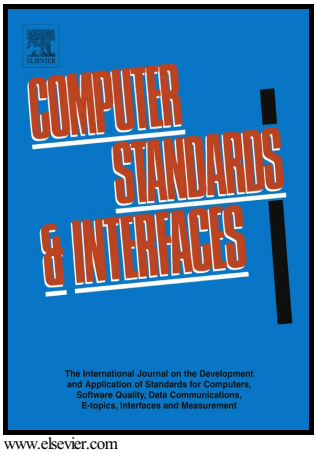
Application of Data Mining techniques to identify relevant Key Performance Indicators

Jesús Peral, Alejandro Maté, Manuel Marco



www.elsevier.com

PII: S0920-5489(16)30092-7
DOI: http://dx.doi.org/10.1016/j.csi.2016.09.009
Reference: CSI3141

To appear in: *Computer Standards & Interfaces*

Received date: 11 April 2016
Revised date: 29 August 2016
Accepted date: 21 September 2016

Cite this article as: Jesús Peral, Alejandro Maté and Manuel Marco, Applicatio of Data Mining techniques to identify relevant Key Performance Indicators *Computer Standards & Interfaces,* http://dx.doi.org/10.1016/j.csi.2016.09.009

Application of Data Mining techniques to identify relevant Key Performance Indicators

Jesús Peral[a], Alejandro Maté[a], Manuel Marco[a]

[a]*Department of Software and Computing Systems, University of Alicante, Spain*

Abstract

Currently dashboards are the preferred tool across organizations to monitor business performance. Dashboards are often composed of different data visualization techniques, amongst which are Key Performance Indicators (KPIs) which play a crucial role in quickly providing accurate information by comparing current performance against a target required to fulfil business objectives. However, KPIs are not always well known and sometimes it is difficult to find an appropriate KPI to associate with each business objective. In addition, data mining techniques are often used when forecasting trends and visualizing data correlations. In this paper we present a new approach to combining these two aspects in order to drive data mining techniques to obtain specific KPIs for business objectives in a semi-automated way. The main benefit of our approach is that organizations do not need to rely on existing KPI lists or test KPIs over a cycle as they can analyze their behavior using existing data. In order to show the applicability of our approach, we apply our proposal to the fields of Massive Open Online Courses (MOOCs) and Open Data extracted from the University of Alicante in order to identify the KPIs.

Keywords: KPIs, Data mining, big data, decision trees, Open Data

1. Introduction

Dashboards and Scorecards [15] allow decision makers to quickly assess the performance of an organization by visualizing aggregated data using different kinds of visualizations. This capability makes Dashboards the preferred tool across organizations for monitoring business performance. From among the different visualizations included within Dashboards, Key Performance Indicators (KPIs) [22] play a crucial role since they provide quick and precise information by comparing current performance against a target required to fulfil business objectives.

However, KPIs are not always well known and sometimes it is difficult to find an appropriate KPI to associate with each business objective [2]. In these cases,

it is common to resort to existing lists of KPIs, such as APQC[1], in order to test candidates over short periods of time until a suitable one is found. However, what happens when an organization adopts an innovative activity or explores a new data source such as the social media? The absence of lists of KPIs forces managers to rely on their intuition in order to select potential candidate KPIs. This has several undesirable consequences. First, some KPIs may be redundant [24], misdirecting the efforts and resources of the organization. Second, the people responsible for the (wrong) KPIs develop a resistance to change once they have found out how to maximize their value [22]. Third, there is a tendency to focus on the results themselves [22, 2] (e.g. Sales) rather than on the actual indicators that can be used (e.g. Successful deliveries/Total deliveries) and that lead to the results obtained.

Therefore, there is currently a need for techniques and methods that improve the KPI elicitation process, providing decision makers with information about the relationships between KPIs and their characteristics. This information can be highly valuable in KPI selection, not only for traditional datasets but also for Big Data where the implications for the company of the data are unknown and, thus, eliciting their relationships with internal KPIs can make these data actionable, adding value to them.

Big Data involves huge volume, complex, and growing data sets with multiple and heterogeneous sources. With the rapid development of networking, data storage, and data collection capacity, Big Data are now rapidly expanding in all domains and scenarios. In [25]a HACE theorem was presented that characterizes the features of the Big Data revolution and proposes a Big Data processing model from the data mining perspective. The authors analyze the challenging issues in the data-driven model and also in the Big Data revolution.

The principle of "What You See Is What You Get" is followed in many human-computer interaction scenarios. Only when the analytical results are displayed in a user-friendly way are they effectively utilized by users. Reports, histograms, pie charts, regression curves, etc. are frequently used to visualize the results of data analysis. This leads to the topic of visualization being seen as one of the main challenges in mining big data [6].

In this paper we present a new approach to combining these two aspects in order to drive Data Mining (DM) techniques to obtain specific KPIs for business objectives in a semi-automated way. The main benefit of our approach is that organizations do not need to rely on existing KPI lists, such as APQC, or test KPIs over a cycle, as they can analyze their behavior using existing data. In order to show the applicability of our approach we apply our proposal to the new field of MOOCs (Massive Open Online Courses) and Open Data extracted from the University of Alicante in order to identify the relevant KPIs.

The remainder of this paper is structured as follows. Section 2 discusses the related literature. Section 3 describes our proposal for KPI elicitation. Sections

[1]American Productivity and Quality Center, http://www.apqc.org/ (visited on 6th of April, 2016).

4 and 5 present our case study, based on a MOOC being run at the University of Alicante. Finally, Section 6 draws the conclusions and discusses future areas of work.

