Studies in Classification, Data Analysis,
and Knowledge Organization

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia
Biagio Aragona · Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Studies in Classification, Data Analysis, and Knowledge Organization

More information about this series at http://www.springer.com/series/1564

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia · Biagio Aragona
Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Editors

N. Carlo Lauro
Department of Economy and Statistics
University of Naples Federico II
Naples
Italy

Biagio Aragona
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Enrica Amaturo
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Marina Marino
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Maria Gabriella Grassia
Department of Social Sciences
University of Naples Federico II
Naples
Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-319-55476-1 ISBN 978-3-319-55477-8 (eBook)
DOI 10.1007/978-3-319-55477-8

Library of Congress Control Number: 2017937698

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data Science is a multidisciplinary approach based mainly on the methods of statistics and computer science suitably supplemented by the knowledge of the different domains to meet the new challenges posed by the actual information society. Aim of Data Science is to develop appropriate methodologies for purposes of knowledge, forecasting, and decision-making in the face of an increasingly complex reality often characterized by large amounts of data (big data) of various types (numeric, ordinal, nominal, symbolic data, texts, images, data streams, multi-way data, networks, etc.), coming from disparate sources.

The main novelty in the Data Science is played by the role of the KNOWLEDGE. Its encoding in the form of logical rules or hierarchies, graphs, metadata, and ontologies, will represent a new and more effective perspective to data analysis and interpretation of results if properly integrated in the methods of Data Science. It is in this sense that the Data Science can be understood as a discipline whose methods, result of the intersection between statistics, computer science, and a knowledge domain, have as their purpose to give meaning to the data. Thus, from this point of view, it would be preferable to speak about DATA SCIENCES.

The Data Science and Social Research Conference has represented an interdisciplinary event, where scientists of different areas, focusing on social sciences, had the opportunity to meet and discuss about the epistemological, methodological, and computational developments brought about by the availability of new data (big data, big corpora, open data, linked data, etc.). Such a new environment offers to social research great opportunities to enhance knowledge on some key research areas (i.e. development, social inequalities, public health, governance, marketing, communication).

Along, the conference has been a crucial issue to discuss critical questions about what all this data means, who gets access to what data, and how data are analysed and to what extent.

Therefore, aim of the conference, and of the present volume, has been to depict the challenges and the opportunities that the "data revolution" poses to Social Research in the framework of Data Science, this in view of building a SOCIAL DATA SCIENCE … Let us own data science!

Naples, Italy

N. Carlo Lauro
Professor Emeritus of Statistics

Contents

Part III On-Line Data Applications

Part IV Off-Line Data Applications

University of Bari's Website Evaluation

**Laura Antonucci, Marina Basile, Corrado Crocetta,
Viviana D'Addosio, Francesco D. d'Ovidio and Domenico Viola**

Abstract Educational websites were studied from many different perspectives. In 2001, Zhang and von Dran developed a theoretical framework for evaluating website quality from a user satisfaction perspective, while Yoo and Jin in 2004 evaluated the design of university websites. In this paper, we assess the quality perceived by the users of the website of the University of Bari using factorial analysis and multiple correspondences analysis (MCA) visual maps. Latent variables resulting from this preliminary analysis were then used to evaluate the most important latent dimensions related to loyalty of the users. A segmentation analysis was performed to study how loyalty is influenced by variables and factors.

Keywords Customer satisfaction · University website · CATPCA · Factorial analysis · MCA · Classification tree

1 Framework and Survey's Description

The university websites are the most important information channel, in fact they provide general information, facilitate contacts between teachers and students, etc. Quality and usability of the websites are, therefore, very important to improve student satisfaction.

