Enhancing Preprocessing in Data-Intensive Domains using Online-Analytical Processing

Alexander Maedche[1], Andreas Hotho[1], and Markus Wiese[2]

[1] Institute AIFB, Karlsruhe University, D-76128 Karlsruhe, Germany
{maedche, hotho}@aifb.uni-karlsruhe.de,
http://www.aifb.uni-karlsruhe.de/
[2] Deutsche Telekom AG , D-76646 Bruchsal, Germany
markus.wiese@telekom.de,
http://www.telekom.de/

Abstract The application of data mining algorithms needs a goal-oriented preprocessing of the data. In practical applications the preprocessing task is very time consuming and has an important influence on the quality of the generated models. In this paper we describe a new approach for data preprocessing. Combining database technology with classical data mining systems using an OLAP engine as interface we outline an architecture for OLAP-based preprocessing that enables interactive and iterative processing of data. This high level of interaction between human and database system enables efficient understanding and preparing of data for building scalable data mining applications. Our case study taken from the data-intensive telecommunication domain applies the proposed methodology for deriving user communication profiles. These user profiles are given as input to data mining algorithms for clustering customers with similar behavior.

1 Introduction

The telecommunication industry is faced right now especially in Germany with a growing competition. "Knowing the customers" is one of the most important steps towards customer-specific pricing and offering special tariffs. Data Mining is a very promising technology to tackle the Customer Relationship Management (CRM) task that becomes more and more important in all industries.

The telecommunications domain is very data-intensive. The immense amount of data makes classical desktop data mining systems hardly applicable. Especially the tasks of data integration and preprocessing are hard to handle. The preprocessing task that is very time consuming in real data mining projects gets more and more difficult.

Recent work emphasize the promising direction of integrating database technology and data mining (see e.g. [7], [9]). We propose merging the scalable database technology using a data mart with classical data mining systems and algorithms. In our approach the interface for this combination consists of an OLAP (online analytical processing) engine that allows a high level of interaction between human and database system. As far as possible the most preprocessing operations are processed in the database system. Additionally, the consistent way of performing preprocessing in a database is scalable and can easily be revoked and reused.

The paper is organized as follows. In section 2 we draft our analysis scenario. The analysis scenario describes our real-world application of handling the complex, high

dimensional and enormous amount of communication data. We further explain in section 3 the general core of our approach called OLAP-based preprocessing. Taken from the telecommunication domain we present a case study using OLAP-based preprocessing. In section 4 we outline how communication profiles are generated using call detail records and show how these profiles are used for customer segmentation. Section 5 gives an short overview over similar works and section 6 concludes with an outlook on further work.

2 A Telecommunication Analysis Scenario

Telecommunication companies make their business by selling communication minutes to their customers. Considering these minutes as the main product different properties must be distinguished that determine the variation of the product. These properties or dimensions include, for example, the daytime, weekday or weekend, the calling distance etc. Calculating the different possible combinations of attribute values for just these three dimensions gives a rather large number. For example, taking every hour of a day as value for the dimension daytime, distinguishing just between weekday and weekend and considering 10 different calling distances (e.g. city, international, ...) gives 24 x 2 x 10 = 480 different combinations called communication features. The Deutsche Telekom distinguishes internally even many more than just 10 calling distances.

From a customer care point of view the Deutsche Telekom aims at identifying groups of customers which typically buy the same or similar products. Understanding the needs of the customers special offers can be made and the tariffs can be adjusted to the customer behaviour. Realizing this difficult task using data mining we build in a first step a communication profile for each customer. A communication profile should represent the typical calling behavior of a customer over a rather long period of time. Having analyzed the domain and the data given in form of call detail records two difficulties arose. First, a customer buys only one product at a time which is contrary to the classical market basket analysis. Second, the large number of single transactions (call detail records) must be aggregated to a more meaningful level.

In our scenario call detail records of around 4500 private customers have been stored in a panel for 5 years. In the average a private customer has between 3 and 7 phone calls a day. Thus, approximately 12.000.000 call detail records were produced and stored in flat files. In addition to this communication data social-demographic data for these private customers was collected.

After having integrated the communication as well as the social-demographic data in one data mart we started to load the data from the data mart into a data mining system. First, we intended to perform all necessary preprocessing within the data mining system. This approach, however had the heavy drawback that all time-consuming preprocessing steps had to be executed again and again using new data. Additionally, this way of preprocessing within the data mining tool was not scalable, reusable and flexible as necessary for our complex and data-intensive domain.

