Knowledge Mining for the Business Analyst

Themis Palpanas[1] and Jakka Sairamesh[2]

[1] University of Trento
[2] IBM T.J. Watson Research Center

Abstract. There is an extensive literature on data mining techniques, including several applications of these techniques in the e-commerce setting. However, all previous approaches require that expert users interpret the data mining results, making them cumbersome to use by business analysts. In this work, we describe a framework that shows how data mining technology can be effectively applied in an e-commerce environment, delivering significant benefits to the business analyst. Using a real-world case study, we demonstrate the added benefit of the proposed method. We also validate the claim that the produced results represent actionable knowledge that can help the business analyst improve the business performance, by significantly reducing the time needed for data analysis, which results in substantial financial savings.

1 Introduction

Data mining has been extensively used in the past for analyzing huge collections of data, and is currently being applied to a variety of domains [9]. More recently, various data mining techniques have also been proposed and used in the more specific context of e-commerce [19, 10]. The downside of the previous discussion is that, despite all the success stories related to data mining, the fact remains that all these approaches require the presence of expert users, who have the ability to interpret the data mining results. We argue that an important problem regarding the use of data mining tools by business analysts is the gap that exists between the information that is conveyed by the data mining results, and the information that is necessary to the business analyst in order to make business decisions.

In this work, we describe a framework that aims at bridging the gap mentioned above. We demonstrate how data mining technology can be effectively applied in an e-commerce environment, in a way that delivers immediate benefits to the business analyst. The framework we propose takes the results of the data mining process as input, and converts these results into actionable knowledge, by enriching them with information that can be readily interpreted by the business analyst. By applying this methodology to the vehicle manufacturing industry, we show that the business analyst can significantly reduce the time needed for data analysis, which results in substantial financial savings. For example, shortening the vehicle warranty resolution cycle by 10 days can save an Original Equipment Manufacturer (OEM) around $300m and reduce the total number of warranty claims by 5%.

In summary, we propose a method that allows the analyst to:

– quickly discover frequent, time-ordered, event-patterns,
– automate the process of event-pattern discovery using a feedback loop,
– enrich these event-patterns with demographics related to the processes that generated them,
– identify the commonalities among these event-patterns,
– trace the events back to the process that generated them, and
– predict future events, based on the history of event-patterns.

Previous work has proposed algorithms for solving the problem in the first step of the above process, and we leverage those algorithms. Yet, there is no work that covers the entire process that we are proposing in this study.

1.1 Related Work

There exists a large body of work in knowledge extraction from huge datasets [8], and many recent studies try to improve on the performance and functionality of the various data mining algorithms. However, these algorithms are targeted to expert users, and are very cumbersome to be used by business analysts. The same is also true for commercial data mining solution packages [3, 4, 2]. The CRISP-DM project [1] has proposed an end-to-end process for data mining. In this paper, we present a specific framework that addresses some of the problems arising in one steps of the above process, namely, the step of organising and presenting to the user the new knowledge gained by applying some data mining techniques.

Several applications of data mining in the e-business environment [19, 10] prove the usefulness of this kind of techniques for improving the business operations. Nevertheless, the proposed solutions are specific to the particular tasks for which they were developed. Previous work has also studied the problem of applying data mining techniques in the context of data warehouses [14, 16, 17, 13], which are used by business analysts during the decision-making process.

2 Knowledge Mining Framework

In this section, we describe in more detail the framework that we propose for knowledge mining. Figure 1 depicts a high level view of our approach.

Pre-Processing: When the data first enter the system, there are a series of pre-processing steps that aim at cleaning the data and bringing them in a format suitable for our system. The purpose of the pre-processing steps is to make sure that all the data conform to the same standard and are semantically comparable. Examples of actions taken during this phase include converting all measures to metric system, all codes to the same standard, and ensuring the semantic equivalence of data.

