Visual Analytics for Supporting Manufacturers and Distributors in Online Sales

Olivier Parisot [1], Gero Vierke [2], Thomas Tamisier [1], Yoanne Didry [1], Helmut Rieder [2]

[1] Public Research Centre - Gabriel Lippmann
41, rue du Brill, L-4422 Belvaux
[2] InfinAIt Solutions SA
2a, Ennert dem Bierg L-5244 Sandweiler

Abstract: This article presents the basic concepts of OPTOSA, a Visual Analytics solution for the optimization of online sales, designed to support manufacturers in all phases of the online sales process from the product specification to the price fixing and more. OPTOSA combines a data processing module that builds and constantly updates operational knowledge related to sales positioning with a decision assistant that uses relevant aspects of the knowledge for helping the tasks of the different teams along the integrated chain of the sales. After reviewing the challenges of the approach, we discuss the first significant experiments with OPTOSA on some formalized use-cases.

1 Introduction

Everybody realize that the future of selling lies on the Internet. Therefore, most manufacturers and uncountable merchants devote themselves in a hard and continuous competition over the online marketplaces. However, many of them do not develop any structured approach to increase their market shares. Others use rather naive procedures, commonly based on lowering prices and thus reducing sales margins. An intelligent approach, based on timely usage of dynamic information directly available from the Internet will give us unique and decisive advantage.

Generally speaking, one of the greatest challenges for the manufacturers is that in online sales the distributors in the online market do not fulfill the same function as in the stationary market. In particular, in the stationary market the distributor, especially the procurement manager, supports the manufacturer with his know how on the consumer needs, namely the optimal specification and presentation of products. In contrast, in the online distribution, the manufacturer is given full freedom to manage all aspects of the sale process.

There is therefore a strong need to provide the sales business with an alternate solution that could make up for the human competence that is not at hand any more. This is the purpose of the OPTOSA system, developed in a partnership with researchers and practitioners, and specified after a thorough operational state of the art driven by three progressive investigation steps:

- In what way it would be possible to rely on data nowadays at disposal with a view to maximizing turnover and margin?
- Based on this knowledge from the data, which processes in online sales could be automated?
- Last, which decisions within the other processes could be supported by a knowledge based assistant?

With regard to the data, there has been a lot of research done and some tools are available in the domain of Search Engine Optimization (SEO), in particular [Sis, Wis], or tools from the search engines themselves like Google Analytics. However, those tools focus on search engines like Google and not on online market platforms. The ranking mechanism of search engines and market places are similar but not identical, the essential difference being that search engines aim to deliver most relevant content while market places aim to deliver products that are most likely to be bought. The goal of OPTOSA goes beyond this current state of the art, in order to figure out those differences and to develop algorithms that for instance analyze the meta-content of successful products and to extract the most relevant key-words and to compute the optimal structure for our meta-content.

Complementary, a lot of data sources like [Adv] propose business information that could be formalized to automatic handle or optimize several tasks in the sale process. Made of static knowledge about common best practices, this information can be outdated when applied to the fast evolving domain of online sales. In contrast, the OPTOSA system is based on the monitoring of live data in view of delivering to the sales collaborators an accurate business handling and a pro-active decision support.

In this paper, we present a software architecture that aims to optimize online sales. This architecture integrates several building blocks for data management, data mining, visualization, knowledge extraction and endly prediction.

The rest of this article is organized as follows. Firstly, challenges are presented. Secondly, the management of data is described in details. Finally, the architecture of the software is presented, and the results of experiments are discussed.

2 Challenges

Within the preliminary state of the art, we identified two main challenges towards a system that could cure the lack of expert knowledge.

The first one is to extract business information from the available data. A lot of sales data are available from online marketplaces, either directly to the general public or to registered users. As such, the raw data cannot be straightforwardly used to solve any concrete problem such as optimizing the sales through a given platform. We need to understand the data in their specific context of use, and then to mine them according to the knowledge we intend to build: remove outliers, find correlations, etc. Moreover, user control and interaction should be possible during the processing of so diversified bunches of data: we follow

here the *Visual Analytics* paradigm, a way of addressing such complexity through an efficient combining of the respective powers of the computer and the human brain [HA08].

