IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4737

Design and Research Implications of Customer Relationship Management on Data Warehousing and CRM Decisions

Colleen Cunningham[1], Il-Yeol Song[2], Jin Taek Jung[3] and Peter Chen[4]
[1,2]Information Science & Technology, Drexel University, Philadelphia, PA, USA
[3]Department of Information Science, Hansung University, Korea
[4]Department of Computer Science, Louisiana State University, Baton Rouge, LA, USA
E-mail: [1]cmc38@drexel.edu, [2]songiy@drexel.edu

ABSTRACT

Customer Relationship Management (CRM) is a strategy that integrates the concepts of Knowledge Management, Data Warehousing, and Data Mining:. This paper addresses the issues related to customer segmentation and design requirements for supporting CRM using data warehouse and data mining technologies. The paper presents a starter multidimensional model for CRM that can address various CRM analyses. The paper also explores some of the areas of research pertaining to data warehousing and data mining as they relate to CRM.

1. INTRODUCTION

According to [5], acquiring new customers can cost five times more than it costs to retain current customers. Furthermore, according to [17], repeat customers can generate more than twice as much gross income as new customers. Companies have come to the realization that instead of treating all customers equally, it is more effective to invest in customers that are valuable or potentially valuable, while limiting their investments in non-valuable customers. In order to manage their customer relationships with customer-specific strategies, companies are turning to Customer Relationship Management (CRM) techniques and CRM-supported technologies.

CRM can be formally defined as a strategy that captures and utilizes organizational knowledge and technology within the context of the organizational structure and culture in order to support and enable proactive and profitable long-term relationships with customers based upon actual customer preferences rather than upon arbitrary general assumptions. At the center of CRM are the knowledge management, data warehouse (DW), data marts and data mining tools. Knowledge management allows us to assemble organizational knowledge and to set up business strategies. The data warehouse and data marts capture knowledge about the customers. Data mining techniques are then used to gain additional insights into customer behaviors and other characteristics that are of interest to the organization.

This paper addresses the issues related to customer segmentation and design requirements for supporting CRM using data warehouse and data mining technologies. The paper presents a starter multidimensional model for CRM that can address various CRM analyses. The paper also explores some of the areas of research pertaining to data warehousing and data mining as it relates to CRM.

The organization of the paper is as follows: the customer segmentation issues are discussed in section 2. Section 3 discusses the e technical requirements for using data warehouse technologies for CRM. Section 4 presents a preliminary starter model; section 5 presents challenges and research areas facing data warehouses, data marts and data mining that are relevant to CRM. Finally, the conclusion is presented in Section 6.

2. CRM RELEVANCE & THE UTILIZATION OF CRM-RELATED ANALYSES

By utilizing a data warehouse and data marts, companies can perform customer profiling, customer segmentation, cross-selling analysis, etc. in order to make decisions about customer-specific strategies. For example, aggregate data marts can help a company to segregate its customers into categories as depicted in Table 1. A company can use an aggregated data mart to determine its customers' historic and future values and to segment its customer base into one of four quadrants: (1) customers that should be eliminated (i.e. they cost more than they generate in revenues), (2) customers with whom the relationship should be re-engineered (i.e. those that have the potential to be valuable, but may require the company's encouragement, cooperation and/or management), (3) customers that the company should engage and (4) customers in which the company should invest. The company could then use the corresponding strategies, as depicted in Table 2, to manage the customer relationships. Tables 1 and 2 are only examples of the type of segmentation that can be performed with a data mart. However, a word of caution should be taken before categorizing a customer into Segment I because that segment can be further segmented into (a) those customers that serve as benchmarks for more valuable customers, (b) those customers that provide the company with ideas for product improvements or efficiency improvements and (c) those that do not have any value to the company.

Table 1: Customer Segments

		Historic Value	
		Low	High
Future Value	High	II. Re-Engineer	IV. Invest
	Low	I. Eliminate	III. Engage

Table 2: Corresponding Segmentation Strategies

		Historic Value	
		Low	High
Future Value	High	Up-sell & cross-sell activities and add value	Treat with priority and preferential
	Low	Reduce costs and increase prices	Engage customer to find new opportunities in order to sustain loyalty

It should be noted that using aggregates would not necessarily allow one to determine the driving factors behind the customer's behavior; however, drilling down to the details can [16]. Moreover, it should be noted that segmentation could be further complicated by the concepts of contemporary and extended households.

Traditionally, most businesses only analyzed customer data for what can be categorized as a traditional household. However, there are actually three types of households: (1) traditional, (2) contemporary and (3) extended household. A traditional household refers to family members that reside at one address. The contemporary household is used to describe the relationship between customers that are family members that reside at different addresses; whereas, the term extended household is used to describe the links between members of a traditional household and a corporation. Figure 1 illustrates the three types of households as described in [1].