3 Our architecture for OLAP-based preprocessing

Practical experiences in the development of data mining applications in general have shown, that the sub-phases data connection and integration, data understanding and

preprocessing take the majority of the whole process time. Whereas preprocessing plays a minor role in scientific research it is of highest relevance in real-world applications. The pure application of data mining algorithms in the development process requires only a petty length of time. However, the quality of the generated models is heavily depending on the algorithm specific data preprocessing.

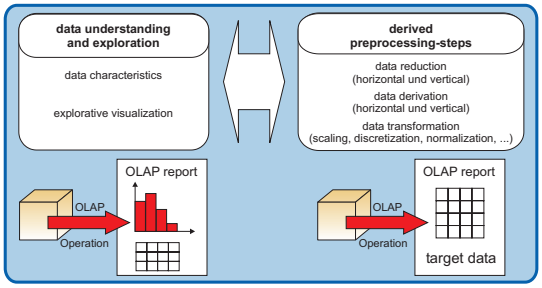

Figure1. Procedure of preprocessing data, in the lower part extend with OLAP functionality

In the upper part of figure 1 the typical procedure of preprocessing the data is shown. Preprocessing is performed in interaction with understanding the data. To understand the data different methods can be applied: One way is to perform multivariate statistical methods to calculate some data characteristics. An other way is the exploratory visualization of the data. After having gained some ideas and hints about the data goal-oriented preprocessing operations can be performed. We distinguish between three groups of preprocessing operations: Data reduction, data derivation and data transformation. The process of data reduction is partitioned into horizontal and vertical reduction. Likewise, data derivation can also be performed in a horizontal and vertical manner. Horizontal derivation adds new individuals to the data set (also called balancing) and vertical derivation means the generation of new attributes by combining existing ones. Data transformation includes operations like applying mathematical functions, discretization or categorization, normalization and replacement of values.

Handling very large amounts of data, the iterative and interactive task of data understanding and preprocessing becomes more and more time consuming. In the following we will show how the preprocessing task described earlier can be performed more efficiently on large amounts of data. The lower part of figure 1 shows the idea of our OLAP-preprocessing framework. We propose using an OLAP-engine to have a flexible way for understanding and preprocessing large datasets. Generating reports with tables and graphs by applying OLAP operations (see lower left part of figure 1) is usually used to obtain valuable information for business decisions. In our case we do not only want to apply OLAP for data understanding or visualization, but also to create a target data set that can be used to generate new hypotheses by applying a data mining algorithm.

The idea described above was realized by the architecture depicted in figure 2. We consider the legacy data to be integrated in a data warehouse or data mart as a necessary prerequisite for efficient data analysis. Therefore, the data integration process is performed only once in our framework and all continuing steps are performed on the data warehouse or data mart. On the one hand side a hyper-cube can be modeled on top of the data mart, on the other hand a knowledge discovery process can be started based on the data in the data mart. An OLAP-engine is used in the modeled hyper-cube.

OLAP is typically performed for the user-driven validation of hypotheses. OLAP functionalities include drill-down, roll-up, slice and dice, pivoting operations for flexible handling and transforming the data (see [1]). The knowledge discovery process consists of various phases that have to be passed (cf. CRISP-DM standard process model for knowledge discovery [3]). The phases typically enclose the tasks of data selection, preprocessing, model generation and interpretation like it is shown in figure 2. Data mining algorithms are executed on the target data set to generate models which have to be interpreted concerning the business relevant questions.

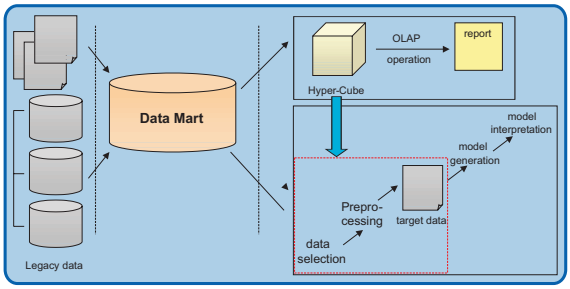

Figure2. Architecture for OLAP-based preprocessing

A target data is derived from applying one of the preprocessing operations described earlier in this section. As indicated in figure 2 by the arrow from the hyper cube to the data preparation part of the knowledge discovery process the data created using the OLAP functionalities on top of the hyper-cube can also serve as preprocessed target data for data mining algorithms. The OLAP-preprocessed data is loaded into the data mining system where further preprocessing can be performed if necessary.

