KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

A DATA MINING FRAMEWORK FOR CITARASA-BASED SYSTEM

Efthimia MAVRIDOU*[a,c], Dimitrios TZOVARAS[a] , Evangelos BEKIARIS[b] , Maria Gemou[b], George HASSAPIS[c]

[a] *Center for Research and Technology Hellas, Informatics and Telematics Institute, Greece*
[b] *Center for Research and Technology Hellas, Hellenic Institute of Transport, Greece*
[c] *Aristotle University of Thessaloniki, Department of Electrical and Computer Engineering, Greece*

ABSTRACT

This paper introduces the data mining techniques and results that have been produced utilizing a series of customer interviews in the automotive industry following the concept of citarasa, an innovative methodology that means emotional intent in Malay. The aim of the surveys were to identify affective needs of car and truck customers and figure out how these are interpreted in design elements of vehicles. Data mining methods were deployed for the discovery of the mapping mechanism between customers' affective needs and design parameters that characterize the design elements of vehicles. The generated mechanism was then used for the provision of personalized vehicle recommendations to customers.

Keywords: Citarasa, *Affective needs, Data mining, Associative Classification*

1. INTRODUCTION

Customers actively seek design features that are important for their emotional satisfaction. Affective needs are defined as user requirements for a specific product, driven by emotions, sentiments and attitudes [1]. Understanding customer affective needs is important to ensure a good fit of affective and functional requirements to design parameters. Citarasa engineering

* **Efthimia Mavridou**: Postal Address: Informatics and Telematics Institute (CERTH / ITI),6th Km Charilaou-Thermi Road 57001 (PO Box 361),Thermi-Thessaloniki, Greece. Email:efi@iti.gr

methodology [1] is an innovative methodology for affective needs elicitation. Citarasa is a Malay word which means emotional intent or a strong desire for a product. It is a synthesis of two words – "cita" meaning intent, aspiration, expectation, hope; and "rasa" meaning taste, feelings, and emotion.

The aim of this paper was to identify the relationship between customers' affective needs (defined by their citarasa) and design parameters related to the design elements of vehicles. As design elements we consider the vehicle parts that customers would like to customize (i.e. steering-wheel, seats). In proportion, as design parameters we consider the features that characterize the design elements (i.e. shape, material).

Data mining techniques were deployed for our purposes. Data mining (DM) enables efficient knowledge extraction from large datasets, in order to discover hidden or non-obvious patterns in data [2]. Our motivation for using DM was based on the hypothesis that the application of the appropriate DM technique on customer surveys could form a suitable mechanism for the knowledge extraction representing the correlation between customer affective needs and design parameters related the various design elements of vehicles. The extracted knowledge was then used for the provision of personalized recommendations to customers in collaboration with the agent-based framework developed in CATER [3].

2. METHODOLOGY

The DM process constituted actually the third step of the citarasa method followed, after the gathering and processing of customer surveys data. The surveys were conducted in the scope of CATER and included interview surveys of 140 truck drivers and 261 car drivers from Europe and Asia (China, Finland, France, Germany, Greece, India, Italy, Malaysia, Netherlands, Singapore, Sweden, Switzerland, the UK). The aim of the data mining process was to discover the mapping relationship between the identified affective needs and the corresponding design parameters of design elements of vehicles. In this paper we present a case study on the application of the methodology on data of car customer surveys. Table 1 includes the design elements (1st column) and their related design parameters (2nd column) that were included in this case study.

Table 1: Design elements and their related design parameters

Design elements	**Design parameters**
Steering – wheel	Material, Number of Spokes
Seats	Material, Shape
Wheels	Material, Number of Spokes
Side Mirror	Shape

Customer survey data was represented by the use of the categorical variables included in Table 2. The variable "Citarasa Descriptor" was used for describing the customers' affective needs. The variables "Design Element" and "Design Parameter" were used for describing the design elements of vehicles and the design parameters that characterize them, respectively. Demographic profile information was also taken into account including the geographic region the customer comes from, his/her age and gender. Table 3 includes the respective variables.

