operatorname{argmax}} \sum_{y \in Y} \hat{p}(x_{i,j},y) \log_2 \frac{\hat{p}(x_{i,j},y)}{\hat{p}(x_{i,j})\hat{p}(y)}$$

- As shown in Figure 6b, the end-user will now get a recommendation for the most informative value within the domain. Accordingly, the end-user should filter out the non-informative (or the low rated) values – an act of move deeper into a more internal dataset.

Considering the concept of a recommender system, the end-user will not be forced to follow the recommendations – but can rather perform any valid navigation act. Figure 7 show the difference between the outcome of a navigation act that follow the recommendations and one that doesn't.

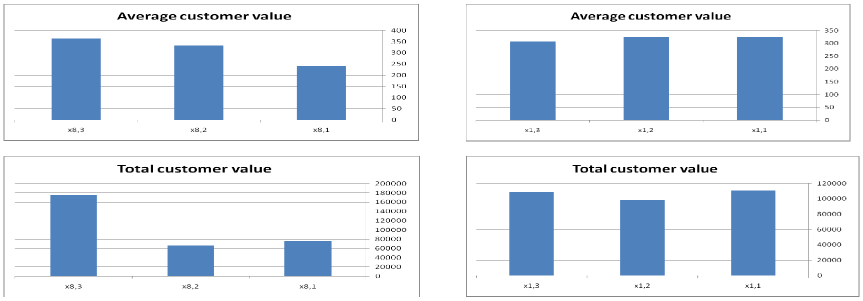


Figure 7. High rated slicing of the data (left) vs. Low rated (right).

3.4 Outside-the-Box Recommendations with the Mutual-Information Metric

One of the main limitations of filters type method for feature selection in generally and of the algorithm presented above particularly it's his greediness. This method doesn't take into account dependencies between features and thus, can cause to loss of information or can be unaware to very informative slicing that could be creating by an integration of two or more *dimension* attributes. Two attributes can be highly relevant together while each of them appears to be poorly relevant once taken individually. As a consequence, these *dimension* attributes could be badly ranked (Akadi et al., 2008). Given this, we now enhance the algorithm by adding "outside the box" recommendations.

Earlier we define "Outside the Box" recommendation as a navigation act that would move the user from a given subset, to another subset, not within it. The main goal of the outside-the-box mechanism is to present the end-user with subsets of *dimension* attributes that are more informative than the subsets currently used. This method required offline calculations due to a costly computation. Given a data set which contain N *dimension* attributes, we suggest ranking subsets of K attributes ($\mathrm{K} \ll \mathrm{N}$) based on their conditional mutual information with the relevant fact (here, we have chosen $\mathrm{K} = 3$).

In order to achieve a subset of attributes that carries as much information as possible, the ultimate goal would be to choose $v(1), \ldots, v(K)$ which maximize $\hat{I}(X_{v(1)}, \ldots, X_{v(K)}; Y)$(Cover and Thomas, 2006; Fleuret, 2004). By estimating this value per each and every subset with K attributes, we can provide the end-user a scale of degree of information. This value is relative to the current *dimension* attributes in use (Figure 8). This mechanism required an offline calculation that will produce a matrix of mutual information with all the possible combinations for K-attribute subsets.

Attributes combination	Information degree
X_8, X_6, X_3	2.8
X_7, X_4, X_2	2.4
…	…
X_5, X_8, X_3	1.7
…	…

Figure 8. *"Outside-the-Box" Recommendations*

The value of $\hat{I}(X_{v(1)}, \dots, X_{v(K)}; Y)$ is hard to estimated with real-world datasets, due to the high computational cost. Fleuret (2004) suggests taking into account the tradeoff between individual power and independence and comparing each new attribute with the ones already picked. He suggest that attribute X' is good only if $\hat{I}(Y\,; X'|X)$ is large for every X already picked. This means that X' is good only if it carries information about Y, and if this information has not been caught by any of the X already picks. More formally, he purposes the following iterative scheme:

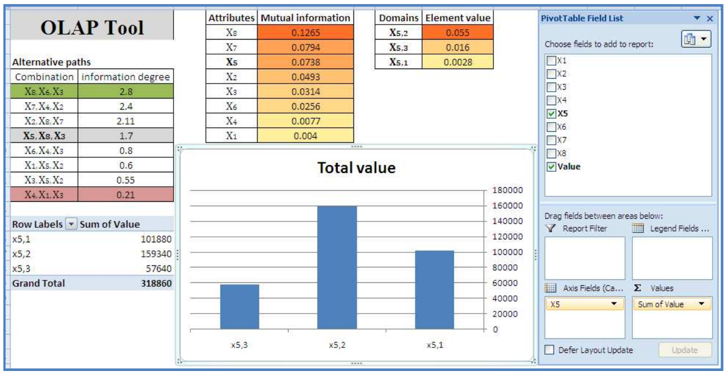
(6) $v(1) = \operatorname{argmax}_n \hat{\mathrm{I}}(\mathrm{Y}\,;\,\mathrm{X_n})$

(7) $\forall k,\ \ 1 \le k < K,\ \ v(k+1) = \operatorname{argmax}_n\{\min_{l \le k} \hat{\mathrm{I}}\left(\mathrm{Y}\,;\,\mathrm{X_n} \mid \mathrm{X}_{v(l)}\right)\}$

In the above method the goal is to find the best subset of K features (one subset only). We suggest using this method to create maximum N subsets of *dimension* attributes, where N is their number in the data set. That's can be made by selecting each time another attribute from the data set as the first one in the subset. Accordingly, we'll get maximum N subsets that each ranked by the total contribution of its features to the reduction of the *fact* uncertainty. The ranking function will be sum of the mutual information of the first attribute in the subset and the conditional mutual information of the rest of them according to Eq. 7.

3.5 Integrating recommendationsinto OLAP tools

Figure 9 illustratesan OLAP tool (based on Excel's Pivot-Table) enhanced with a recommendation mechanism. The tables above the chart present the recommendations given to the end-user regarding highly-ranked next moves, relative to the selected data cube. The table, at the left hand side, present recommendations of attributes for the next preferred navigation act. Once to end-user has chooses an attribute, the table at the right hand side presents value domains that are possible better for the next filter operation. Those recommendations refer to the inside-the-box algorithm.



OLAP Tool

Alternative paths

Combination	information degree
X_8, X_6, X_3	2.8
X_7, X_4, X_2	2.4
X_2, X_8, X_7	2.11
X_5, X_8, X_3	1.7
X_6, X_4, X_3	0.8
X_1, X_5, X_2	0.6
X_3, X_5, X_2	0.55
X_4, X_1, X_3	0.21

Attributes	Mutual information
X_8	0.1265
X_7	0.0794
X_5	0.0738
X_2	0.0493
X_3	0.0314
X_6	0.0256
X_4	0.0077
X_1	0.004

Domains	Element value
$X_{5,2}$	0.055
$X_{5,3}$	0.016
$X_{5,1}$	0.0028

Row Labels	Sum of Value
x5,1	101880
x5,2	159340
x5,3	57640
Grand Total	318860

PivotTable Field List

Choose fields to add to report:

☐ X1
☐ X2
☐ X3
☐ X4
☑ X5
☐ X6
☐ X7
☐ X8
☑ Value

Drag fields between areas below:

Report Filter | Legend Fields ...

Axis Fields (Ca... : X5

Values: Sum of Value

☐ Defer Layout Update

Figure 9. *OLAP tool with recommender system based on Microsoft Excel 2007*

In addition – the user is also provided with outside-the-box recommendations, presented in the top-left side of the panel. The best dimension combination appears at the top of the list, while the current configuration is highlighted in grey for comparison.