2. Related Work

In [15] the authors propose the Balanced Scorecard, a tool that consists of a balanced list of KPIs associated with objectives covering different business areas. The usefulness of the Balanced Scorecard has led to its rapid adoption by companies all around the world. However, while the structure of the Balanced Scorecard is clear, its content is not. Given that many companies struggle to succeed with their KPIs, in [16] the authors propose the use of Strategy Maps. Strategy Maps describe the way in which the organization intends to achieve its objectives, by capturing the relationships between them in an informal way. Recently, the concepts included within Scorecards and Strategy Maps have been further formalized into business strategy models. Business strategy models [14] bring KPIs, objectives, and their relationships together in a single formal view. Despite these efforts, it is still unclear whether the KPIs included in these objective models are adequate, or even if the relationships between objectives perceived by decision makers are indeed reflected by the KPIs selected to measure their degree of attainment.

Therefore, in [22], the author focuses on the design and implementation of KPIs within Dashboards. The author differentiates between Key Result Indicators (KRIs) and KPIs in order to differentiate between results and actual performance and highlight the importance of relationships between indicators, and also discusses the characteristics and target public that each KPI should have. However, there is no discussion about how KPIs or their relationships could be elicited from data. In addition, in [24], the authors propose the QRPMS method to select KPIs and elicit relationships between them. The method starts from a pre-existing set of candidate KPIs and performs a series of analytical steps using data mining techniques, such as Principal Component Analysis (PCA), alternated with human intervention in order to identify potential KPI relationships and help decision makers select those KPIs that seem most relevant for the business.

Big Data is a relatively new but already commonly used term used to identify datasets that we cannot manage with current methodologies or data mining software tools, principally due to their size and complexity. Big Data mining is the capability to extract useful information from these large datasets or streams of data. New mining techniques are necessary due to the volume, variability, and velocity of such data. The Big Data challenge is becoming one of the most exciting opportunities for the years to come. The work of Fan and Bifet [9] is a good reference point as it offers a broad overview of the topic, its current status, controversial aspects, and a vision of the future. They introduce four articles, written by influential scientists in the field, covering the most interesting and state-of-the-art topics on Big Data mining. There are many tools for big data mining and analysis, including professional and amateur software, expensive

commercial software, and open source software. In [6] there is a brief review of the top five most widely used pieces of software, according to the survey "What Analytics, Data mining, Big Data software that you used in the past 12 months for a real project?" received from 798 professionals by KDNuggets in 2012.

There is no clear consensus on what Big Data is. In fact, there have been many controversial statements about Big Data, such as "Size is the only thing that matters." In [17] the authors try to explore the controversies and debunk the myths surrounding Big Data.

In [26] there is a discussion of the major differences between statistics and data mining before moving on to look at the uniqueness of data mining in the biomedical and healthcare fields. It gives a brief summary of various data mining algorithms used for classification, clustering, and association as well as their respective advantages and drawbacks.

In short, there are a number of works focused on monitoring performance using KPIs, but most of the works that tackle the problem of KPI selection require a pre-existing set of KPIs. Obtaining this set of KPIs can be a difficult task in already established organizations [2], and becomes a challenge when the business activity is developed in an innovative environment.

3. Methodology proposed

In this paper we propose a new methodology for extracting the relevant KPIs from the business strategy model of a particular enterprise/activity. The six steps comprising the methodology are shown in Figure 1. Below we discuss the different stages:

Stages 1 and 2: Definition of the business strategy model. First of all, we start by focusing on modeling the business strategy and known KPIs (if any) to guide the process. In many organizations, some of these goals and KPIs are already listed in the Balanced Scorecard [15]. However, the business strategy model can offer us more information. Specifically, it includes the relationships between the different business objectives to be achieved and (optionally) the processes that support them. If a more thorough analysis is required, one can consider including a SWOT analysis (Strengths, Weaknesses, Opportunities and Threats) [13] within the business strategy. This analysis identifies those elements, external and internal to the company, that affect one or more of the objectives established as priorities. Therefore, SWOT analysis allows us to quickly identify the possible reasons for deviations in the indicators and then make decisions accordingly. A result similar to the concept of Strategy Maps [16] will be obtained on completing this first step. Once we have modeled the business strategy view, we list and prioritize those objectives that do not have any associated KPIs so that we can measure these. Each of these objectives is related to one or more business processes that support them.

For example, our first step in tackling the analytical challenges of UniMOOC at the University of Alicante was to carry out several interviews

