

Segmento: An R-based Visualization-rich System for Customer Segmentation and Targeting

Anestis Fachantidis
Department of Informatics
Aristotle University of
Thessaloniki
54124 Thessaloniki, Greece
afa@csd.auth.gr

Grigorios Tsoumakas
Department of Informatics
Aristotle University of
Thessaloniki
54124 Thessaloniki, Greece
greg@csd.auth.gr

Athanasios Tsiaras
Department of Informatics
Aristotle University of
Thessaloniki
54124 Thessaloniki, Greece
tsiarasa@csd.auth.gr

Ioannis Vlahavas
Department of Informatics
Aristotle University of
Thessaloniki
54124 Thessaloniki, Greece
vlahavas@csd.auth.gr

ABSTRACT

Customer segmentation is one of the most efficient and promising tools in a marketer's toolbox. In this paper, we introduce Segmento, an R-based customer segmentation system that uses clustering techniques to discover customer segments and offers tools to design and evaluate marketing campaigns. We present the features and the functionality of the system, as well as some of its unique, state-of-the-art visualizations.

CCS Concepts

•Information systems → Data analytics; Clustering;
•Computing methodologies → Cluster analysis; •Applied computing → Business intelligence; •Applied computing → Marketing;

Keywords

clustering, customer segmentation, visualization

1. INTRODUCTION

Customer segmentation deals with dividing a company's customer base into distinct and internally homogeneous groups of customers based on some of their characteristics (behavioural, demographic etc.). This allows marketers to create targeting strategies tailor-made for each segment. As a result, the risk in deciding to whom a product or service should be marketed is reduced and marketing efficiency can be increased.

Customer segmentation can be mainly achieved through business rules and data mining techniques [9]. The former

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SETN '16, May 18 - 20, 2016, Thessaloniki, Greece

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-3734-2/16/05...\$15.00

DOI: <http://dx.doi.org/10.1145/2903220.2903245>

approach can only handle a few customer features and is not always objective, as it depends heavily on the perceptions of a business expert. In contrast, data-driven analysis can be far more reliable and efficient. Clustering techniques can discover natural groupings of customers, which can lead to deeper insights and more efficient marketing strategies.

This paper presents Segmento¹, a customer segmentation system, developed with the Shiny Web application framework for R². Segmento can easily connect to a data warehouse, retrieve customer transaction data, aggregate them per customer for any user-specified time period and apply clustering techniques to discover customer segments. In addition, it features tools to observe the progress of the discovered segments in time and as a result, estimate their financial value. Finally, Segmento produces several key performance indicators (KPIs), which can be used to design and evaluate marketing campaigns based on customers associated to any of the discovered clusters.

The main innovative aspect of Segmento is the fact that, to the best of the authors knowledge, it is the only integrated customer segmentation tool that is entirely developed using the R language. It also represents the state of the art in terms of single and multi-variate visualization, featuring a total of 25 plots, created using the powerful plotting capabilities of R. Moreover many of the plots are interactive, and most of them are used for the first time in a customer segmentation software.

2. IMPLEMENTATION

Segmento is implemented as a Web application which can be hosted in a typical Web Server running the Shiny Server. It was developed entirely using R, SQL and the R Shiny Web framework. For its multi-variate visualizations Segmento uses `ggplot2`³ and `D3.js` plots⁴. In its current version, Segmento can connect to any SQL RDBMS through the use

¹A demo of Segmento can be found at: <http://segmento.csd.auth.gr:7169/>

²<http://shiny.rstudio.com/>

³<http://ggplot2.org/>

⁴<http://d3js.org/>

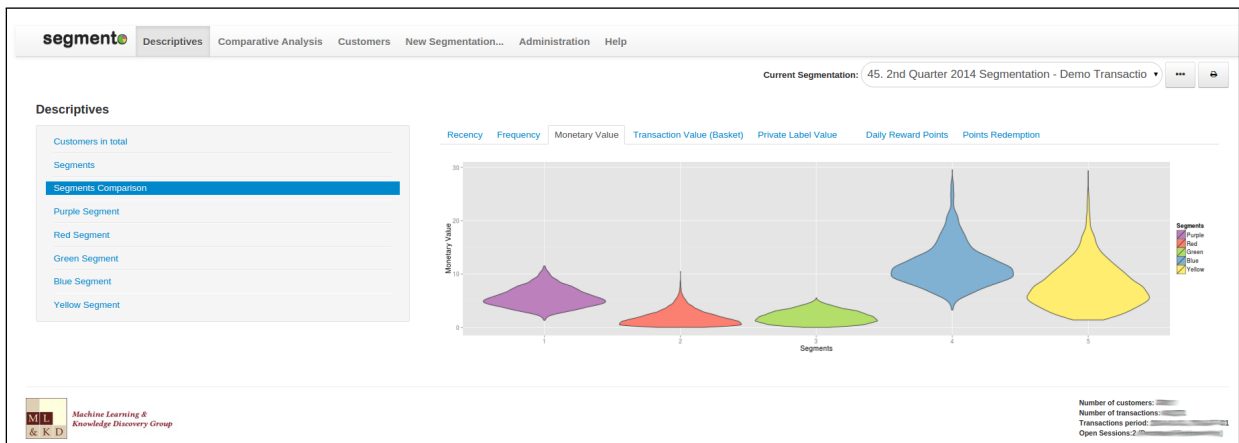


Figure 1. Violin plot comparing the *monetary* distribution in five segments.