4 A case study

As mentioned in section 2 the Deutsche Telekom uses a panel to analyze the communication behavior of its customers. For 5 years every phone call of around 4500 private customers has been logged with their aggreement. The information stored in anonymous form includes the number of calls per day, the duration of each call, the calling distance, the daytime, weekday etc. Additionally, a well-known market research institute was charged with inquiries at these private customers to get social-demographic information in order to better classify and describe different customer groups. In the present time of a very competitive and unsteady telecommunication market in Germany the Deutsche Telekom tries to extract valuable information from these data sources for goal-oriented marketing actions and new pricing activities. To solve the desired requirements for a quick and efficient analysis of the different information sources it was decided to develop a data mart called PAS (for **p**anel **a**nalysis **s**ystem) that contains all relevant information. Data mining methods should be applied to the PAS such that the data mart serves as a decision support system for marketing and pricing actions.

In a case study using the PAS data mart it was the goal to describe each panel customer by a communication profile and build groups of customers that have similar profiles. A communication profile consists of typical characteristics (e.g. a similar number of international calls at certain times on a weekday) of the customer. Generating the

desired communication profiles was realized using the idea of OLAP-preprocessing that was introduced in section 3.

4.1 Application of OLAP-Preprocessing

Our approach presented in section 3 was applied to the case study described in the last section. To derive the target data we performed the most preprocessing steps by applying various OLAP functions. Because of the complexity indicated in section 2 we restricted the communication data to 3 month for building a customer profile.

In a first approach, the calling minutes of each customer were aggregated over the hour in which the call started, every day of the week and the calling distance. However, this approach was cancelled since the number of communication features (over 1000!) was too complex to handle. An exploratory analysis revealed that the data was too detailed. In a second approach, the calling minutes were summed up over 6 hour slots starting from 0am to 6am, 6am to 12pm and so on. Furthermore, we distinguished just between weekdays (Monday thru Friday) and weekend (Saturday and Sunday). The values for the dimension calling distance was also cut to three. As a result we got the number of 24 communication features (4 x 2 x 3 = 24) used to describe the communication behavior of a single customer. On the left side of figure 3 an average of all panel customers with the chosen 24 communication features is depicted. The middle of the x-axis separates weekdays from weekends, while the first 12 communication features represent the weekdays section. Weekdays are further separated in three areas depending on the calling distance (city, regional, national). Furthermore every calling distance is subdivided into the 4 time windows mentioned earlier. The y-axis represents the average communication minutes in dependence of the communication feature.

The desired aggregated information was returned by an OLAP tool in form of a single table that contained all communication transactions of the panel customers within 3 months. In a next OLAP preprocessing step the “pivot” functionality transformed the column representation of a single customer to the more appropriate form of a row representation. A single row consists of a customer id and the calling minutes aggregated for each of the 24 communication features. Using simple visualization techniques to view the distribution of values for each communication feature it became evident that all communication features are more or less left-slanted distributed. As a prerequisite for applying the cluster algorithm the data had to be transformed from a left-slanted distribution to a symmetric distribution. This was achieved by transforming the calling minutes of each communication feature by the logarithmic function.

4.2 Clustering and Interpretation

Using the well-prepared and preprocessed data the task of building various customer segments with its characteristic communication behavior could be addressed. The well-known and successfully approved k-means clustering method (see [6]) was selected and applied to the data. The number of clusters k was set to 10. To determine the number k, we executed a hierarchical clustering algorithm on a sample set (cf. [8]). The absolute number of elements in the 10 clusters varies from 109 to 777 private customers. Analyzing the smallest cluster revealed that these private customers were identified as outliers. Their calling minutes were significant above the average communication behavior of

private customers and look more like small enterprises. The remaining nine clusters yielded a concentration of private customers with rather homogeneous communication behavior within each cluster. Indeed, the found clusters built a very good basis for segmenting different groups of private customers. The average profile of each cluster (see left side of figure 3) could be visualized by using so called error diagrams. For each of the 24 communication features the mean value and a 95-% confidence interval is calculated and represented for each cluster. By using this simple visualization technique very interesting differences between the clusters could be detected concerning the usage of special communication features.

On the right hand side of figure 3 a cluster with 777 members is represented. The communication behavior of the cluster members significantly differs from the average communication behavior of all 4500 panel customers. This cluster is very similar in city calls on weekdays and weekends compared to the average of all customers. However, all other communication features are obviously different (see figure 3).