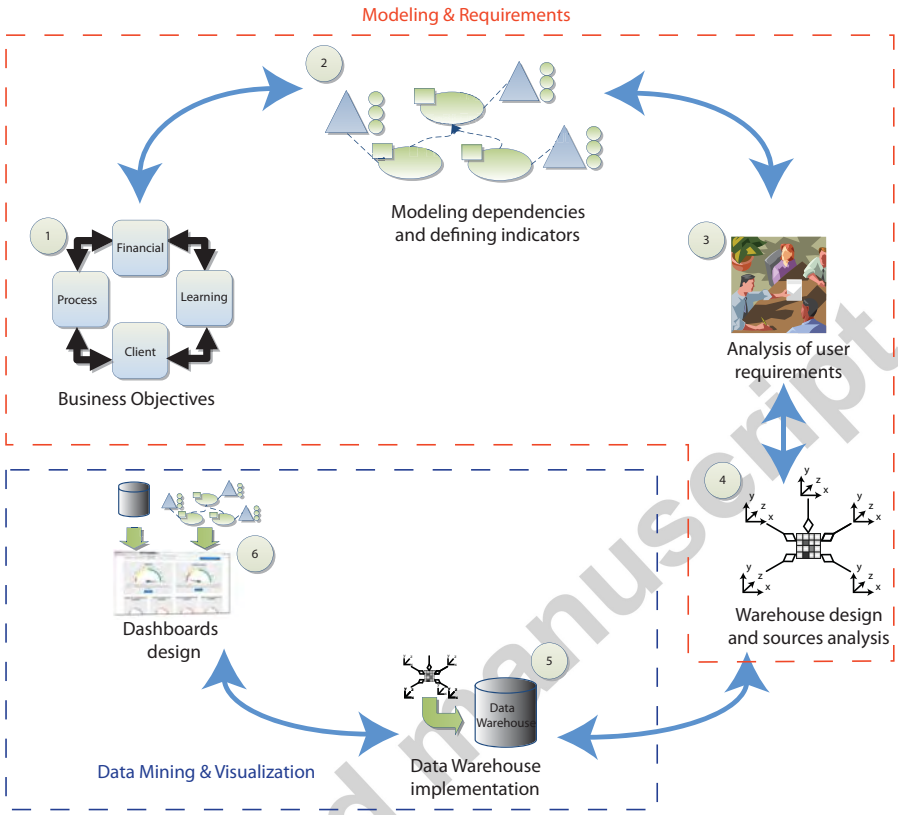


Figure 1: Generic strategic map. In stages 5 and 6 data mining is applied to provide visualization and KPIs.

with the course organizers. This provided us with some abstract and high level information to about the goals and objectives of the course managers, allowing us to produce a first set of indicators and create an initial version of the multidimensional model for analysis.

Stages 3 and 4: Definition of KPIs and multidimensional model. In this phase, and using the indicators obtained in the previous stage, we create a multidimensional model to support their calculation and provide additional analytical capabilities. The model allows the mapping from the indicators to DW elements, making it possible to generate the DW schema automatically. Our multidimensional model is composed of two analysis cubes: Enrollment and Activity. The first one, Enrollment, allows us to analyze whether the characteristics of the students, such as country, interests and expectations, present certain patterns.

For example, in the case of UniMOOC the "Increase the number of Students" objective is related to the "Enrollment" business process. When-

ever possible, candidate KPIs for each objective should be extracted from the business processes that support them, as it is the running of these daily processes that leads to the success or failure of the company.

Each of the business processes listed has one or more decision makers responsible for analyzing the information produced in its daily activity. By interviewing these decision makers we can create new user requirement views or review existing ones that have already been specified for implementing the companys data warehouse. The aim of this step is to associate business objectives with entities and measures that are related to their performance. In this way we move from abstract objectives to pieces of information that we can combine in order to propose KPIs to measure the performance of the organization.

Using the entities and measures identified during the requirement analysis, we work with stakeholders to elaborate a set of candidate KPIs for each objective listed during the first step. To ensure the usefulness of the defined KPI, each of these should follow the SMART criteria [19], being Specific, Measurable, Achievable, Relevant, and Timely available. If a target does not have an associated indicator or does not follow the SMART rules, then it cannot be properly tracked. Therefore, there is a risk that the target will deviate from its expected evolution and that this will go unnoticed until eventually it fails without having had the option to take corrective action.

Stages 5 and 6: Use of DM to extract relevant KPIs. The following step is to analyze the candidate KPIs through data mining techniques to ensure that they reflect the relationships identified during the business strategy modeling. The main objective of these stages is to check the relationships/correlations between the KPIs to determine real and relevant KPIs.

The method consists of five processes:

1. Preprocessing. We first preprocess the data. During preprocessing we determine the availability and characteristics of the data, including the existence of missing values and flat data within the time series. Indicators for which data is unavailable, largely missing, or flat (i.e. their values do not change at all during the period of time being analyzed) are marked and discarded from further analysis since we cannot derive any information from them. Furthermore, during preprocessing we determine whether we are working with pure time series or panel data[2].
2. Detection of potential anomalies. Second, we perform a basic analysis of indicators in order to detect potential anomalies. This analysis

[2]Panel data refers to data that is two-dimensional, most often containing time and another dimension, such as geographical or product information.

includes basic statistics, including maximum, minimum, and standard deviations. If we are working with panel data, measures that present very large deviations can be addressed in two ways:

- One possible solution is to separate each instance of the data into its own time series and perform independent analyses if enough data is available[3] for each instance. This will lead to us analyzing operational KPIs (by product or by region), rather than strategic level KPIs. The advantage of this method is that we will be able to identify diverging behaviors across instances of the data. However, once the analysis has been conducted, we must extract the factors and relationships common to all regions in order to provide a strategic view and update the strategic model.
- Alternatively, we can normalize the data and add the new normalized measures into the analysis. This allows us to focus on the relative behavior of the variables without the size of instances affecting the analysis. However, we must take into account that by following this approach we are weakening the correlation effect between variables which may make it harder to identify potential relationships. Furthermore, we are assuming that all instances will behave in a similar way, leading to a confirmation of the correlation which may not be the case.