The second challenge is to process this business information within concrete usecases by means of a collaborative expert system and a transparent data model. The OPTOSA system primarily intends to generate recommendations for operating on an online platform, dedicated to the sales collaborators. The recommendations must be in any case supported by a clear and suitable justification. The justification exhibits both the trace of the reasoning and available references to knowledge extracted from the data. The traceability brings constraints on the architecture and the technology used to implement the system. In particular, OPTOSA will use a referencing mechanism associates rules with documentary sources. Also, when updating the knowledge base, the references allow identifying the impact of a given knowledge source on the rules model, in order to replace them.

Sales processes are modeled through different procedures and algorithms. We structured the functionalities of the OPTOSA system according to separate optimization needs that could be refined independently:

- Direct optimization of the sale of a product.
- Optimization in relation to sales and delivery channels: How to deal with competitors?
- Reporting and analysis: How to improve the offer through the consolidation of the sale channels?

As mentioned above some of the functionalities can be fully automated, given the values of parameters asked by the system. An example is the dynamic price adjustment: for each product to be sold, the user can specify a minimal and maximal price. OPTOSA will autonomously monitor the market situations, especially the prices of the competitors, and finally adapt the product price in such a way that turnover and margin is optimized. Other functionalities can be only partially automated and serve as decision support for the user. For instance, when description text or search tags are to be defined, OPTOSA will scan other, comparable product descriptions on the sales platforms, and evaluate which key words correspond to sales success. From this it computes suggestions that are presented to the user. In any case, the input data processed by the rules is information extracted through data mining from recorded characteristics of past sales. The business knowledge to process the data is specified and tested in partnership with intended users of the system.

3 Data handling

In order to tackle the optimization of online sales, data about sellers and products have to be collected, preprocessed, analyzed and finally used for extracting knowledge models for decision support.

In fact, the OPTOSA platform aims at providing decision support features [TPD+11] by combining:

- The insights that are automatically discovered from data (for instance: 'the price of a given product should be lower than the price of similar concurrent products, with at least a difference of X%').
- The business know-how that is brought by the user (for instance: some rules on prices to take into account margin).

As a result, the goal is to provide *predictive models* to help *price positioning*: a typical output can be a model that evaluates the price of given product (according to the product features and the similar products prices). The following sections briefly describe each step of the process that supplies the OPTOSA platform (Figure 1).

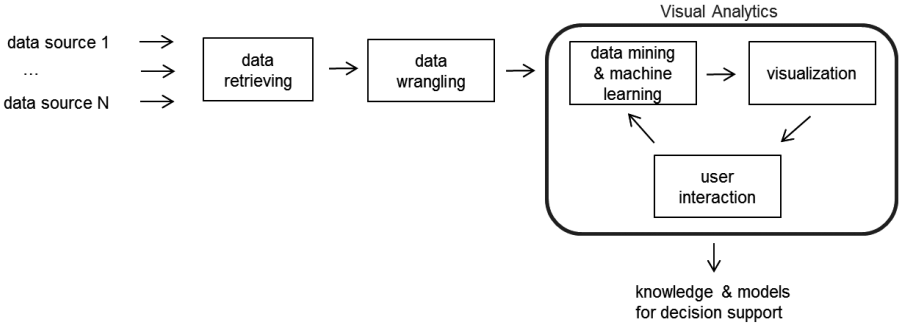

Figure 1: The worflow that is applied into the OPTOSA platform for data handling.

Recent papers about *data analysis* argue that *Visual Analytics* is needed to build accurate models by considering both data and *business knowledge* [KAF+08, KHP+11]. But is it realistic to apply the straightforward *Visual Analytics* approach in a real-time system which continuously manages huge heterogeneous data? The OPTOSA platform aims at combining state-of-the art techniques and pragmatical choices in order to provide operational results.