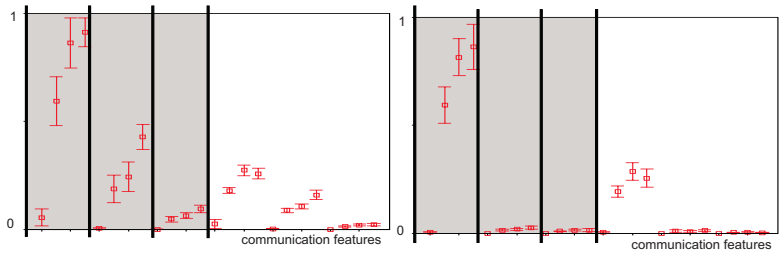

Figure3. Communication profiles on left side of all customers, on right side of one cluster

As mentioned earlier in this section the PAS data mart contains additional social-demographic data derived from the customer inquiries. These include, for example, information concerning the size of the household or the net income of the customer. The demographic data can be accessed by an OLAP tool like the communication data and is added in form of columns to the table described in section 4.1 such that it can be used for further analysis. It is intended to use this information for generating intensional rules that explain better each derived cluster.

5 Related Work

Few methodologies have been proposed for the preprocessing task. In [4] it is described how recommendations for preprocessing steps can be derived for classification tasks that base on the calculation of data characteristics. Calculating data characteristics on complex and very large data sets is very difficult and in some cases impractical and rely on statistical assumptions that do not hold in real world domains.

The relationship between OLAP and data mining is examined by Parsaye (see [9]). He describes an architecture that combines OLAP and data mining applications and shows how data mining applications depend on different aggregation levels. Han's group (see [7]) focuses its research on the area of OLAP mining. One of the motivations for OLAP mining was the perception that different aggregation levels were necessary for their pattern analysis. Both approaches concentrate on data mining algorithms but hardly on other parts of the knowledge discovery process.

6 Conclusion & Further Work

In this paper we presented an approach for efficient and scalable data understanding and preprocessing in real world domains. Using OLAP technologies for preprocessing offers quite a lot of advantages: OLAP applications are usually interactive and can handle very well large data sets.

Our approach presented in this paper aims at simplifying the problem of preprocessing real world data based on an existing OLAP environment. The performed preprocessing steps within the OLAP environment leads to well prepared target data for the following clustering task. That way we were able to derive interesting clusters. The communication profiles derived in the case study are based on the assumption that 24 communication features are suited to describe the customer behavior. Existing works that build user profiles (e.g. [2], [5]) in a similar way make pretty much the same assumptions, but use different aggregation levels. Applying our architecture for OLAP-based preprocessing we are able to reconstruct different aggregation levels in a very efficient and easy way. The determination of the right aggregation level is a nontrivial task. In our further research we intend to investigate the influence of different aggregation levels on clustering algorithms and on other data mining methods.

Acknowledgements The work presented in this paper was partially financed by an internship of the Deutsche Telekom AG.

References

1. S. Chaudhuri and U. Dayal: An Overview of Data Warehousing and OLAP Technology. SIGMOD Record, Volume 26, p. 65-74, 1997.
2. Qiming Chen, Umesh Dayal, Meichun Hsu: OLAP-based Scalable Profiling of Customer Behavior. In Proccedings of the First International Conference on Data Warehousing and Knowledge Discovery, (DaWaK '99), 1999.
3. P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, R. Wirth: The CRISP-DM Process Model, (http://www.crisp-dm.org), 1999.
4. R. Engels and C. Theusinger: Using a Data Metric for Offering Preprocessing Advice in Data Mining Applications. Proceedings of the 13th European Conference on Artificial Intelligence (ECAI 98), Springer, Brighton, p. 430-434, 1998.
5. T. Fawcett and F. Provost: Combining Data Mining and Machine Learning for Effective User Profiling. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996.
6. K. Fukunaga: Introduction to Statistical Pattern Recognition. San Diego, CA, Academic Press, 1990.
7. J. Han: OLAP Mining: An Integration of OLAP with Data Mining. Conference on Data Semantics (DS-7), 1997.
8. L. Kaufman and P.J. Rousseeuw: Finding Groups in Data: An Introduction to Cluster Analysis. New York, Wiley, 1990.
9. K. Parsaye: OLAP and Data Mining: Bridging the Gap. Database Programming and Design, Volume 10, p. 30-37, 1998.