3. The calculation of difference series. Third, we can calculate difference series. Difference values are calculated by subtracting the first term from the second. Their purpose is two-fold. First, they allow us to calculate the trend for the indicators and potential thresholds in the specific case of measures. Second, they allow us to calculate sentinel-like rules, as specified in [20].
4. Analysis of pair-wise relationships. Using all the pre-calculated data, we proceed as follows. First, we analyze pair-wise relationships between series using correlation, time series analysis, and linear regression. Then, we analyze the existence of compound relationships by considering multiple indicators at the same time and applying multiple DM techniques.
   For the pair-wise analysis, if enough data is available for the time series analysis we start by analyzing the correlation between indicators in order to obtain a pair-wise list of candidate relationships within the data. Next, on the basis of this list, we calculate the cross-correlation coefficients between the indicators of each candidate relationship. Cross-correlation coefficients give us the time difference that leads to obtaining the best correlation coefficient for each relationship identified. Finally, using this information we calculate

[3] As a rule of thumb, at least 30 to 50 points is recommended for statistical models. However, this value actually depends on the model to be built. For the interested reader, we refer to [11] for a more in-depth discussion.

ARIMA models [4] to estimate the confidence of the relationships identified by measuring the predictive power of the variables.
Alternatively, if there is not enough data for the time series analysis, it will not be feasible to calculate complex regression models and cross-correlation. Therefore, we resort to simpler analysis techniques which are more robust for small data samples. In this case, we first calculate the correlation values between indicators to obtain a pair-wise list of candidate relationships. On the basis of this list, we calculate linear regression models for each relationship in order to compare the behavior across data instances and the predictive power of the variables. Finally, we calculate the sentinel-like relationships using the difference series calculated in the previous step to evaluate the confidence in the relationship. These relationships are calculated by counting the number of times a positive (negative) difference in the first indicator has a direct effect on the second indicator of the relationship, and subtracting the number of times that the opposite happens.

5. Analysis of compound relationships. Pair-wise analysis allows us to detect strong relationships between indicators. This is useful for identifying both potential cause-effect relationships as well as the existence of unknown redundant indicators. In order to capture more complex relationships we make use of DM techniques, especially classification ones, which can represent non-linear functions between the components under study. For this task we first align the data according to the results of the cross-correlation in the previous step (if available) and the domain knowledge available. For example, in the case of marketing campaigns there is a delay between the launch of the marketing campaign and the moment that it starts having an effect, so the data would have to be aligned in order to account for this time lag. Once the data has been aligned, we start training classifiers with the set of relationships between objectives and their associated indicators captured in the strategic model. Progressively, we add more indicators to the model and remove those that reduce the predictive power of the model (add noise). When the set of candidate indicators is able to predict the value of the target to a certain threshold of confidence, defined on a case-by-case basis, then this set is selected for inclusion in the next update of the analysis views.

Finally, we define or update the analysis views for different roles, embodied in dashboards that will allow decision makers to access and monitor the new KPIs.

4. Previous Case Study and Limitations

In recent years, the effect of globalization along with the proliferation of open online courses has radically changed the traditional education sector. While new

technologies offer many opportunities, there are significant challenges that must be overcome to take full advantages of these [1].

More recently, a new kind of online course has appeared: MOOCs. A MOOC is an online course with the objective of interacting and promoting participation and open access via the web. In addition to traditional resources such as slides, MOOCs provide video lectures, both off-line and on-line, and user forums that help to build an expert/professional community for the students and instructors. These advantages have seen MOOCs quickly gaining in popularity, and thus they have been increasing their number of students exponentially over the last few years.

Following the methodology proposed in section 3, we can identify the following stages:

Stages 1 and 2. In particular, we present the process followed to elicit and model the critical information from the MOOC named UniMOOC[4], as well as the results of this procedure. UniMOOC is a MOOC that currently has over 20,000 unique students registered and focuses on entrepreneurship. The course includes several units and modules as well as links to social networks where students can exchange opinions. Some of the UniMOOC course objectives are defined in Figure 2 following the methodology proposed in the previous section.

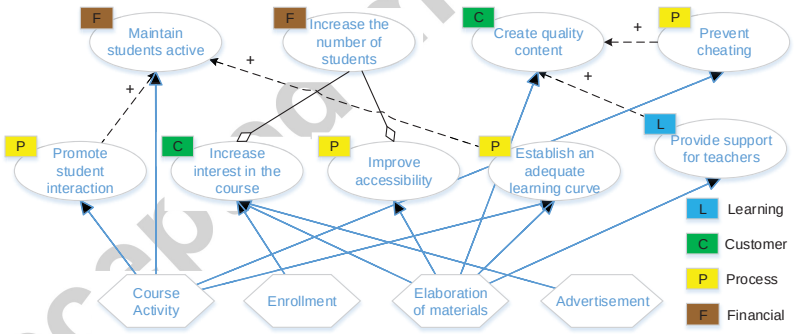


Figure 2: Diagram for the MOOC course objectives.

First, UniMOOC aims, as do many other MOOCs, to have the largest possible number of active students (top-left objective). MOOCs suffer from a relatively large number of students who just come to check the course content but do not intend to follow it. Therefore, making the course engaging to students is key to its survival and popularity. One

[4]UniMOOC can be accessed at http://unimooc.com/landing/ (visited on 6th of April, 2016).

way of making MOOCs more engaging is to promote student interaction. There are multiple ways of fostering communication between students, from simple message boards included in some activities that foster the exchange of ideas, to fully-fledged social networks dedicated to one or more MOOCs. All of these alternatives must, however, be considered within the scope of the course activity in order to avoid generating conversations that are unrelated to the course itself. Additionally, in order to keep students active a course must not only be engaging, but also appropriate. The lack of direct professor-student interaction makes it difficult to pinpoint problems in the lessons, where there are sudden jumps in the learning curve that make students who are struggling with the course give up and abandon it.