Table 2: Variables representing customer survey data

Name	**Values**
Citarasa Descriptor	Cute, Cool, Classic etc.
Design Element	Wheel-rims, Seats etc.
Design Parameter	Material, Shape etc.

Table 3: Variables representing demographic information for car customers

Name	**Values**
Region	Europe, Asia
Gender	Male, Female
Age	18-24,25-54,55-above

The above data preparation actions resulted in a data set of 261 records where each record corresponded to an individual car customer and his/her selections on specific design parameters of design elements that are included in Table 1. The purpose of this work was to discover the mapping mechanism between customers' affective needs and design parameters. To sufficiently deal with this task, a subdivision in *7* equal subtasks was performed. Each task was related to a pair of design parameter and design element (i.e. material/ steering wheel). As a result, there was a translation held in *7* implementation mechanisms that provide a mapping between customers' affective needs and the specific pair of design parameter and design element.

Towards this direction, the customers' survey data set was divided to *7* subsets, each one related to a pair of design parameter and design element. Table 4 includes a snapshot of the subset related to the design parameter "Material" and the design element "Steering-wheel". Each row in the subset corresponds to an individual user response related to the specific design parameter and design element. For example, row 1 corresponds to a male car customer who comes from Asia, his age is above 55, his affective needs are described by the citarasa descriptor "Classic" and would be satisfied by a steering-wheel made by wood.

Table 4: Snapshot of subset for the design parameter 'Material' of the design element 'Steering-wheel'

Region	Gender	Age	Citarasa Descriptor	Steering wheel / Material
Asia	Male	55-above	Classic	Wood
Asia	Male	55-above	Classic	Wood
Asia	Female	25-54	Modern	Aluminium
Asia	Female	25-54	Cool	Vinyl
Europe	Female	25-54	Cool	Vinyl

Association rule discovery (AR) techniques were applied to each one of the subsets for identifying associations among data. Association rule discovery refers to the discovery of the relationships among a large set of data items [4]. Let $I = \{i_1, i_2 ....., i_n\}$ be a set of items. Let D be a set of records, where each record R is a set of items such that $R \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I, Y \subset I$ and $X \cap Y = \varnothing$. X is the head of the rule and Y is the body. For each rule the confidence and the support are provided. The confidence c of a rule is defined as the number of records that contain X and also Y ($count(X \cap Y)$) divided by the number of records in D that contain X ($count(X)$).

$$c = \frac{count(X \cap Y)}{count(X)} \tag{1}$$

Confidence can be interpreted as an estimation of the probability of $P(X|Y)$. The support s of a rule is defined as the number of records that contain X and also Y ($count(X \cap Y)$) divided by the total number of records in D ($count(R)$).

$$s = \frac{count(X \cap Y)}{count(R)} \tag{2}$$

Associative classification (AC) is a special case of association rule discovery in which only the class attribute is considered in the rule's right-hand side (consequent); for example, in a rule such as $X \Rightarrow Y$, Y must be a class attribute [5]. In our case, the class attribute corresponds to the variable "Design parameter". The main task of AC is to construct a set of rules (model) that is able to predict the classes of previously unseen data, known as the test data set, as accurately as possible. Recently, several associative classification algorithms have been proposed. These include CBA [6], CPAR [7], L3 [8] amongst others. The general steps for the implementation of an associative classifier are the following [5]:

1. All frequent rules are discovered.
2. Class association rules are selected amongst the rules produced in the previous step. The most popular technique applied for this purpose is pruning. The rules that remain after the application of pruning posses a minimum required confidence level.
3. One subset of class association rules is selected to form an appropriate classifier, based on the rules in step 2.
4. The quality of the derived classifier is assessed by the use of testing data.