3.1 Data retrieving

By using online marketplace platforms like *eBay* or *Amazon*, it is possible to retrieve and aggregate a lot of data about products and sellers. The best way is to use the API that are provided by these systems. Generally stable and well-documented, these API allow to get data easily in order to store them into a *local* database for further analysis. In addition, the produced database can be completed by using services provided by dedicated data providers like icecat ([ice]).

3.2 Data wrangling

Traditionally used as a preliminary step in data visualization and knowledge extraction, data wrangling is a semi-automated process that aims at transforming data in order to facilitate further usage of them [KHP+11]. More precisely, data wrangling consists in iteratively applying different kinds of preprocessing tasks like *inconsistent data cleaning*, *missing values imputation* or *sampling* [ET98]. As it induces modifications and information loss, data wrangling has to be carefully applied; in addition, it is a painful and time-consuming task that requires efficient tools [KPHH11].

We plan to use different tools to wrangle data: Excel for quick transformation prototyping, and specific modules for more complexes tasks into the OPTOSA system.

3.3 Decision support & Visual Analytics: data mining, machine learning, visualization, user interaction

Once the data are ready to be analyzed, a classical *cascading process* can be implemented in order to build efficient predictive models [CTTB06]: in order words, the idea is to mix unsupervised and supervised machine learning methods in order to obtain efficient models for each kind of products. In a few words, the process is the following:

- *Features selection* in order to filter the data that are the most significant [YL03].
- *Clustering* in order to regroup similar objects and to detect/ignore outliers. [Jai10].
- Computation of *predictive models* for each cluster. In this field, regression trees are helpful and popular because they use an intuitive formalism that is easy to understand for domain experts [B+84, Mur98]. Moreover, they can be built from data by using well-known algorithms like CART [B+84] and M5 [Qui92] (Figure 2).

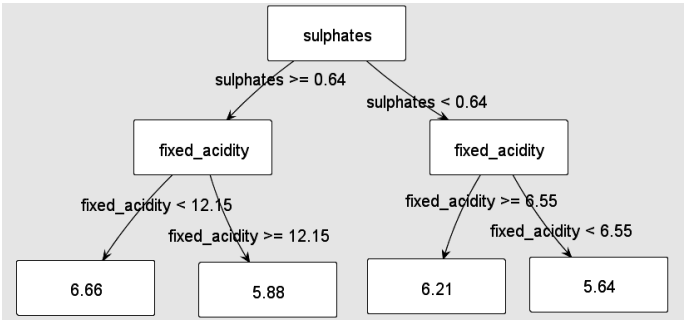

Figure 2: A regression tree obtained from the 'winequality-red' dataset [BL13], to evaluate the wine quality score: a score of 0 represents a poor wine and a score of 10 represents an excellent wine.