Second, the course has the objective of increasing the number of students taking it (top center-left objective). Despite the fact that a large proportion of the students may not see out the entire course, it is necessary for new students to join in order for the course to continue to have a meaning. New students will join a course if their attention is captured by the benefits of following it. UniMOOC does this in several ways, by means of advertising, having high quality materials, emphasizing the skills it provides and positions that have been achieved by those who have passed it, and by identifying the related companies and renowned experts that collaborate in giving the course. Additionally, another way of increasing the number of new students is by improving the accessibility of the course. In many cases, students who wish to enroll on a course do not finish the process when the entrance barriers (data to be provided, excessive requirements, pre-tests, etc.) are too time consuming for them, abandoning the course before they even start.

Third, UniMOOC aims to create high quality content. Creating high quality content is positive for MOOCs not only because it makes the course more satisfying for the students and professors, but also because it benefits other courses within the same platform, as it creates a positive reputation for those in charge of the course and for any other course within the same group. Furthermore, high quality content is easier to maintain. Generating high quality content can be done by providing teachers with enough support to develop this, ensuring not only that they are given the tools to produce the material, but also that they are provided with the knowledge and skills to use the platform, and with assistance whenever they struggle in their interaction with the technology. Finally, but most importantly, a high quality course ensures that certifications and badges provided by the course are meaningful for students. In this sense, one of the key aspects in achieving high quality content is to prevent cheating.

Preventing cheating is a main objective in itself. Given the vast number of students enrolled in MOOCs, it can be challenging to detect and keep track of cheaters. However, due to its multiple negative effects on the course, this is an objective that needs to be focused on. At its most basic

level it nullifies the whole point of the course: transferring knowledge and skills to students. In traditional courses, cheating can be dealt with by means of in-situ exams, where it is hard to cheat. In some MOOCs however, badges and certifications are given for following the course and paying a fee. If students are able to cheat on that course, those badges and certifications will have no validity at all, thus the importance of putting in place mechanisms to prevent cheating.

Stages 3 and 4. In our case study, the indicators obtained in a generic way, which may be applicable to other online courses, are:

- increment in the number of students,
- dropout ratio,
- student recovery ratio,
- % of active students,
- % of students who fail the course,
- % of students passing the exams without seeing the corresponding lessons,
- % of students taking the course in a continuous or sequential pattern.

The multidimensional model was created by using the conceptual modeling proposal described in [18], where the information is organized according to Facts (center of analysis) and Dimensions (context of analysis), as shown in Figure 3.

Stages 5 and 6. We started by applying the classical data mining techniques to the course database. However, due to the large amount of data on this course, these techniques are not very suitable because they are difficult to interpret. For instance, they produce a lot of rules in association rules and decision trees; they also produce many hidden neural connections in the artificial neural networks, etc. The best way to analyze these data is by using visualization methods. In addition, visualization techniques allow us to see how they graphically grow dynamically. In particular, we use Google Analytics (GA) since it offers a free tool for measuring and analyzing several useful statistics [21] [23] [7] [10].

Figure 4 shows the number of sessions per user throughout the course. We can identify different days that are higher. In addition, it also shows the users and new users per month. This is important to identify the days where there is more traffic and to make a more efficient use of the resources. Figure 5 shows the users per country; this is very useful when promoting these courses in the areas where there are no students.

There are many other parameters to measure such as gender, age, type of navigator, expectations, interests, etc. about the students in order to improve the courses, especially in future editions such as Figure 6 and Figure 7.

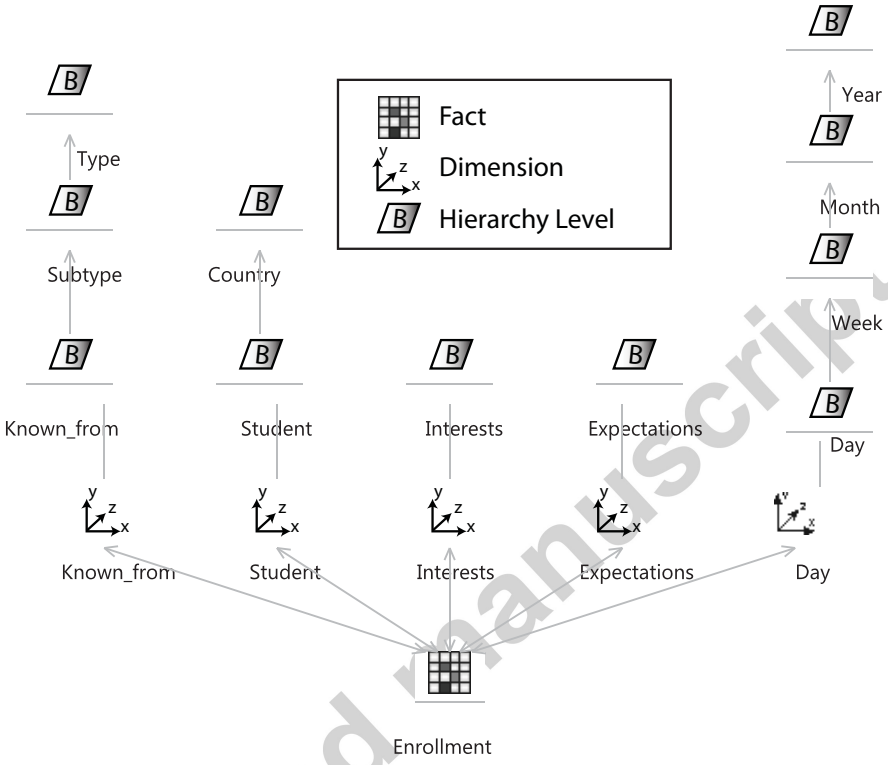


Figure 3: MOOC Multidimensional modeling for the enrollment analysis.