of the ODBC or JDBC drivers from within R. In terms of efficiency, a significant part of data pre-processing is executed using SQL on the company’s database server whereas the application can be hosted separately and distantly, in the company’s Web server.

Typically, users within the company’s Intranet can connect through a standard Web browser to the IP associated with Segmento and use it, however restrictions can be applied so that Segmento will be used only on specific PC’s in the company’s Intranet. To minimize the network utilization and the number of queries to the database server, we have implemented a novel caching mechanism within Segmento.

Finally, all the data mining techniques used are implemented in R using efficient data structures such as the Data Table⁵. From our experimental setup on real transaction data, Segmento can efficiently handle million transactions and tens of clustering features.

3. FUNCTIONALITY

3.1 Customer Segmentation

The main functionality of Segmento is customer segmentation. Users have the option to perform a new clustering on a subset of customers of their choice, or inspect a previously created clustering. Users can freely select the data that they want to be used for the clustering by defining parameters such as the time period of the transactions. They can also exclude newly acquired customers from the clustering process, that is exclude customers whose profile has not been strongly established yet, or also exclude dormant customers that are not of any particular interest to them any more.

In addition, there is a large number of available segmentation criteria (clustering features) for the user to choose from. Available clustering features include the three attributes of the RFM analytic model (recency, frequency, monetary value), as proposed by Hughes [3], as well as other attributes like the average transaction value and the percentage of private label product value of each customer. The RFM model is proven to be very useful in customer segmentation processes, in that it efficiently differentiates important

customers [5], while at the same time, it is a model that marketers and decision makers can easily understand [4].

Another clustering parameter that is freely chosen by the users, is the weight of each clustering feature. As indicated by Stone [8], the three parameters of the RFM model, as well as any other segmentation parameter, are sometimes different in terms of importance, depending on the characteristics and the nature of the industry they are applied to. The capability of the users to define the feature weights, not only gives them the opportunity to include their prior field knowledge into the segmentation process, but it also makes Segmento suitable for use in many different fields of the industry sector.

For the segmentation process, users can define the number of clusters they want to produce. Then Segmento uses the k-means clustering algorithm [2] to produce the clusters, based on the previously defined weighted clustering parameters. The suitability of the k-means algorithm on clustering RFM customer data has been verified in the past and is well studied [6, 1].

Segmento offers a total of 22 plots and tables to visualize the exploratory and inferential analysis performed by the software, as well as the clustering results and the profile of each one of the customer segments. Figure 1 shows one of these plots: a violin plot comparing the *monetary* value of five discovered customer segments.

For the visual representation of the discovered segments, Segmento uses a color scheme assigning a unique color to each segment. The main advantage of such a color representation is that it discourages the one-dimensional interpretation of the produced segments. Since the clustering process is multi-variate and there is no order in the R^n space where $n > 1$, the segments of the customers are not just better or worse customers but *different* ones. The user should explore the descriptives and plots produced for each segment to find the key characteristics of each.

Finally, along with the descriptives produced for each segment and the total customer base, Segmento also produces correlation values between all the pairs of the available clustering features (such as the RFM values or the average basket value) as a useful indicator of their relation within each segment and how this relation changes between segments. For this, Segmento uses the non-parametric Spearman correlation.