Within the specific framework the L3 (Live and Let Live) algorithm [8] has been deployed, which has been proven to appropriately classify data that were usually not covered or erroneously assigned to the default class by previous associative classifiers [8]. L3 extracts classification rules from data and then applies a lazy pruning technique in order to select high quality rules and build an accurate model of them. This technique performs a reduced amount of pruning by eliminating only rules that misclassify training data. For the prediction, high – quality rules (which are rules used in the classification of training data) are considered first, and if there is no match then spare rules are used (which are rules generated but not used in the training phase).

For each pair of design parameter and design element, a classifier based on association rules was constructed. The accuracy of the classifiers was assessed by a k-fold cross validation [9] process which is described in the following section. Indicative results that underline the importance of the methodology are also presented in what follows.

3. INDICATIVE RESULTS

The rule discovery process led to the generation of a set of rules for each pair of design parameter/ design element. Table 5 provides the rules generated for the design parameter "Material" of the design element "Steering-wheel". The rules provided an overview of the associations among data. For example, rule 1 implies that a customer whose affective needs are described by the citarasa descriptor "Modern" s/he would be satisfied with a steering-wheel of "Aluminum" material.

Table 5: Rules for Material /Steering – Wheel

No	Confidence	Support	Rule
1	0,461	0,031	Citarasa Descriptor = Modern ==> Design Parameter = Alumininum
2	0,414	0,015	Region = Europe and Citarasa Descriptor = Modern ==> Design Parameter = Alumininum
3	0,293	0,016	Age = 24-54 and Citarasa Descriptor = Cute ==> Design Parameter = Vinyl
4	0,281	0,000	Gender = Male and Citarasa Descriptor = Classic ==> Design Parameter = Wood
5	0,263	0,011	Region = Europe and Age = 24-54 and Citarasa Descriptor = Sporty ==> Design Parameter = Alumininum
6	0,234	0,011	Age = 18-24 and Citarasa Descriptor = Cool ==> Design Parameter = Alumininum

For each pair of design parameter/ design element a classifier based on the association rules was constructed by the application of a lazy pruning technique according to L3 algorithm. The predictive accuracy of the classifiers was validated by a k-fold process. According to this method, the dataset is divided into k subsets. Each time one of the k subsets is used as the test set and the other $k-1$ form the training set. The advantage of this method is that it does not depend on how the data gets divided as each one of the data instances takes part in the test set once and in the training set $k-1$ times. In this work, we have used 10 as a k value.

The accuracy (AC) of the classifiers is measured by the proportion of the total number of items that were correctly classified. It is determined using the equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The *True Positive* (TP) is the number of positive cases that were correctly classified. And the *false positive* (*FP*) is the number of negatives cases that were incorrectly classified as positive. In proportion, the *true negative* (*TN*) is defined as the number of negatives cases that were classified correctly and the *false negative* (*FN*) is the number of positives cases that were incorrectly classified as negative. Figure 1 includes the calculated predictive accuracy of the classifiers generated for each one of the pairs of design parameter / design element.