However, one limitation for MOOCs is that they have only recently started to be run and, as such, platforms, practices, and courses offered are not yet stable. We ourselves had to face this problem as the underlying technical platform has been completely changed, rendering all the previous data gathered inaccessible and requiring a complete re-engineering of the statistical analysis. Our initial idea was to build upon the previous analysis to provide more in-depth insights. However, the constant changes to the platform technology makes this idea unfeasible. Thus, we opted to analyze another interesting case also within the education sector, Open Data provided by universities.

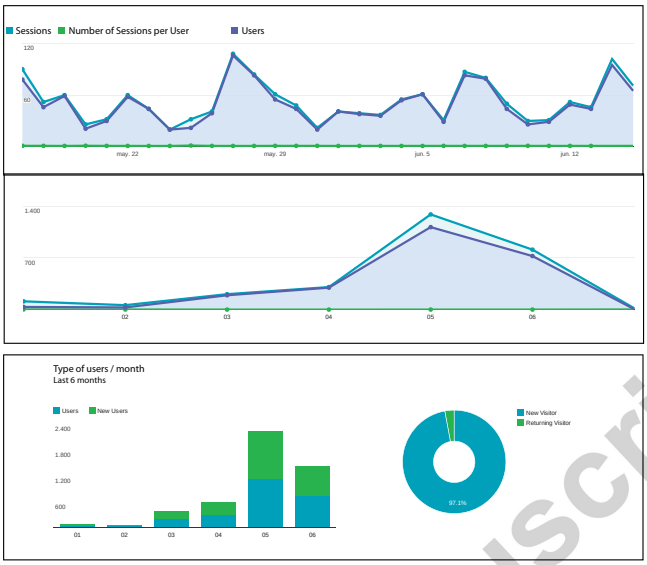


Figure 4: Visualizations of sessions and type of user.

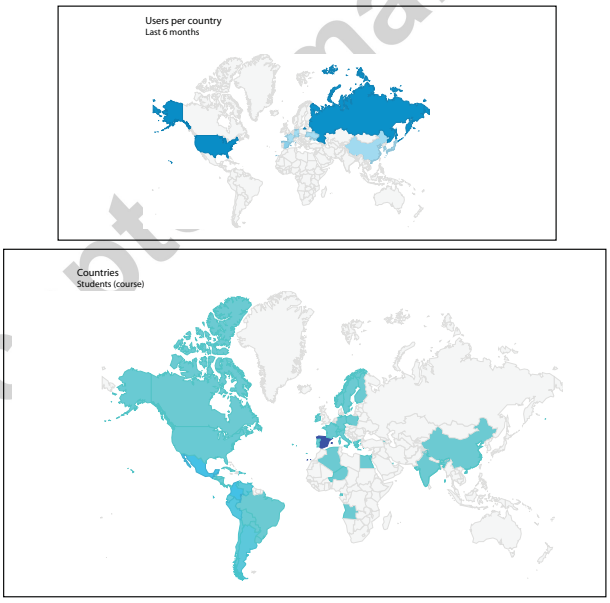


Figure 5: Visualizations of users per country.

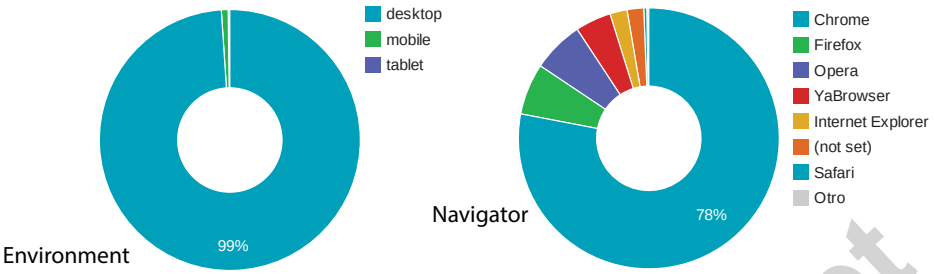


Figure 6: Statistics on environments and type of navigator.

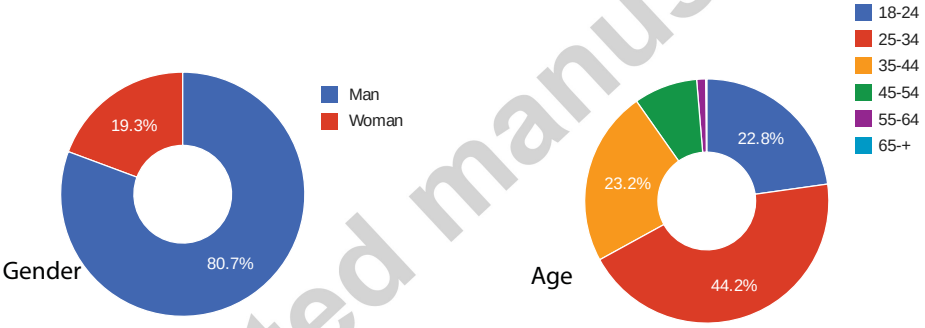


Figure 7: Statistics on gender and age.

5. New Case Study

While MOOCs gather extensive data from interactions between students and the course due to their highly technological nature, they are not the only education-related Big Data source available. Together with the rise of MOOCs, more and more universities have been adopting Open Data policies. Most universities host their own page for Open Data downloads in an effort to promote transparency in their operations. The data available may vary from one university to another, depending on how much investment the university has made in the Open Data initiative.