Discovering Patterns: In the next series of steps, we identify patterns of interest in the data. The computed patterns should also conform to some user-defined

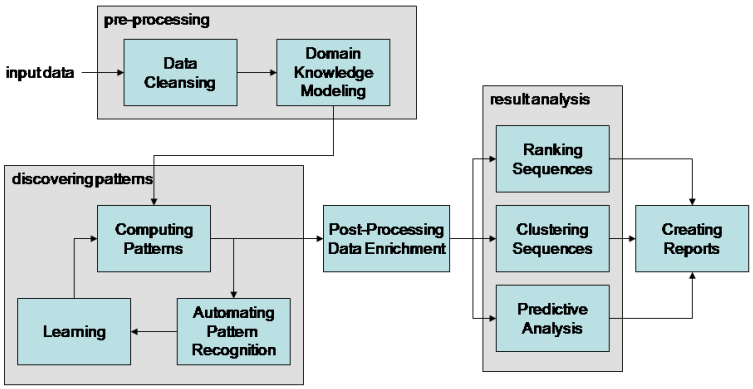

Fig. 1. Process overview.

constraints. These constraints determine the form of the patterns that are going to be computed. For example, the user may set the minimum and maximum lengths for each of the reported patterns, in the case where we are interested in mining for frequent sequences. The user may also define the maximum elapsed time between the start and the end of a sequence, as well as between two consecutive items in the sequence. The proposed framework, allows the users to try out the various parameter alternatives (a principle similar to exploratory mining [12]).

Data Enrichment: The frequent sequences that have been produced during the previous step only hold some minimal, vital information in order to identify the items that participate in each one of the frequent sequences. The goal of data enrichment is to take as input the computed frequent sequences, and correlate them with all the relevant bits of information that are stored in the system. Data originating from different parts of the business are gathered and integrated with the data mining results. The data may refer to various phases of the lifecycle of each specific item, and they enrich the discovered sequences with contextual information pertaining to the processes that generated them.

Result Analysis: The enriched sequences that were produced during the previous phase can then be analyzed in a variety of ways. The diagram of Figure 1 depicts three different analysis modules that we are proposing.

- **Ranking Sequences:** This module uses various knowledge models and utility functions in order to rank the event-patterns according to different criteria. The results of this methodology capture a macro view of the business performance issues based on a small but important fraction of the available information (for details see [6]).
- **Clustering Sequences:** The purpose of this module is to use the contextual information associated with each event-pattern in order to identify clusters of similar event-patterns. When a clustering algorithm (for example, [7]) is run

against the enriched event-patterns, it produces groupings of those patterns that are semantically meaningful within the business context, and help the analyst to gain insight on the root causes for each behavior.
- **Predictive Analysis:** This module aims at using the history of event-patterns to predict future events. The identified patterns represent an approximation of the historical behavior of the items under examination. Given these data, we can make projections for the future behavior [15].

Note that apart from the above three analysis modules that we have implemented in our system, other modules can be incorporated in the proposed framework as well.

Report Creation: In the final phase of the framework we propose, we produce a series of reports that summarize the results of the previous data analysis phases. In our framework we have developed a graphical user interface, through which the analyst has access and can customize several different report types.

3 Case Study

In this section, we describe a case study with one of the two vehicle manufacturing companies we collaborated with. The manufacturer has data relevant to the characteristics of each vehicle. The data refer to warranty claims made for vehicles of particular models during the period of the year 2005. The first dataset we examined includes almost $2,000,000$ records of warranty claims, referring to almost $1,000$ different failure reasons. These claims were referring to approximately $250,000$ unique vehicles, corresponding to more than 100 different vehicle models.

3.1 Proposed Process

In this section, we elaborate on the process that we propose for extracting new knowledge from the available data, for the vehicle manufacturing company. In the following presentation, we focus on a single stream of knowledge extraction and management, namely, that of *frequent sequences*.

Pre-Processing: The input to our system are the warranty claims of the vehicles. Each warranty claim comes with the following pieces of information.

- The vehicle identification number (VIN).
- The date that the vehicle visited the mechanic.
- The mileage of the vehicle at the time of the visit to the mechanic.
- The ids for all the failed vehicle parts, as identified by the mechanic.
- The root cause for each one of the failures, as identified by the mechanic.
- The cost of the visit to the mechanic, broken down by part and labor cost.

All these claims are gathered from different sources, normalized, and stored in a database called *warranty claims data store*. The organization of these data in a database helps in the subsequent steps of data analysis.

