Semantic Technologies used in the Data Analysis Process: Measuring Customer Satisfaction in context

Clemens Forster

Vienna University of Economics and Business
cforster@a1.net

Abstract. The objective is to transform indicator-coded variables extracted from open-ended questions by using semantic technologies rather than manual coding, in order to compare two forecast models and evaluate which yields the best results. Experiments are carried out within the domain of customer satisfaction. For this purpose, a leading Austrian financial institution provides a large sample of survey data. In the questionnaire, respondents are asked explicitly about their overall customer satisfaction. This dependent variable is to be predicted by use of statistical analysis techniques. For the process of model building, a procedure is required for generating coding categories in order to enable the use of unstructured verbatim data on critical incidents. One approach is based on traditional methods, while the other uses ontologies and reasoning.

Keywords: Semantic Technologies, Customer Satisfaction, Forecast Models, Survey Data, Verbatim, Model Building, Ontologies and Reasoning.

Introduction

Semantic technologies enable new forms of collaboration beyond the boundaries of applications. Through this inventive kind of technology, computers are able to put information to a wider variety of uses and provide a novel basis for reasoning, and this results in a direct effect on quality.

Quantitative research methods can be supplemented with semantic technologies to make use of unstructured data from qualitative research. Companies might use the integration of such procedures in knowledge management systems as a strategic advantage [1]. This can result, for example, in better understanding of customer perspective as well as enhancing controlling-correlated key performance indicators tied to balanced scorecards.

Problem

The coding of open-ended responses to survey questions by applying semantic technologies to utilize the information in the data analysis process is a new, and interdisciplinary area of application. Hardly any explicit literature referring to this topic exists. Combining "data and text mining" in business forecasting is also something which has only relatively recently been demonstrated experimentally. There have so far been no documented or published evaluations on the application of semantic technology in the coding of open questions, or its integrated usage within the scope of data mining utilized in building predictive modeling methods. Although applications for such combinations have been presented by well-known software companies, as solutions with high potential for the near future, the scientific evaluation of the integration of semantic technologies in forecast analytics is missing.

State of the Field

Terms of reference and state of the field vary between the borders of standard quantitative and qualitative research procedures.

Quantitative research methods in context

In the field of quantitative research methods, there is a wide variety of statistical and mathematical analysis procedures to choose from. The concept of "data mining" [2], is deployed for the purpose of pattern recognition i.e. in gathering new levels of understanding to make forecasts [3], frequently in connection with algorithms which are used on large databases and show the most efficient possible asymptotic consumption of computer resources.

Connection with qualitative research methods

In analytics and interpretation within qualitative research procedures, computer-aided techniques are becoming increasingly more used in order to analyse and interpret methodically evaluated data such as text files [4]. Despite the fact that critics warn of a possible adverse affect to survey design, creativity [5] and the procedure of exhaustive analysis, their use is nonetheless becoming more and more commonplace.

Transcribed text files from qualitative interviews, such as expert interviews, are becoming more commonly analysed through QDA (qualitative data analysis) applications in correlation with concepts like "the stages of open, axial and selective coding in the grounded theory", "thematic framework" or the "application of the most essential technique in objective hermeneutics" [6].

Responses to open questions determined by a specific survey design are regularly coded (founded on similar concepts used in QDA applications) to take advantage of statistical analysis. ESOMAR (European Society for Opinion and Marketing

Research) defines the process of coding as: “The organising of responses into categories and the assignment of a unique numerical code to each response prior to data entry.”

This transformation into quantitative components is subsequently used for analysis and can lead, in principle, based on the resultant additional information, to an improvement in predictive modelling. Manual coding efforts are nevertheless time consuming and subject to human error when coders develop certain rules for classifying ambiguous cases. Part of the conceived work is to evaluate a new method for improving this coding procedure through the application of semantic technologies.

Proposed Approach

The aim of the thesis is to develop and deploy an application by assembling open-source components capable of handling both the mathematical and statistical evaluation of analytical methods, as well as the classification of answers to open-ended questions. This will be carried out by the use of semantic technologies, in contrast to the use of manual classification and coding. In order to demonstrate a comparison between these experimental results, the quantification of any differences will be shown empirically.

Mathematical and statistical procedures

Most procedures for automatic and quantitative content analysis track the occurrence of specific words in n-dimensional space [7], or rather the co-occurrences [8] of words, and then use the information on frequencies for further clustering.

Rather than solely use these mathematical and statistical procedures, this paper will focus on the extended possibilities offered by semantic technologies, such as the classification of open answers to “meaningful higher-level unities” through reasoning by rules and ontologies in combination with Natural Language Processing methods.

Pre-existing knowledge in the form of upper-level ontologies

Automatic word sense disambiguation using a “light-weight ontology” like GermaNet [9] together with opinion mining and sentiment analysis methods [10] similar to the sentiment polarity classification will be applied.

The classification process should be extended to pre-existing knowledge in the form of upper-level ontologies like e.g. DOLCE, PROTON, SUMO [11] or DBpedia [12] to improve identifying key concepts for the further statistical analysis.