In our new case study we make use of Open Data extracted from the University of Alicante[5]. We have selected the data that contain the evaluation of

[5]http://datos.ua.es/busqueda-de-datos.html (visited on 6th of April, 2016).

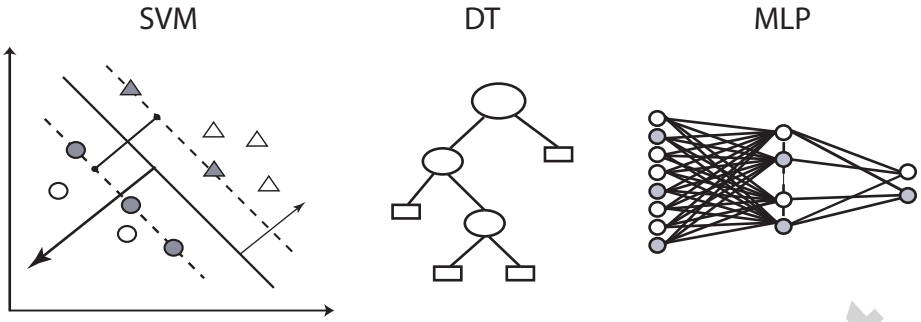


Figure 8: Techniques used to analyze the relationships in the data.

405 students in Degree courses and 847 students in Postgraduate courses. The data contain information including the age, city, Degree, matriculated courses and final result, among other personal data.

Our aim with this case study is to test our approach to identifying performance indicators through the use of data mining and visualization techniques, which provides us with insights into potential predictor attributes on which companies and organizations can focus to improve their results. In this case, the area we focus on is the course result itself, and the potential predictors of particular interest to us are those aspects that the university can influence (such as the number of matriculated courses, degrees that are apparently more difficult than they should be, etc.).

As mentioned, we have focused on the objective of "student achievement" which is measured by their results. We have applied the different processes in the method described in section 3 to extract the relevant KPIs. After the preprocessing, anomalies detection, difference series calculation, and pair-wise relationships analysis we concluded that there were no significant correlations. Therefore, we have analyzed the existence of compound relationships. In order to explore the data we have used data mining techniques which use the algorithms provided by Weka [12]. This application generates predictions and produces the most suitable visualizations for the analysis. The different techniques tested can be seen in Figure 8. The first technique tested was Support Vector Machines (SVM) [8], which tries to construct the vector that best separates the classes (in our case, results obtained) by using data points from the remaining data (age, city, and so on). The second technique tested was random forest of decision trees (DT) [5]. This technique creates multiple decision trees and tries to predict the result obtained by using attributes in a hierarchical way. Finally, the last technique tested was neural networks [3], specifically multilayer perceptron (MLP). Neural networks create a set of connections between units called neurons that are grouped into neuron layers. Each of these connections is assigned a weight, in such a way that from the set of input criteria (values related to age, degree, etc.) the output, the result, can be predicted.

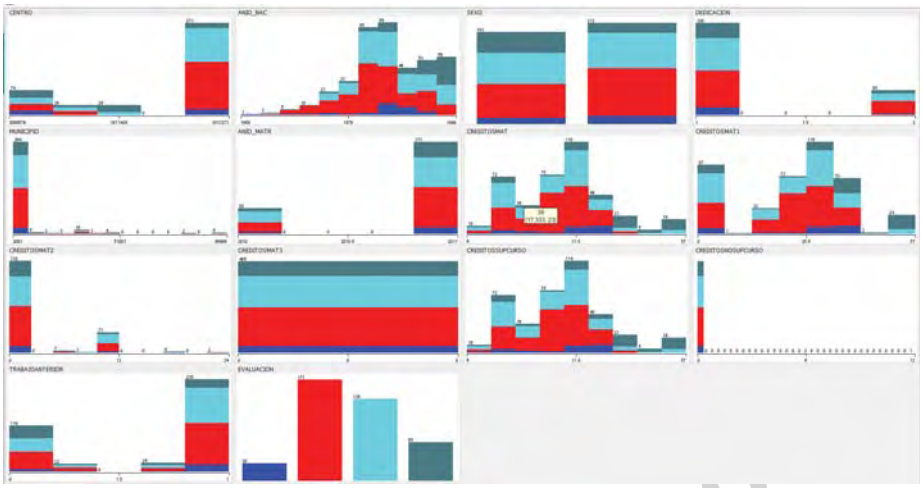


Figure 9: Weka output: correlation between the different attributes.

All the techniques we used provided similar results, with 80% to 84% accuracy, which provides us with reasonable evidence that at least some of the factors used as an input can determine the results obtained and could be used as indicators. In order to gain further insights into the most important factors we analyzed the structure of the data mining models built from the data. In Figure 9 we can see that the two most determinant attributes for the result are the number of credits for which the student has signed up (the attribute "CREDITOSMAT") and the geographical location where they come from (the attribute "MUNICIPIO"), the seventh and the fifth attribute respectively. This is further confirmed in Figure 10 where we can see the rules generated by the random forest algorithm which give more weight to these two characteristics. As can be seen in the rules, the abovementioned attributes are very crucial because depending on their values the students will get different results. From these results we can extract an indicator that can be influenced by universities, the number of credits signed up for, and another one which cannot be influenced directly, which is the city where students come from. The first factor seems intuitive, the more credits signed up for, the bigger the workload for the student and the greater the likelihood of poor results. In this sense, universities can recommend that students do not sign up for more than a certain number of credits, or can stop them from doing so, and the benefit of this can be determined on a university-by-university basis through analyzing their own data, thereby converting this attribute into a KPI for their decisions. On the other hand, the second factor is a bit less intuitive, since geographical origin should not be so dominant for the results obtained. Universities that obtain a similar result with their data may wish to analyze what factors differentiate a group of cities from others when it comes to student results.