Frequent Sequence Mining: We define failure set $f = (f_1, f_2, ..., f_m)$ to be a nonempty set of failures (i.e., a set describing all the known events in which a particular vehicle failed), and failure sequence $s =< s_1, s_2, ..., s_n >$ to be an ordered list of failure sets. A failure-sequence which contains all the failures of a vehicle (identified by the VIN) ordered by the claim date, is called a vehicle-failure sequence. We say that a vehicle supports a failure-sequence if this failure sequence is contained in the vehicle-failure sequence of this particular vehicle. The problem of mining failure patterns [18, 11, 5] is to find the failure sequences that are supported by many vehicles.

Post-Processing Data Enrichment: The output of our frequent failure patterns analysis is a long list of failure sequences, which we augment with statistics relating to several of the attributes contained in the warranty claims database. More specifically, with each failure pattern we associate the following information (related to the vehicles that failed): Number of vehicles supporting the failure pattern; l most common vehicle models; l most common engine types; l most common manufacturing plants; l most common makers; l most common build-years. In addition to the above information, which refers to the entire pattern, we also associate with each particular failure of each failure pattern the following information: l most common cause-codes for the failure; Minimum, maximum, average, and standard deviation of the mileage at which the failure occurred; Minimum, maximum, average, and standard deviation of the replacement part cost for the failure; Minimum, maximum, average, and standard deviation of the labor part cost for the failure.

Result Analysis - Report Creation: The wealth of this information makes the use of a database imperative, in order to organize all the results and help in their analysis. Even though the database can be used to produce a listing with all the failure patterns along with the additional statistics outlined above, the real power of this approach stems from the fact that the analyst can query the database, and get results that are relevant to the particular aspects of the failure patterns she is focusing on.

The failure pattern database can be used to answer any query that correlates any combination of the attributes that the database captures (listed in the previous paragraphs). A small sample of the questions that this database can answer is presented in the following sections.

3.2 Evaluation Results Using Aggregated Behavior Reports

We first present sample results on failure sequence statistics related to the aggregated behavior of vehicles. This is a way to direct the analyst to examine some problems that are considered more important than others. Our framework can be useful in answering many other queries as well.

Ranking by the difference of part and labor cost for a specific type of failures.

In this case, we are looking for failure sequences that involve engine failures, for which the labor cost is more than the part cost. This query is interesting, because it shows that for some repairs under warranty the part cost is very small, while the labor cost is much higher (around $\$1,200$). When redesigning an engine, it may be beneficial to take these cases into consideration so as to make sure that that the labor cost for repairing this kind of failures is reduced (e.g., access to a particular part is made easier).

Ranking by frequency of occurrence for a specific engine model.
This query reports the most frequent failure sequences, for which the most common engine model is "A". These sequences are interesting, because they inform the analyst what are the most frequent recurring problems related to a specific engine model. Our data show that more than $2,300$ vehicles that are mounted with the specific engine model exhibit the same problems. These results can help identify potential problems in the design or manufacturing of this engine model.

3.3 Evaluation Results Using Focused Reports

In the examples that follow, we isolate some frequent failure sequences of interest, and analyze the detailed characteristics of the vehicles that exhibit these failure sequences.

Once again, it is important to note that the following are just two examples we used during our study. The approach we propose allows the user to focus on any aspect that is important to the team of data analysts.

Failures in brakes and electrical components.
This example focuses on vehicles that visited the mechanic two different times within the same year, for problems related to the brakes and the electrical components. Table 1 lists the most common causes for each failure in the sequence. As we can see, in the majority of the cases the failure cause is the same. This indicates that there may be a problem with the design or the manufacturing of the failed parts.

failure X cause code	%	failure Y cause code	%
inoperative	72	leaking	64
shorted	13	rubs	9
leaking	7	broken	4

Table 1. Example 1: Cause code break-down. (All the code ids and descriptions have been replaced by dummy values for reasons of anonymity.)