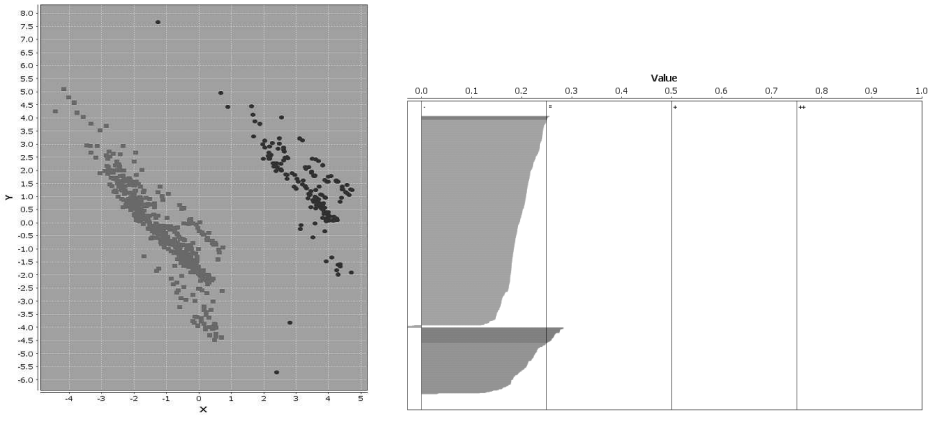

Figure 3: Visualizations to inspect a clustering result ('*credit-a*' dataset, clustered with k-means). At the left, a colored PCA projection: the x axis represents the first principal component, and the y axis represents the second principal component. At the right, a Silhouette plot: the x axis represents the Silhouette value, and the y axis represents the elements (sorted by clusters and Silhouette values). In practice: larger Silhouette values indicate a better quality of the clustering result.

Even if this process is well-known, obtaining operational results is not straightforward: in fact, various parameters can have a great influence on the final results. For instance, the usage of clustering is not trivial and depends of the final application [AR13]: what kind of clustering algorithms should be used? Which settings have to be applied? Should some data be manually processed? Generally speaking, applying machine learning can be considered as a *black art* that requires a strong background and experiment [Dom12].

However, pragmatical solutions can be considered. For instance, by following the *Visual Analytics* approach [KAF+08], and by providing interactive features to the final user, results can be produced, evaluated, visually inspected and refined in order to obtain clusters are simple to interpret and predictive models that are efficient.

More precisely:

- Clustering can be adapted to refine the predictive models: in an automatic way [PDBT14], or by taking into account the user domain knowledge.
- Predictive models can be simplified by using various simplification techniques: for instance, a recent technique has been proposed to simplify them by using data preprocessing [PBDT13].
- Results can be visualized using simple scatter plots or more sophisticated visualizations like MDS projections [KW78].

As a result, the predictive models can be used to support decision regarding the *price positioning* objective.

4 The OPTOSA platform

A first prototype has been implemented in Java as a standalone tool. The core system is based on several libraries that provide building blocks for *Visual Analytics* and Decision Support:

- Data wrangling, data mining and machine learning features are provided by Weka, a widely-used data mining library [WFH11]. More precisely, it provides several implementations for features selection, regression trees induction and clustering.
- As the Weka licensing policy can cause further issues in the case of commercial uses, some alternatives to Weka have been tested and integrated: Sandia Cognitive Foundry ([BBD08]) and apache-commons ([Com]). These libraries are two main advantages: they provide implementations of the state-of-the-art techniques, and they are distributed under licences that are permissible and compliant to potential commercial use.
- For the graphical representation of the price evaluation model, the tool uses Jung, a robust and well-known graphical library [OFWB03].

In addition to these building block components, the prototype is completed with a graphical interface for manipulating data, computing predictive models, exploring the results (statistics about the datasets, visual representations of the clusters and the predictive models, etc.) and taking into account the user knowledge (Figure 4).

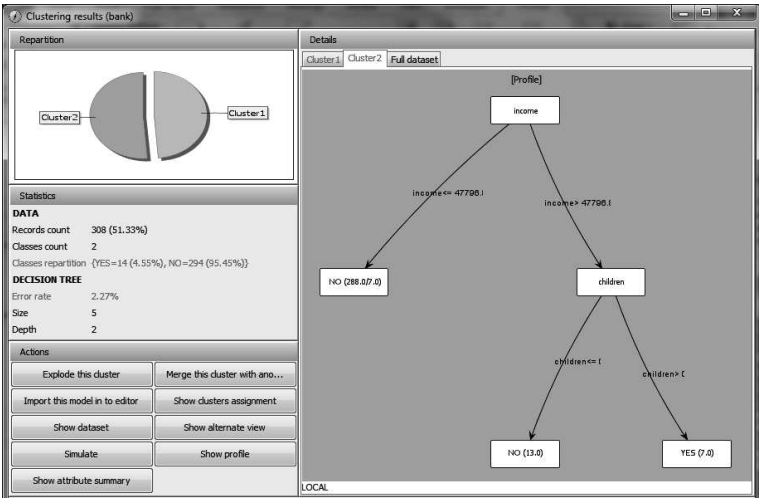

Figure 4: The prototype module that allows to check and refine if needed the clustering results.