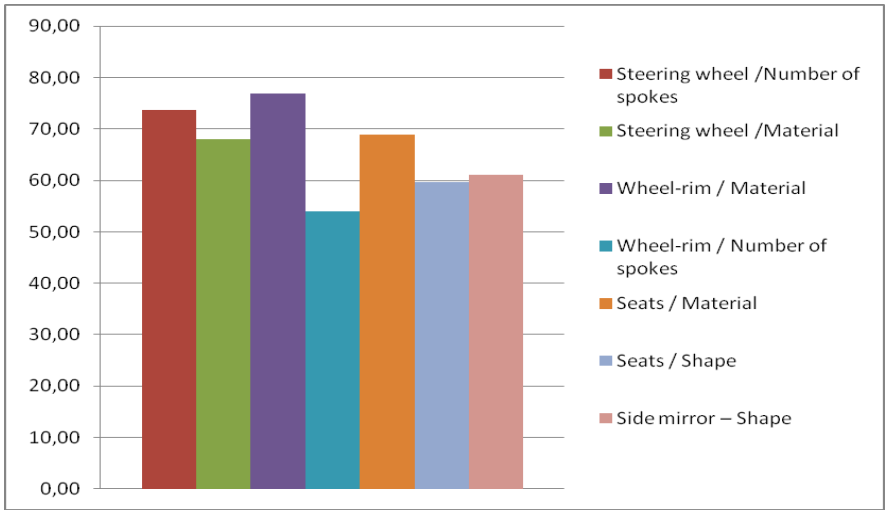

Figure 1: Predictive accuracy of classifiers

As it is depicted in Figure 1, most of the classification models have achieved a level of predictive accuracy above 60%. The highest value of predictive accuracy (76.89%) is observed for the model "Wheel-rim/Material", whist the lowest (53,95%) for the model "Wheel-rim/Number of Spokes". The use of more customer data from customer web surveys carried out in CATER is expected to lead in improvement of the DM outcome and hence accuracy of predictions.

The generated classification models form the prediction mechanism which generates for each design parameter a specific prediction based on the generated models. A complete recommendation is then proposed to the customer. Table 6 shows the predicted values for an individual user for each pair of design parameters and design elements. The example refers to a female car driver from Europe, who belongs to the age range of 25-54 and would like to have a "Cool" car.

Table 6: Predicted design parameters for a customer

Design element/Design parameter	**Predicted Values**
Steering-Wheel/ Material	Aluminum
Steering-Wheel/ Number of Spokes	Multiple
Wheel-rim /Material	Aluminum
Wheel-rim/Number of Spokes	Six
Seats/Material	Canvas
Seats/Shape	Curved
Side-mirrors /Shape	Angular

The predicted parameters are provided as input to the agent-based framework developed in CATER and are "interpreted" to configuration elements by the use of the configuration ontology. A complete vehicle recommendation is then presented visually to the user.

4. CONCLUSIONS AND FUTURE WORK

This paper presented a Data mining framework based on citarasa principles. The methodology followed provided a mapping mechanism of customers' affective needs described by their citarasa to design parameters related to vehicle design elements. Results derived on the application of the methodology on user survey data showed that the framework is capable of providing recommendations to the customers based on the generated mechanism. However, the need for more customer data and larger training datasets will be always a desirable option because it results in improvement of the DM outcome and hence accuracy of user recommendations. Future experiments will be conducted in order to evaluate the generated mechanism and measure the improvement introduced, compared to the initially evaluated rules.

Acknowledgments

Work reported in this paper was partially funded by the EU funded project CATER (http://www.cater-ist.org/), contract number: IST-035030.

The open source software package of data mining algorithms WEKA (http://www.cs.waikato.ac.nz/ml/weka/) was in our experiments.

REFERENCES

1. Khalid H.M., and Helander M.G, Customer emotional needs in product design. International Journal on Concurrent Engineering: Research and Applications, 14(3), pp.197-206, 2006.
2. Witten I.H. and Frank E., Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, June 2005.
3. Annex I-"Description of Work", CATER project, CN. 035030, Sixth Framework Programme, Priority 2-IST, Information Society Technologies.
4. Agrawal R., Srikant R., Fast algorithms for mining association rules In the Proceeding of the 20th International Conference of Very Large Data Bases, Santiago de Chile, Chile, 1994.
5. Thabtah F., A review of associative classification mining, The Knowledge Engineering Review, 22, pp 37-65, 2007.
6. Liu B., Hsu W. and Ma Y., Integrating Classification and Association Rule Mining, In the proceedings of the 4rd International Conference Knowledge Discovery and Data Mining (KDD-98), New York,1998
7. Yin X. and Han J., "CPAR: Classification based on predictive association rules", Proceedings 2003 SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, May 2003.
8. Baralis E. and Torino P., A lazy approach to pruning classification rules, In the proceedings of the IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, 2002.

9. Kohavi R., A study of cross-validation and bootstrap for accuracy estimation and model selection, In the proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo, 1995.