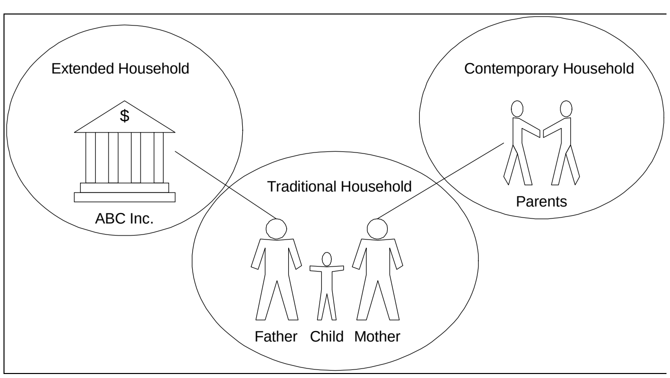
The significance of the three types of households is that companies would like to manage these lines of potential influence such that the company's decision to treat a member of one segment does not have a negative impact on an associated customer in another segment. For example, if a customer is in a non-profitable segment, then the company may decide to increase the customer's price. However, if the company is aware that the same non-profitable customer has influence over another customer (e.g. a parent or a small business) that is in a more profitable segment, then the company may decide to not increase the customer's price rather than to risk loosing both of the customers.

Household analysis, which would include identifying a customer's extended household(s), entails grouping individuals by household or relationship patterns. Some of the benefits of performing Household analyses are:

1. The identification of cross-selling and up-selling opportunities
2. The identification of significant life events
3. The identification of lines of potential influence

Clearly, in order to successfully implement a true CRM strategy, organizations must have a consistent, unified enterprise-wide view of its

Figure 1: Non-Traditional Household [1]



customers. This means that data from disparate systems must be integrated. This requirement for a source of integrated quality data means that the core technology for CRM will be the data warehouse and data mining technologies.

3. IMPLICATIONS OF CRM REQUIREMENTS ON DATA WAREHOUSING

Careful consideration must be given to the customer dimension:. The data in the customer dimension should be in its most atomic form in order to facilitate robust customer analysis and browsing. For example, separating the area code from phone numbers allows us to perform a telemarketing campaign based upon the customer's area code. Similar arguments can be made for decomposing the zip code into the primary zip code and the secondary zip code. Additionally, the data warehouse must be able to group customers into extended households. The value of household analysis is to determine the different lines of potential influence that exist between customers that are in different segments, where one segment may be more profitable than the other segment.

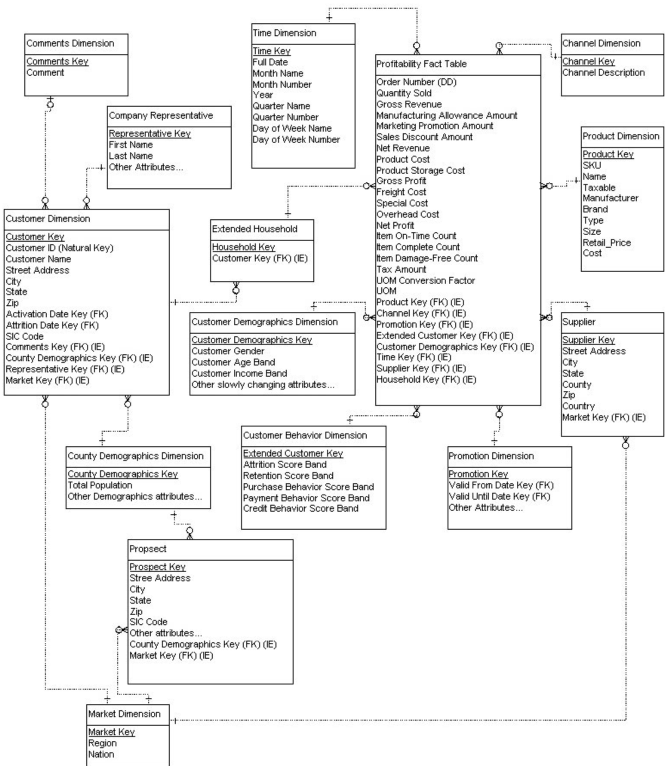
The ability to measure customer retention and the ability to identify root causes of customer defection are key factors for improving customer retention [6]. This means that the design of the data warehouse must provide the company with the ability to track and analyze these key factors. Based upon the customer's profile, the customer can be categorized as a member in one or more segments, where the membership can change over time (i.e. membership mobility). Additionally, CRM analyses involve predicting customer behavior. The prediction is normally represented as a numerical score value that indicates the likelihood that the customer will exhibit a particular behavior. For example, a high customer attrition score means that a customer is likely to leave. Thus, careful design considerations must be given to modeling the customer segmentation to allow for efficient updating and tracking of customer classification. Since the segment to which the customer can belong may change over time, the dynamic mobility of the customer should be modeled using either minidimension technique or outrigger technique [3]. Since most companies will want to track the history of the changes in the customer classification, the best approach (i.e. the most flexible approach) would be to use a combination of the Type 3, 2 and 1 techniques (see [3] for these three methods) for maintaining the dimensions that need to be updated. By taking this approach, companies can know the complete history of the customer's classification as well as its current classification. Additional considerations are the handling of complexity that comes from different data collection channels and different refresh rates into data warehouses.