⁵<https://cran.r-project.org/web/packages/data.table/>

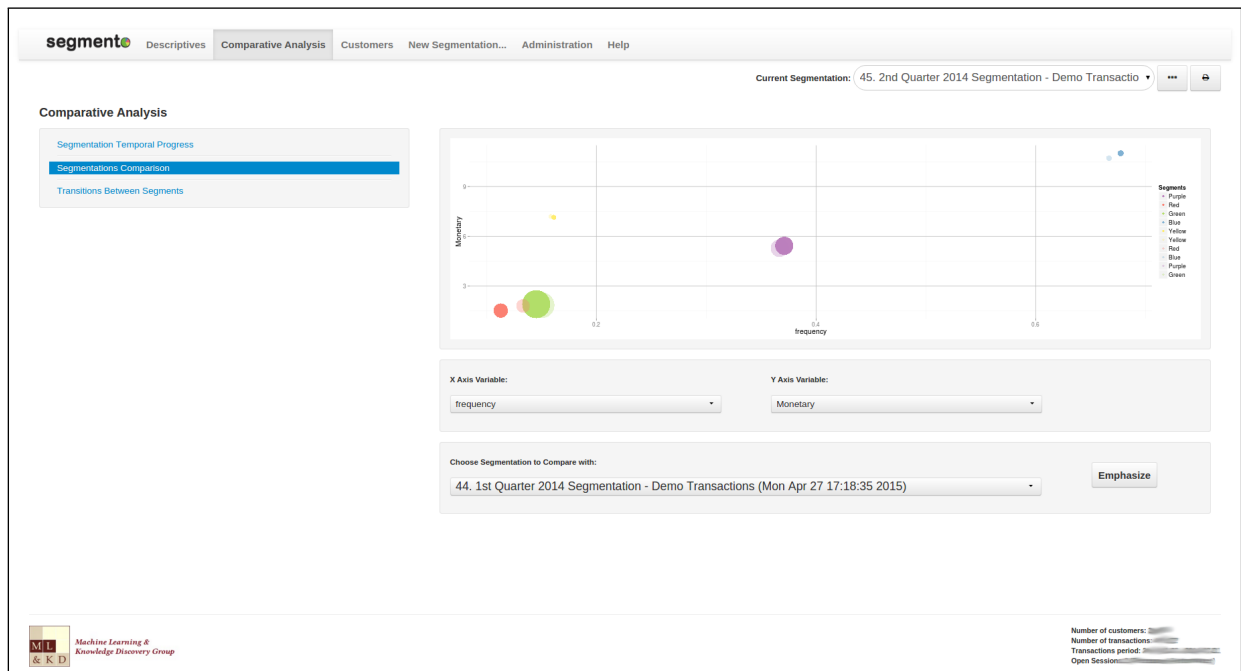


Figure 2. Bubble plot comparing two clusterings in terms of *frequency* and *monetary*.

3.2 Comparative Analysis

After the creation of customer segments, users can track their progress and behaviour over time using the comparative analysis functionality. Segment progress for each of the available KPI's (not only those available as clustering features) can be tracked for the whole time period for which relevant data are available. The histograms produced, showing the progress of each segment over time, can also be augmented with a 95% Confidence Interval(CI) useful for the visual assessment of significant differences between the progress of different segments.

Furthermore, users can also compare a newly created clustering with a previously created one (with similar clustering criteria), for any of the features used in them, as shown in Figure 2. An example on the importance of this comparison is that it allows the user to assess any shift of the center of the clusters, so that although a cluster X of good customers has become larger, its center may have been moved to smaller (worse) values for some KPIs.

Finally, Segmento detects transitions of customers between segments produced by different period clusterings and evaluate these transitions in terms of their financial profits or losses (see Figure 3).

To visualize all of the comparative analysis information, Segmento offers a total of 15 plots and tables, many of which are interactive. These visualizations can be easily used by the users to reveal complex patterns in the customer data, provide a clear understanding of them and the relationships between the distinct customer segments and their progress over time.

3.3 Customer Targeting and Campaign Evaluation

Segmento allows marketers to target customers and evaluate marketing campaigns based on the discovered market

segments. Users can select a subset of the customers from one or more segments that they want to target at. Then they can tag this customer subset with a label and export it either to the company's database or in an MS Excel format. The selected and exported Customer IDs can then be used directly for any marketing action (e.g. send specialized offers via mail, e-mail and/or the company's mobile app).

Following such a targeted a campaign, Segmento can keep track of the buying behaviour of this specific subset of customers and assist the user on evaluating the impact of the campaign. As an example, in Figure 3 the bar in the center of the figure shows the distribution of a customer's subset between segments, before and after a campaign. Moreover, Segmento produces a table with an estimation of the financial loss/gain of the selected campaign.

Existing targeting models and mining techniques for targeting, sometimes lack the ability to visualize and present the mined knowledge to the decision-maker. Especially for senior managers, visualization tools are very useful because they can provide quick but insightful knowledge without preconceptions [10]. By using Segmento, users can track changes in the values of the clustering features of the targeted customers and observe transitions of these customers between market segments through state-of-the-art visualizations, as shown in Figure 3. Furthermore, user can evaluate the financial impact of these transitions.

This is an important functionality of Segmento, as a typically overlooked situation in Business Intelligence projects is missing an environment to consume the modelling results, monitor the model's performance and overall results/ROI tracking⁶. Moreover, this functionality enables the users to easily perform A/B testing (randomized experiments with two variants, A and B) and statistically evaluate their marketing strategies.