L. Antonucci · C. Crocetta
University of Foggia, Foggia, Italy

M. Basile
Language/Techno-Economic State High School "Marco Polo", Bari, Italy

V. D'Addosio
Professional State High School "Ettore Majorana", Bari, Italy

F.D. d'Ovidio (✉) · D. Viola
University of Bari Aldo Moro, Bari, Italy
e-mail: francescodomenico.dovidio@uniba.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_11

This work aims to evaluate the user satisfaction of the website http://www.uniba.it, using a ten-section CAWI questionnaire: *User profile*, *Graphics of the website*, *Website contents*, *Services*, *Error Handling*, *Website management*, *Interruptions management*, *Usability*, *Security/privacy* and, finally, *Overall Satisfaction*. The first nine sections contain several items, measured with a four- or five-level scale.

2 Explorative Analysis

Table 1 reports the average scores given by the 1,049 respondents to the main aspects considered, according to the frequency of access to the website. This frequency has an important role because it allows to distinguish occasional users from expert ones.

21.9% of respondents access the website only in few occasions, but 10.7% declare that they browse the website several times a day. 67.4% of respondents visit the website one to several times a week. In most cases students are quite satisfied, the average mark ranges from 3 to 4 in a five-point scale, and there are not great differences between occasional users and expert ones, but expert users are a little more satisfied than the others.

An exception concerns, obviously, the item "reporting of errors/malfunctions during browsing", because frequent users are presumably annoyed by errors/malfunctions more often than occasional users.

3 Identification of the Website Quality's Dimensions

The Bartlett's test of sphericity for the observed 46 items was very significant (p-value < 0.0000001), allowing the use of principal component analysis (PCA) to explore the dimensions of website's quality.

Because some observed variables are measured on few level categories and not normally distributed, the ALSOS CATPCA was applied instead of PCA .[1] By using a backward stepwise procedure, only factors with eigenvalues higher than 1.1 were selected, iteratively removing all items with communality lower than 0.51. As final result, we obtained a correlation matrix with 25 optimally scaled items, identifying six principal components that explain 70.2% of the overall variance.

[1] The CATPCA (categorical principal component analysis) algorithm is due to the Data Theory Scaling System Group of the Leiden University, NL (De Leeuw et al. 1976; Meulman et al. 2004). It belongs to the PRINCALS family, based on *Alternative Least Squares Optimal Scaling* procedures, allowing researcher to use categorical variables, while PCA requires at least interval-scaled variables and normal distribution of residuals. Incidentally, also classic PCA was performed in explorative way, providing almost the same results than CATPCA.

Table 1 Average rate of significant items, according to the user's frequency of access to the website of the University of Bari Aldo Moro; percentages of users access frequency

Statistically significant items* ($p < 0.001$)	Frequency of access				*All users*
	Never/at times	About once a week	Several times a week	Several times a day	
Utility level of the published information	3.36	3.55	3.71	3.64	*3.57*
Level of depth and detail of the content	3.07	3.15	3.22	3.32	*3.17*
Comprehensibility of the used lexicon	3.85	3.94	4.03	3.87	*3.94*
Reporting of errors/malfunctions during browsing	3.38	3.23	2.99	2.93	*3.16*
Duration of the service interruptions	3.05	3.06	3.10	3.09	*3.07*
Download time	3.65	3.78	3.91	3.85	*3.80*
Viewing the site on any browser	3.63	3.71	3.87	3.76	*3.75*
Appropriateness of the content discussion	3.42	3.64	3.49	3.62	*3.55*
Comprehensible and unambiguous terminology	3.39	3.57	3.63	3.67	*3.56*
User recognition	3.82	4.01	4.10	4.08	4.00
Overall assessment about the website	**3.42**	**3.51**	**3.48**	**3.71**	**3.50**
% by access frequency	21.9	37.2	30.2	10.7	100.0

*Statistics significances were obtained by using the test of maximum likelihood ratio ($\alpha = 0.05$)

The Kaiser-Meyer-Olkin value is very high (0.92), ensuring excellent fitting of the model to data.

Starting from the identified principal components, a factor analysis (Cattell 1952) was conducted by using non-orthogonal promax rotation, in order to obtain a simpler solution. The promax rotation allowed to identify the most characterizing variables for each latent dimension, preserving relationships between the factors (Manly 1986).

Table 2 shows the residual correlations not due to direct relationships among the observed items. Only the first four factors have high correlation coefficients showing a *structural relation* among factors.