CREDITOSMAT <= 39 AND MUNICIPIO = 3014.0: A (128.0/9.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 2003.0: A (4.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3065.0: A (42.0/2.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 99999.0: A (4.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3122.0 AND CENTRO > 3010545: A (16.0/1.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3121.0 AND SEXO = M: N (3.0/1.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3140.0: A (12.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3079.0: A (5.0/2.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3009.0: A (10.0/1.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3059.0: A (3.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3066.0: A (9.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3139.0: A (3.0/1.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3119.0: A (9.0/1.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3015.0: A (3.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3090.0: A (7.0/1.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3122.0: S (3.0/1.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3031.0: A (6.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3133.0: A (3.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3099.0: A (5.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3123.0: A (3.0/1.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3019.0: A (5.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3121.0: A (2.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 30043.0: A (4.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 46250.0: A (2.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3050.0: A (4.0)	CREDITOSMAT <= 39 AND MUNICIPIO = 3089.0 AND SEXO = M: N (2.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3044.0: A (4.0)	CREDITOSMAT <= 39: A (53.0/6.0)
CREDITOSMAT <= 39 AND MUNICIPIO = 3093.0: A (4.0)	

Figure 10: Weka output: Rules generated with RandomForest Tree.

6. Discussion

Dashboards are the preferred tool across organizations for monitoring business performance. They are often composed of different data visualization tech-

niques, amongst which Key Performance Indicators (KPIs) play a crucial role in quickly providing accurate information by comparing current performance against a target required to meet the business objectives.

Dashboards and Key Performance Indicators are important given the crucial role they play in providing quick and precise information. This is produced by comparing current performance against a target required to meet the business objectives.

Very often it is difficult to find an adequate KPI to associate with each business objective and this is where our proposal comes into play.

The main objective is to obtain specific candidate KPIs for business objectives in a semi-automated way. In this paper we have proposed a new methodology for extracting the relevant KPIs from the business strategy model of a particular enterprise/activity. In particular, we have defined a new process based on Data Mining techniques to identify the relevant KPIs. This consists of five processes: (1) Preprocessing, (2) Potential anomalies detection, (3) Difference series calculation, (4) Analysis of pair-wise relationships between series, and (5) Analysis of compound relationships. We have illustrated our approach with two case studies, one on MOOC courses, which is a very new area and therefore very suitable for this task, and another on Open Data from the education sector. In terms of future work, we plan to continue researching Big Data environments using visualization methods.

Acknowledgments

This work has been funded by the Spanish Ministry of Economy and Competitiveness under the project Grant SEQUOIA-UA (TIN2015-63502-C3-3-R). Alejandro Maté is funded by the Generalitat Valenciana (APOSTD/2014/064).

[1] C. Allison, A. Miller, I. Oliver, R. Michaelson, and T. Tiropanis. The web in education. *Computer Networks*, 56(18):3811 – 3824, 2012.

[2] Angoss. Key performance indicators, six sigma, and data mining. white paper., 2011.

[3] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[7] B. Clifton. *Advanced web metrics with Google Analytics*. John Wiley & Sons, 2012.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

[9] W. Fan and A. Bifet. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14(2):1–5, 2013.

[10] W. Fang. Using google analytics for improving library website content and design: A case study. *Library Philosophy and Practice*, 9(2):22, 2007.

[11] S. B. Green. How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, 26(3):499–510, 1991.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[13] T. Hill and R. Westbrook. Swot analysis: it's time for a product recall. *Long range planning*, 30(1):46–52, 1997.

[14] J. Horkoff, D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida, and J. Mylopoulos. Strategic business modeling: representation and reasoning. *Software & Systems Modeling*, 13(3):1015–1041, 2014.

[15] R. S. Kaplan, D. P. Norton, and P. Horváth. *The balanced scorecard*, volume 6. Harvard Business School Press Boston, 1996.

[16] R. S. Kaplan et al. *Strategy maps: Converting intangible assets into tangible outcomes*. Harvard Business Press, 2004.

[17] A. Labrinidis and H. Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[18] S. Lujan-Mora, J. Trujillo, and I.-Y. Song. A uml profile for multidimensional modeling in data warehouses. *Data and Knowledge Engineering*, 59 (3):725 – 769, 2006.

[19] P. J. Meyer. Attitude is everything: If you want to succeed above and beyond. *Waco, TX: The Meyer Resource Group*, 2003.

[20] M. Middelfart and T. B. Pedersen. Implementing sentinels in the targit bi suite. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, pages 1187–1198. IEEE, 2011.

[21] H. Pakkala, K. Presser, and T. Christensen. Using google analytics to measure visitor statistics: The case of food composition websites. *International Journal of Information Management*, 32(6):504–512, 2012.

[22] D. Parmenter. *Key performance indicators: developing, implementing, and using winning KPIs*. John Wiley & Sons, 2015.

[23] B. Plaza. Google analytics for measuring website performance. *Tourism Management*, 32(3):477–481, 2011.

[24] R. R. Rodriguez, J. J. A. Saiz, and A. O. Bas. Quantitative relationships between key performance indicators for supporting decision-making processes. *Computers in Industry*, 60(2):104–113, 2009.

[25] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1):97–107, 2014.

[26] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.