The procedural model is shown in the following figure:

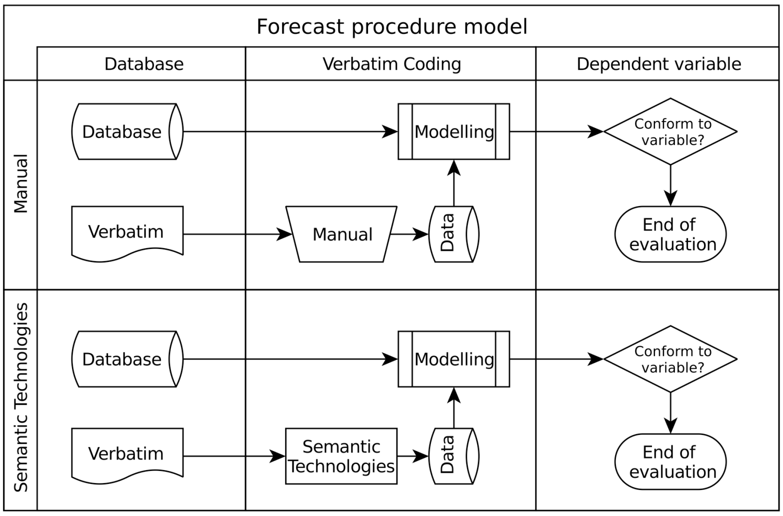

Fig. 1 Forecast procedure model and verbatim coding process

Quality assurance (with respect to the code-plan, inter-coder reliability and intra-coder reliability) will be implemented in the manual coding process. The forecasting method chosen for the experiment will be specified and applied in accordance with the ceteris paribus clause. The other coding method is based on ontologies and reasoning. The results of the experiment will yield data to determine which approach offers greater accuracy in predicting the overall satisfaction variable.

Methodology

For this purpose, a leading Austrian financial institution will provide a large sample of survey data. An extensive dataset with 21,146 interviews from 35 customer satisfaction projects/waves are available for analysis. The critical incident technique was applied using questionnaires. This resulted in 8,242 positive comments and 9,854 negative comments, which can potentially be used for the coding.

In the questionnaire, respondents were asked explicitly about their overall level of customer satisfaction. This dependent variable is to be predicted through statistical analysis techniques. The objective of this research is the theoretical foundation, analysis and development of a coding process for forecast models with semantic methodologies.

Results

These experiments have not been performed as of yet, and the results therefore remain to be determined. The concept for the work has, however, been developed. The means of obtaining the data have been established and the authorisation granted - a confidential agreement has been signed and several discussions have been held regarding the data-quality, as well as the procedure itself.

Conclusions and future work

The next steps will involve: development of the code-plan; performing the coding process with several coders; and applying quality assurance. Data understanding, data preparation and performing the modelling tasks are to follow subsequently. The work's conclusion will provide an answer to the research question of whether or not it is possible to substitute and/or enhance manual coding through the use of semantic technologies.

References

1. Dietz, J.L.G.: Enterprise ontology: theory and methodology. Springer, Berlin; NY (2006)
2. Berry, M.J.A., Linoff, G.: Mastering data mining: The art and science of customer relationship management. Wiley, New York, u.a. (2000)
3. Mertens, P., Albers, S.: Prognoserechnung. Physica-Verl., Heidelberg (2005)
4. Kuckartz, U.: Einführung in die computergestützte Analyse qualitativer Daten. VS, Verl. für Sozialwiss., Wiesbaden (2005)
5. Lindsay, V.J.: Handbook of qualitative research methods for international business. In: Marschan-Piekkari, R., Welch, C. (eds.). Elgar, Cheltenham [u.a.] (2004)
6. Wernet, A.: Einführung in die Interpretationstechnik der objektiven Hermeneutik. VS Verlag (2000)
7. Lourenço, A., Carreira, R., Glez-Peña, D., Méndez, J.R., Carneiro, S., Rocha, L.M., Díaz, F., Ferreira, E.C., Rocha, I., Fdez-Riverola, F., Rocha, M.: BioDR: Semantic indexing networks for biomedical document retrieval. Expert Systems with Applications, 37(4): 3444-3453, (2010)
8. Mazanec, J.: Eine "Landkarte der Werbeforschung": Schlagwortvisualisierung am Beispiel der Zeitschrift "transfer - Werbeforschung & Praxis". In: Strebinger, A., Kurz, H., Mayerhofer, W. (eds.): In: Werbe- und Markenforschung: Meilensteine - State of the Art - Perspektiven. Gabler, (Hrsg.) Wiesbaden (2006)
9. Finthammer, M., Cramer, I.: Exploring and navigating: Tools for germanet. (2008)
10. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2): 1-135, (2008)
11. Mascardi, V., Cordì, V., Rosso, P., Lemnitzer, L., Gupta, P., Wunsch, H.: A comparison of upper ontologies. Acquisition of a New Type of Lexical-Semantic Relation from German Corpora.: WOA07, Genova, Italy. Citeseer, IOS Press (2007)
12. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking open data on the web. 4th European Semantic Web Conference. (2007)