Failures in driving axle, wheels, and brakes.
In this case, we examine vehicles that visited the mechanic three different times during the same year, for problems related to the driving rear axle, the wheels, and the brakes. Note that all these problems relate to the same sub-system of the vehicles, and have occurred one after the other. When we look at the causes of these failures (see Table 2), it is obvious that the main problem is leaking

failure X cause code	%	failure Y cause code	%	failure Z cause code	%
leaking	100	leaking	47	leaking	21
		loose	5	broken	11
				loose	11

Table 2. Example 2: Cause code break-down. (All the code ids and descriptions have been replaced by dummy values for reasons of anonymity.)

parts. Furthermore, it turns out that all the vehicles that had those failures were manufactured in year 2004, and the vast majority of them, almost 90%, in the same factory (see Table 3). These data provide a clear indication to the analyst as to where to direct the efforts necessary for resolving the problems in the vehicles.

bld_dte	%	model	%	plant	%	engine	%
2004	100	M1	74	P1	89	E1	79
		M2	21	P2	5	E2	16
		M3	5	P3	5	E3	5

Table 3. Example 2: Demographics break-down. (All the code ids and descriptions have been replaced by dummy values for reasons of anonymity.)

3.4 Discussion

By following the proposed process, the analyst (or in this particular case, the engineer responsible for the design and manufacturing of engine type "E") can quickly focus on the most important problems that are relevant to her work. Actually, the same analyst can view, prioritize, and evaluate the corresponding information according to different criteria, such as cost, which relates to the financial aspect of the business, or frequency of failures, which relates to customer satisfaction and the marketing aspect.

These benefits of the presented method were also validated by different analysts from the two vehicle manufacturing companies that provided us with their data. By using our framework, they were able to not only cut down the time spent on data analysis and interpretation to a small fraction of the time they used to spend (from more than 45 days down to a few days for specific types of analysis), but they were also able to perform more focused analysis and deliver reports with a high impact factor.

4 Conclusions

In this work, we described a framework that enriches the results of the data mining process with information necessary for the business analyst. This information pertains to different aspects of the data mining results, and can help the analyst manipulate and interpret these results in a more principled and systematic way.

As our case study with a real-world problem demonstrates, the proposed framework has a great value in the e-commerce context. It converts the data

mining results into actionable knowledge, that the business analyst can use to improve the business operations. In our case study, this meant changing the design and manufacturing processes in order to avoid expensive warranty claims for specific failures.

References

[1] Cross Industry Standard Process for Data Mining. http://www.crisp-dm.org/.

[2] DB2 Intelligent Miner. http://www-306.ibm.com/software/data/iminer/.

[3] Microsoft SQL Server Business Intelligence. http://www.microsoft.com/sql/ solutions/bi/default.mspx.

[4] Oracle Data Mining. http://www.oracle.com/technology/ products/bi/odm/.

[5] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential Pattern Mining Using a Bitmap Representation. In *International Conference on Knowledge Discovery and Data Mining*, 2002.

[6] M. Chen and J. Sairamesh. Ranking-Based Business Information Processing. In *E-Commerce Technology*, 2005.

[7] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *SDM*, 2004.

[8] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[9] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2000.

[10] Y.-H. Li and L.-Y. Sun. Study and Applications of Data Mining to the Structure Risk Analysis of Customs Declaration Cargo. In *ICEBE*, pages 761–764, 2005.

[11] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. Technical Report C-1997-15, Department of Computer Science, University of Helsinki, 1997.

[12] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory Mining and Pruning Optimizations of Constrained Associations Rules. In *ACM SIGMOD International Conference*, Seattle, WA, USA, June 1998.

[13] T. Palpanas. Knowledge Discovery in Data Warehouses. *ACM SIGMOD Record*, 29(3):88–100, 2000.

[14] T. Palpanas, N. Koudas, and A. O. Mendelzon. Using datacube aggregates for approximate querying and deviation detection. *IEEE Trans. Knowl. Data Eng.*, 17(11):1465–1477, 2005.

[15] E. Pednault. Transform Regression and the Kolmogorov Superposition Theorem. Technical Report RC-23227, IBM Research, 2004.

[16] S. Sarawagi. User-Adaptive Exploration of Multidimensional Data. In *VLDB International Conference*, pages 307–316, Cairo, Egypt, Sept. 2000.

[17] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes. In *International Conference on Extending Database Technology*, pages 168–182, Valencia, Spain, Mar. 1998.

[18] R. Srikant and R. Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In *EDBT*, pages 3–17, Avignon, France, Mar. 1996.

[19] X. Yang, W. Weiying, M. Hairong, and S. Qingwei. Design and Implementation of Commerce Data Mining System Based on Rough Set Theory. In *ICEBE*, pages 258–265, 2005.