5 Use-case

For our first technical trials, we processed datasets relevant to different business use cases. As explained in the previous sections, the knowledge is formalized at the end on the shape of decision trees and regression trees [Kot13], based on a user readable formalism, the trees being directly extracted from real life data.

Data considered as input are: the list of products and information about each product (features description, prices, sales ranks, etc.). The information can be structured in the shape of clusters of products, and for each cluster the products profile is defined (core features, prices ranges, sales rank ranges, etc.). The use cases are groups of tasks defining some concrete sales activity, such as the following examples. Some use cases are principally handled by data mining (on relevant data) such as 'Price fixing' based on similar products. Other ones contain more domain related knowledge (from vendor or buyer point of view) such as 'Calculation of a similarity index between products'.

In Figure 5, we illustrate the 'Price positioning' use case for a laptop. First, from a dataset, a decision tree is computed to determine the price range according to the features of the laptop. The decision model is then applied to evaluate and justify the price of a given product. Second, the knowledge is used for the computation of a clustering of the laptops recorded on a dataset, according to their main features. A planar projection of the laptops shows the correspondence between the features (horizontal axis) and the price (vertical).

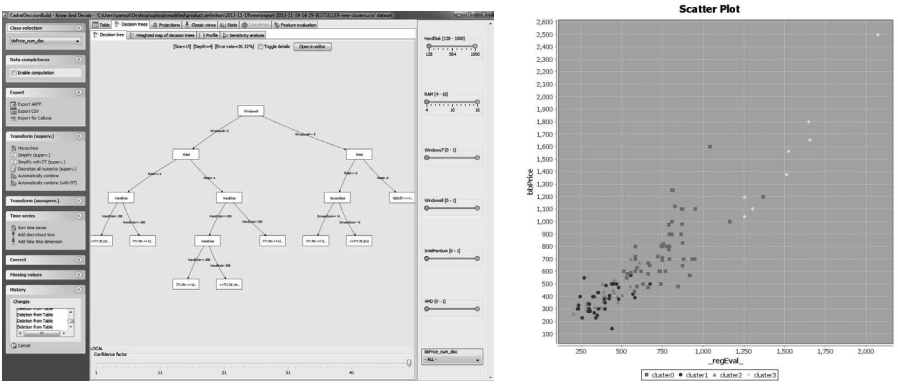

Figure 5: *Visual Analytics* for price positioning.

By using these models, a user interface has been built to interactively evaluate the price of a given product (Figure 6). In this interface, the evaluation is completed by a justification: here, the justification corresponds to the rules that are *verified* for the given product. In addition, several data mining features have been added to complete the evaluation: for instance, the 'k-nearest neighbours' technique allows to automatically find the products that are the most similar to the given product, so the price positioning result can be checked.