4 Conclusion

This study presented a quantitative method, based on the mutual-information metric, for assessing the relative importance of different data subsets within a large dataset, and generating recommendation for the end-user regarding the next steps to take, when navigating over a large dataset with an OLAP tool. The solution aims at improving two important information-quality aspects in BI tools – *accessibility*, conceptualized as the ability to reach the relevant data subsets within a reasonable time, and *understandability*, conceptualized as the ability to comprehend the relevant data subsets and gain important business insights from analyzing them.

The concept of generating recommendations with OLAP, and BI applications in general, is novel and requires further investigation. The technique presented here is one among many possible for assessing whether a certain data subset is more interesting than another – and others should be explored as well (e.g., metrics such as the variance, or as the Gini coefficient which reflects the extent of inequality). Beyond the analytical development – some empirical assessment is required as well, toward assessing whether the extension proposed here will indeed improve the end-user's decision-making capabilities. This study is currently in a phase of designing a lab experiment that will test the contribution of recommendation based on the mutual information metric in a simulated setting. Finally, there is still a need to rethink the design of OLAP tools that will be able to make valuable recommendations, without reducing the tool's usability and visual effectiveness. Figure 9 only illustrates a possible design that integrates some recommendation – however, turning the recommendations into an applicable tool, is likely to require some significant design efforts.

References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Tr. on Knowledge and Data Eng., 17 (6), 734-749.

Akadi A.E., Ouardighi A.E. and Aboutajdine D. (2008).A Powerful Feature Selection approach based on Mutual Information.Intl. Journal of Computer Science and Network Security, 8 (4), 116-121.

Cover M.C. and Thomas J.A. (2006).Elements of Information Theory.John Wiley & Sons Inc.

Davenport, T.H. (2006). Competing on Analytics. Harvard Business Review. 84 (11), 99-107.

Fayyad U., Piatesky-Shapiro G. and Smyth P. (1996).From Data Mining to Knowledge Discovery in Databases.Communications of the ACM. 39 (11), 37-54.

Fleuret F. (2004).Fast Binary Feature Selection with Conditional Mutual Information.Journal of Machine Learning Research 5, 1531-1555.

Guyon I. andElisseeff A. (2003).An Introduction to Variables and Feature Selection. Journal of Machine Learning Research 3, 1157-1182.

Han J. andKamber M. (2006). Data Mining: Concepts and Techniques.Morgan Kaufmann publishers.

Kolodner Y. and Even A. (2009).Integrating Value-driven Feedback and Recommendation mechanisms into Business Intelligence systems.In Proceedings of the 17th European Conference on Information Systems

Malinowski J., Weitzel T. andKeim T. (2008). Decision support for team staffing: An automated relational recommendation approach. Decision Support Systems. 45 (3), 429-447.

March, S. T., and Hevner A. R. (2007).Integrated Decision Support Systems: A Data Warehousing Perspective. Decision Support Systems. 43 (3), 1031-1043.

Pipino L.L., Lee Y.W. and Wang R.Y. (2002).Data quality assessment. Comm. of the ACM.45 (4), 211-218.

Sarawagi S., Agrawal R., Meggido N. (1998). Discovery Driven Exploration of OLAP Data Cubes.EDBT Conference Proceedings.

Tremblay M.C., Fuller R., Berdt D. and Studnicki J. (2007). Doing more with more information: Changing healthcare planning with OLAP tools. Decision Support Systems. 43 (4), 1305-1320.

Wang H. and Wang S. (2008). A knowledge management approach to data mining process for business intelligence.Industrial Management & Data Systems. 108 (5), 622-634.

Wang R.Y., Dai W. and Yuan Y. (2008). Website browsing aid: A navigation graph-based recommendation system.Decision support systems. 45 (3), 387-400.

Watson H.J. and Wixom B.H. (2007).The Current State of BI.*Computer.* 40 (9), 96–99

Wixom B.H., Watson H.J. and Hoffer, J.A. (2008). Continental Airlines Continues to Soar with Business Intelligence. Information Systems Management. 25 (2), 102-112.