In Table 3, the *communalities* column indicates the variability explained by the factorial system, or in other words, the importance of the observed item. The factor loadings express the intensity of the relationship between variables and factors.

Table 2 Correlation among factors in the promax solution*

Factors	F1	F2	F3	F3	F4	F5
F1	1	**0.480**	**0.628**	**0.434**	*0.270*	**0.397**
F2		1	**0.553**	**0.474**	0.089	*0.282*
F3			1	**0.496**	*0.187*	**0.343**
F4				1	0.084	*0.183*
F5					1	*0.104*
F6						1

*Statistical significance = Bold font: $p < 0.01$; Italic font: $p < 0.05$

Table 3 Factor loadings and communalities of the items of the promax rotated solution*

Items	Factors						*Communalities*
	F1	F2	F3	F4	F5	F6	
Clarity of the site map	0.949						*0.793*
Information's accessibility in a few clicks	0.918						*0.785*
Map accessibility	0.857						*0.696*
Categories classification while browsing	0.822						*0.705*
Understandable terminology	0.683						*0.555*
Useful information on the site	0.521		0.350				*0.604*
Services/activities simplification	0.482		0.366				*0.571*
Opening speed of the pages		0.910					*0.839*
Website load speed		0.908					*0.815*
Download speed		0.879					*0.776*
Scrolling speed		0.836					*0.758*
Viewing the site on every browser		0.827					*0.705*
Comprehensibility of the used lexicon			0.868				*0.699*
Utility of the published information			0.849				*0.703*
Clarity of the contents			0.809				*0.730*
Level of depth and detail of the content			0.795				*0.713*
Adequacy of the contrast between font and background colour				0.855			*0.776*
Font size				0.808			*0.733*

(continued)

Table 3 (continued)

Items	Factors						*Communalities*
	F1	F2	F3	F4	F5	F6	
Visibility of the website features				0.712			*0.556*
Language selection					0.890		*0.760*
Responsiveness/alerts of technical inefficiency in the contact form					0.789		*0.633*
Accuracy/correctness of the translation					0.648		*0.606*
Error messages/corrective action						0.891	*0.811*
Alerts of errors or malfunctions						0.785	*0.640*
Error/data recovery						0.659	*0.593*

[*]Factor loadings lower than 0.33 have been omitted in this table

By evaluating such relationships, the factors can be then interpreted as follows:

- Factor 1: Accessibility and usability;
- Factor 2: Access speed;
- Factor 3: Information and content;
- Factor 4: Graphics and readability;
- Factor 5: Interactions;
- Factor 6: Error handling.

4 Proximity Map of the Observed Items

In order to confirm factorial similarities and to identify the main relationship, a visual map was used. Figure 1 shows the first two dimensions resulting from the multiple correspondence analysis obtained using the ALSOS algorithm: HOMALS (De Leeuw and Van Rijckevorsel 1980).

The position of the 25 centres of gravity of the observed variables highlights the relationships among the factors to which these variables are related (de Leeuw 1984; Gifi 1990). The points related to each factor are inserted in a shape with the corresponding number of the factor.