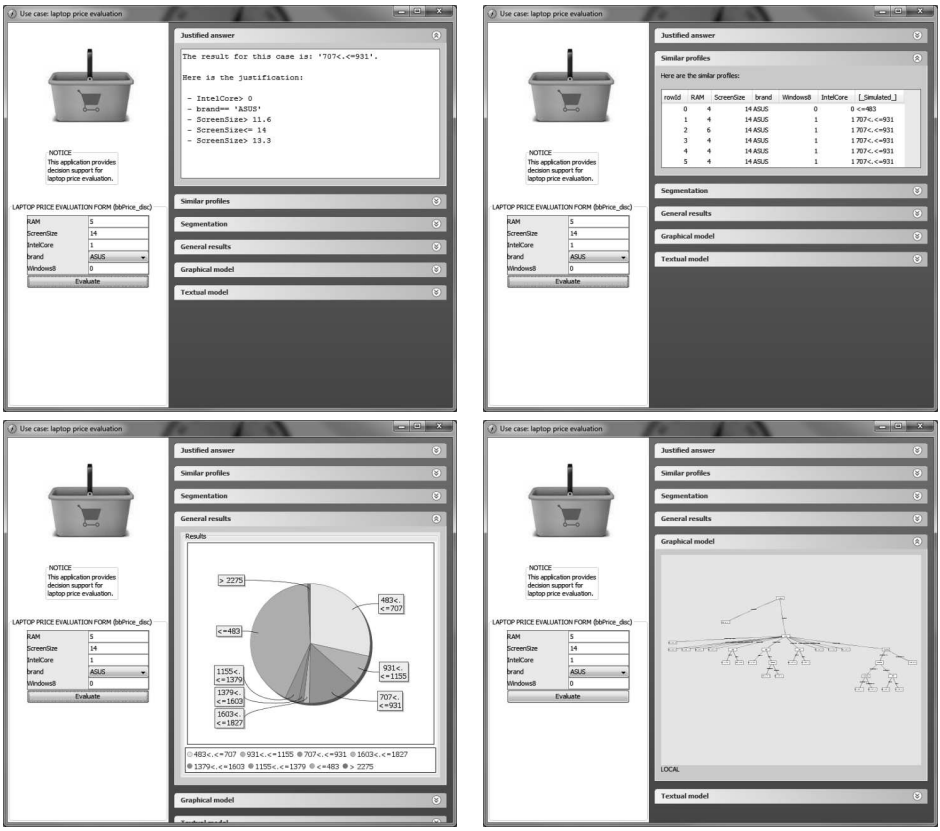

Figure 6: User interface to evaluate the price of a given product. At the top left, the price and the justification are given. At the top right, the similar products are shown, in order to check if the evaluated price is compliant with the others. At the bottom left, the clusters are represented into a pie chart. At the bottom right, the graphical model is presented.

The processing of the use cases shows the possibility to extract knowledge, track evolutions and build decision models. From the debriefing on the data mining results, it was also clear that a lot of business information must be taken into account to reinterpret the results, in order to be able to formalize rules that could automate the handling of marketing activities.

Finally, the integration of expert of business knowledge is a critical step for decision support: to handle this issue, the prototype provides a module to edit the price evaluation model (Figure 7). By using it, the user is able to adapt and refine the price positioning module according to the business requirements.

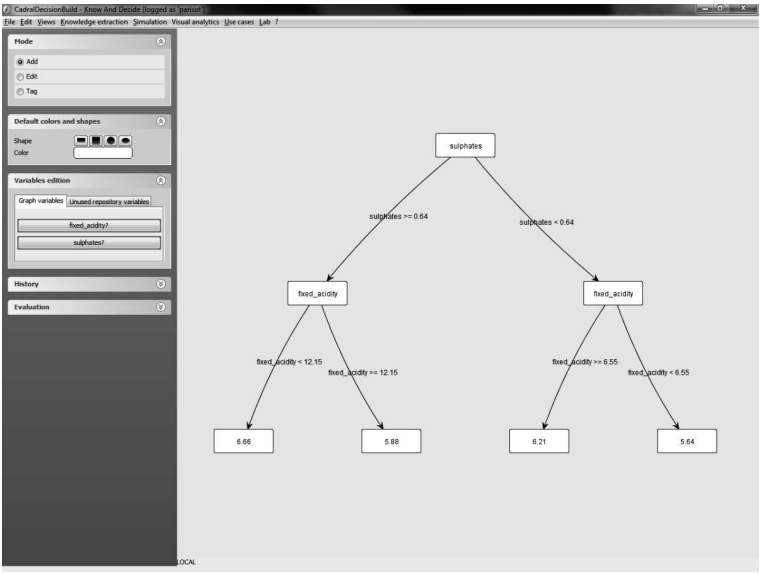

Figure 7: Graphical user interface for the edition of the price evaluation model: the structure of the regression tree can be changed and/or the rules can be modified by the expert.