⁶<http://goo.gl/tCZGYb>

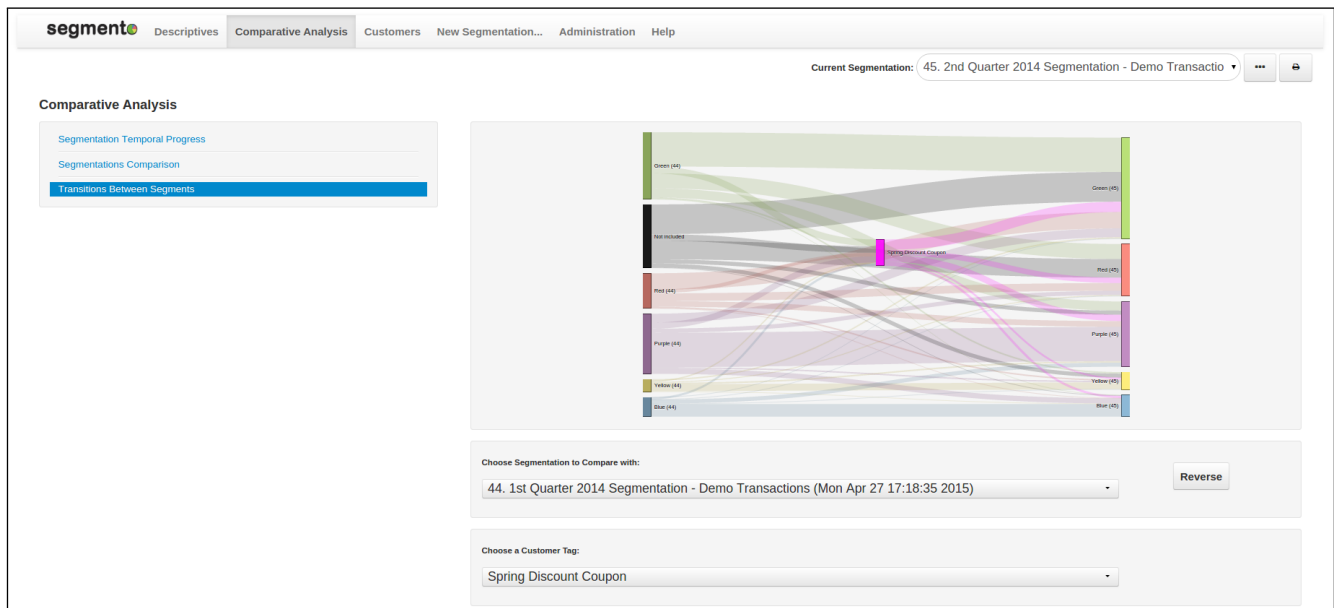


Figure 3. Sankey plot showing targeted customers' transitions between segments of two different clustering periods.

4. CONCLUSIONS AND FUTURE WORK

Segmento responds to the needs of a marketing department by discovering customer segments, indicating their significance, observing their progress over time and evaluating the segment-based marketing campaigns. It is an easy-to-use customer segmentation system that allows companies to produce marketing analytics in-house, without the need for specialized data scientist personnel. It features state-of-the-art visualizations that are unique among all relevant software and it is developed entirely using open source technologies, thus easily expandable and open to community improvements.

Future work for the improvement of Segmento includes the ability to perform clustering based on the cart profile (product history) of the customers and their demographic characteristics. Another promising addition could be mining for associations between purchased products, based on the buying habits of the customers, and finally sales prediction for specific products or services.

Acknowledgements

Anestis Fachantidis gratefully acknowledges financial support from the Hellenic Artificial Intelligence Society (EETN) for attending this conference.

5. REFERENCES

- [1] C.-H. Cheng and Y.-S. Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.
- [2] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.
- [3] A. M. Hughes. *Strategic Database Marketing*. McGraw-Hill Pub. Co., 1994.
- [4] C. Marcus. A practical yet meaningful approach to

customer segmentation. *Journal of Consumer Marketing*, 15(5):494–504, 1998.

- [5] F. Newell. *The new rules of marketing: how to use one-to-one relationship marketing to be the leader in your industry. Las nuevas reglas del marketing. Use el marketing de relaciones personales y será el líder de su industria..* Number 658. 802 N544E. 1997.
- [6] E. Ngai, L. Xiu, and D. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592 – 2602, 2009.
- [7] P. Samaras, A. Fachantidis, G. Tsoumakas, and I. Vlahavas. A prediction model of passenger demand using avl and apc data from a bus fleet. In *Proceedings of the 19th Panhellenic Conference on Informatics*, pages 129–134. ACM, 2015.
- [8] B. Stone and R. Jacobs. *Successful direct marketing methods*. NTC Business Books Lincolnwood, IL, 1988.
- [9] K. Tsipstis and A. Chorianopoulos. *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons, 2011.
- [10] J. Y. Woo, S. M. Bae, and S. C. Park. Visualization method for customer targeting using customer map. *Expert Systems with Applications*, 28(4):763–772, 2005.