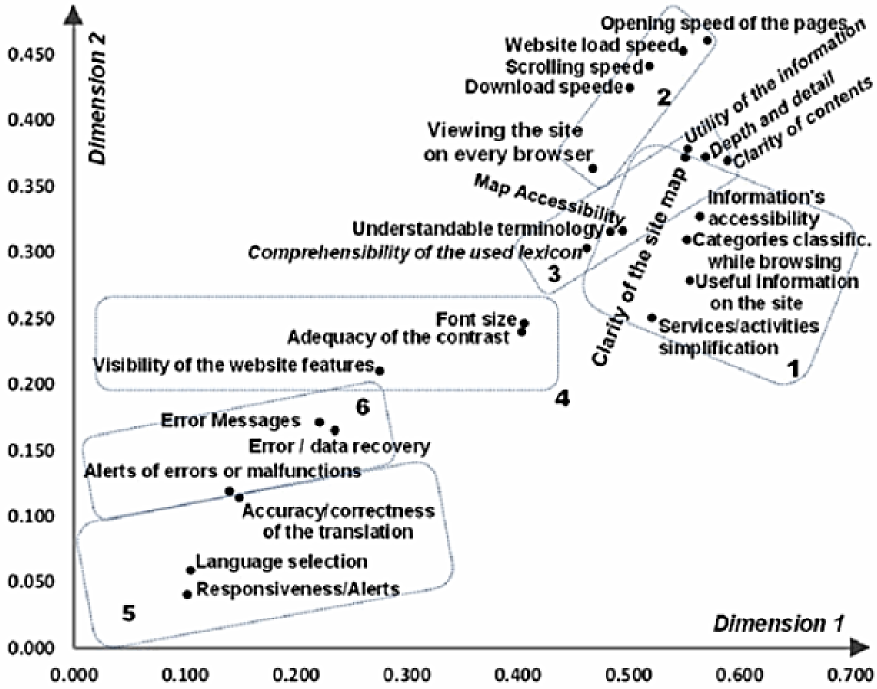

Fig. 1 Multiple correspondences map of observed items (first two dimensions)

The first two dimensions of MCA explain more than 70% of the total inertia. Figure 1 shows that the results of the factorial analysis are quite congruent with the two dimensions of the MCA.

The centres of gravity are concentrated along the main diagonal, ranking variables, and factors according to their importance with respect to the unidimensional concept of quality. The lower end of the diagonal (the less important items) is identified by the variables corresponding to factor "interactions", while the factor "access speed" identifies its upper end, i.e. the most important variables.

5 Quality Dimensions and Loyalty Elements

Loyalty can be predicted through classification methods. After many attempts, we choose to try a classification tree using the binary variable "access frequency" as response, where *high frequency* grouped the answers "several times a week" and "several times a day", while *low frequency* was associated with the other answers.

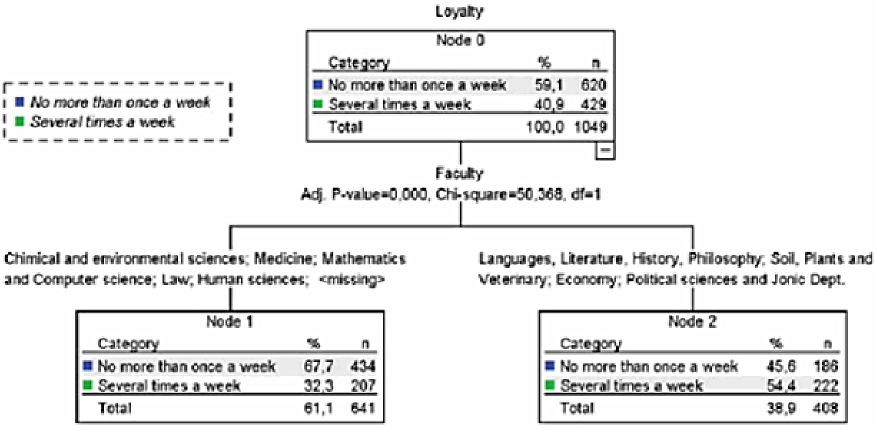
Loyalty
Node 0
Category % n
No more than once a week 59,1 620
Several times a week 40,9 429
Total 100,0 1049
No more than once a week
Several times a week
Faculty
Adj. P-value=0,000, Chi-square=50,368, df=1
Chimical and environmental sciences; Medicine; Mathematics and Computer science; Law; Human sciences; <missing>
Languages, Literature, History, Philosophy; Soil, Plants and Veterinary; Economy; Political sciences and Jonic Dept.
Node 1
Category % n
No more than once a week 67,7 434
Several times a week 32,3 207
Total 61,1 641
Node 2
Category % n
No more than once a week 45,6 186
Several times a week 54,4 222
Total 38,9 408

Fig. 2 Classification tree to predict the frequent access to the UNIBA website

All the interviewees characteristics (gender, residence, faculty, etc.) were selected as predictive variables, as well as the six quality factors identified above.[2]

The best known classification methods, CRT (Breiman et al. 1984) and CHAID (Kass 1980), were used, fixing 30 cases as minimum frequency of child nodes, expanded on maximum five levels of classification, and assessed by using cross-validation with 25 subsamples.