6 Conclusion

In this paper, a software architecture has been proposed to optimize the online sales: the OPTOSA platform aims at integrating both clever methods for data processing and visualization techniques in order to provide an efficient and useful decision support.

Drawing from these promising results, the innovative approach of OPTOSA will be carried on and refined to efficiently support the work of both manufacture companies and distribution channels. The monitoring of online sales include many time consuming tasks for the human operator that could be automated by enriching OPTOSA with business knowledge. The user will thus be free to concentrate on crucial parts of the process through relevant interaction with the system. Thanks to our operational partnership, further steps towards such integrated tool will continue to be regularly assessed on live data from market places.

7 Acknowledgments

The project OPTOSA is supported by a grant from the Ministry of Economy and External Trade, Grand-Duchy of Luxembourg, under the RDI Law.

References

[Adv] Channel Advisor. `http://www.channeladvisor.com`.

[AR13] C.C. Aggarwal and C.K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC D.M. and K.D. Series. Taylor & Francis, 2013.

[B+84] Leo Breiman et al. *Classification and Regression Trees*. Chapman & Hall, 1984.

[BBD08] Justin Basilico, Zachary Benz, and Kevin R Dixon. The cognitive foundry: A flexible platform for intelligent agent modeling. In *Proceedings of the 2008 Behavior Representation in Modeling and Simulation (BRIMS) Conference*, 2008.

[BL13] K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.

[Com] Apache Commons. `http://commons.apache.org/`.

[CTTB06] Laurent Candillier, Isabelle Tellier, Fabien Torre, and Olivier Bousquet. Cascade evaluation of clustering algorithms. In *Machine Learning: ECML 2006*, pages 574–581. Springer, 2006.

[Dom12] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[ET98] Robert Engels and Christiane Theusinger. Using a Data Metric for Preprocessing Advice for Data Mining Applications. In *ECAI*, pages 430–434, 1998.

[HA08] Jeffrey Heer and Maneesh Agrawala. Design considerations for collaborative visual analytics. *Information visualization*, 7(1):49–62, 2008.

[ice] icecat. `http://icecat.lu/`.

[Jai10] Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[KAF+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[KHP+11] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Inf. Vis.*, 10(4):271–288, 2011.

[Kot13] S.B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283, 2013.

[KPHH11] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on H.F.C.S.*, CHI '11, pages 3363–3372, NY, USA, 2011.

[KW78] Joseph B Kruskal and Myron Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[Mur98] Sreerama K. Murthy. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Min. Knowl. Discov.*, 2(4):345–389, 1998.

[OFWB03] J. O'Madadhain, D. Fisher, S. White, and Y. Boey. The JUNG (Java Universal Network/Graph) Framework. Technical report, UCI-ICS, October 2003.

[PBDT13] Olivier Parisot, Pierrick Bruneau, Yoanne Didry, and Thomas Tamisier. User-Driven Data Preprocessing for Decision Support. In Yuhua Luo, editor, *CDVE*, volume 8091 of *LNCS*, pages 81–84. Springer Berlin Heidelberg, 2013.

[PDBT14] Olivier Parisot, Yoanne Didry, Pierrick Bruneau, and Thomas Tamisier. Data visualization using decision trees and clustering. In *5th International Conference on Information Visualization Theory and Applications (IVAPP 2014), Lisbon, Portugal*, 2014.

[Qui92] John R Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian joint Conference on A.I.*, volume 92, pages 343–348. Singapore, 1992.

[Sis] Sistrix. `http://www.sistrix.com`.

[TPD+11] Thomas Tamisier, Olivier Parisot, Yoann Didry, Jérôme Wax, and Fernand Feltz. Adapting decision support to business requirements through data interpretation. In *Cooperative Design, Visualization, and Engineering*, pages 82–85. Springer, 2011.

[WFH11] Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

[Wis] Wiseseo. `http://www.wiseseo.de`.

[YL03] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.