The chosen model, performed by using CHAID, can correctly predict the 62.5% of cases according the Faculty/Department (Fig. 2). The classification tree points out that students attending humanistic courses use the website more often than their colleagues of scientific courses.

The quality factors (precisely, "access speed", "information and content", and "interactions") appear at the second and third level of the classification tree, but without any effect on the predictive power of the model and thus they were removed by manual pruning.

The outcomes for the two cases "not more than once a week" and "several times a week" are quite different (see Table 4), because the latter response seems to be more difficult to identify.

The results here obtained are very good and robust, given that crossvalidation provides exactly the same risk values than the main classification (Table 5).

[2] The user's evaluation of the website could influence the frequency of access, because satisfied users tend (*cæteris paribus*) to browse the site more often than unsatisfied ones.

Table 4 Confusion matrix (classification table)

Observed website access frequency	Predicted website access frequency		
	Not more than once a week	Several times a week	Correct classification (%)
Not more than once a week	434	186	***70.0***
Several times a week	207	222	***51.7***
Total (%)	*61.1*	*38.9*	***62.5***

Table 5 Risk table

Method	Risk estimate	Std. error
Resubstitution	0.375	0.015
Crossvalidation	0.375	0.015

6 Concluding Considerations

This study showed a hierarchy of the variables, connected to the six dimensions of quality. Among them, the technical dimensions ("accessibility and usability" and "access speed") seem to be the most important, while the main mission of a website (providing *information and content*) has only the third position.

These findings were used, in addiction to the interviewees' characteristics, to analyse variables with respect to the loyalty proxy "access frequency to the website", by using segmentation analysis. Only a strong Faculty/Department effect was found, and this appears logical because, as it is known, the services are usually provided by these institutions following rules fixed at central level.

The main conclusion of this study is that the website quality has a weak influence on the "users loyalty", despite the current opinion "the higher the quality, the higher the loyalty".

Certainly, the analysis of the websites quality can not be limited to the few aspects described in the previous pages. This study should be considered just a first approach to the problem. Further analyses can start by the structural relationships here found among the quality dimensions, in order to find a causal model able to better explain the user behaviour.

Acknowledgements The authors jointly designed and realized the work here described. However, L. Antonucci wrote the Sect. 5, M. Basile wrote the Sect. 3, C. Crocetta wrote the Sect. 6, V. D'Addosio wrote the Sect. 2, F. D. d'Ovidio wrote the Sect. 4, and D. Viola wrote the Sect. 1. The Authors thank the University of Bari Aldo Moro (Italy) that provided the data. The usual disclaimer applies.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York-London: Chapman & Hall.

Cattell, R. B. (1952). *Factor analysis*. New York: Harper.

de Leeuw, J. (1984). *Canonical Analysis of categorical data* (2nd ed.). Leiden (NL): DSWO Press.

De Leeuw, J., & Van Rijckevorsel, J. (1980). Homals and princals—Some generalizations of components analysis. In: E. Diday, Y. Escoufier, L. Lebart, J. P. Pages, Y. Schektman, R. Tomassone (Eds.), *Data analysis and informatics* (pp. 231–241). Amsterdam, NL.

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternative least squares method with optimal scaling features. *Psychometrika, 41,* 471–504.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Manly, B. F. J. (1986). *Multivariate statistical methods: A primer* (p. 77). London: Chapman & Hall.

Meulman, J. J., Van der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 49–70). London: Sage.

Yoo, S., & Jin, J. (2004). Evaluation of the home page of the top 100 university web sites. *Academy of Information and Management Sciences, 8*(2), 57–69.

Zhang, P. & von Dran, G. M. (2001). Expectations and ranking of website quality features: Results of two studies on user perceptions. In *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS34)*, January.