Table 3: Summarizes Some of Requirements for a Data Warehouse Design to Support Robust CRM Analyses.

Table 3: Design Requirements for CRM DW

1.	Ability to track retention
2.	Ability to identify root causes for customer attrition
3.	Ability to score customers
4.	Ability to associate customers with multiple extended household accounts.
5.	Ability to segment customers into multiple customer segmentations
6.	Ability to maintain the history of customer segments and scores.
7.	Ability to evaluate different campaigns and responses over time
8.	Ability to predict future customer behavior based upon past behavior
9.	Ability to understand loyalty patterns among different relationship groups
10.	Ability to perform demographic analysis
11.	Ability to perform trend analysis
12.	Ability to perform customer profitability analysis
13.	Ability to perform product profitability analysis
14.	Ability to integrate data from multiple sources, including external sources
15.	Ability to efficiently update/maintain data.

Figure 2: Starter Model for CRM DW



4. A PRELIMINARY STARTER MODEL FOR A DATA WAREHOUSE DESIGN

Figure 2 illustrates a preliminary starter model for a data warehouse design to support the CRM requirements listed in Table 3. Table 4 defines the entities represented in the starter model.

The starter model depicted in Figure 2 can be used for a variety of profitability analyses, including customer profitability analysis, household profitability analysis, demographics profitability analysis, product profitability analysis, channel profitability analysis and promotion profitability analysis. For example, the profitability for any transaction in the fact table can be calculated as follows:

(a) Gross Profit = Gross Revenue – Manufacturing Cost – Marketing Cost – Product Storage Cost
(b) Net Profit = Gross Profit – Freight Cost – Special Cost – Overhead Cost
(c) Gross Margin = Gross Profit/Gross Revenue

With the exception of the gross margin measure, all of the measures on both sides of the equations above are stored in the fact table to facilitate the accessibility of accurate measures. While the gross profit and gross revenue are stored in the fact table, the gross margin measure is not explicitly stored because it is non-additive. The profitability analysis can be grouped by customer, product, channel, promotion or any other dimension. The results of the profitability analyses can then be used to segment (or categorize) customers into one of the quadrants discussed in Section 2, which in turn can be used to make strategic (and tactical) decisions about how to manage the customers. Additionally, the model depicted in Figure 2 can be used to calculate key performance indicators (KPIs) for delivery, such as the number of on-time items and the number of damage-free items. These KPIs are important to track and manage because they can influence customer satisfaction and possibly customer retention.

The Time dimension that would be joined to the dimensions that results in the date-related foreign keys in the dimensions have been omitted from the diagram for the sake of clarity in the starter model. The dynamic customer dimensions (i.e. customer demographics dimension and customer behavior dimension) were split into two dimensions because the information would change over time. Additionally, these dimensions were implemented as minidimensions as opposed to outriggers in order to allow the user to readily browse the fact table and to build customer segments based upon the extended customer key in the fact table. In Figure 2, the profitability fact table would be an accumulating snap shot because it captures measures at different points in the process.

Other analyses that can be done using the starter model include being able to determine which products sell well together. This could then be used to define promos by industry and/or by geographical regions. The starter model can also support the ability to identify customers that are in the same industries as some of the existing good customers, but who are not purchasing the same types of products. This could be used to determine what other products customers would potentially purchase were the company to introduce those customers to the products that their counter parts are purchasing that they, however, are not purchasing. The outcome of such an investigation can be used to help companies make decisions about product offerings, marketing strategies and how they interact with their customers.

5. CHALLENGES FOR CRM TECHNOLOGIES

Data Mining and Customer Classification

An interesting challenge for CRM technologies is discovery-driven data mining. According to [9], data mining algorithms can be classified into three categories: (1) math-based methods such as neural networks and linear discriminant analysis, (2) distance-based methods and (3) logic-based methods such as decision trees and rule induction. . One of the advantages of decision trees and rule induction techniques is that they explicitly model the expert's knowledge and reasoning involved in the classification process, and are therefore more easily understood and interpreted by the users. On the other hand, it is more difficult to interpret the model's reasoning process in the neural networks technique. While it may be possible to categorize a customer as a member of a customer classification based upon an activation value greater than 0.5, the rationale for the classification would not be apparent to the user [9]. An additional drawback of the neural network technique is that it tends to lead to a higher proportion of false negatives, where false negatives are defined as the observation of unpredicted (or unexpected) characteristics [9]. Some advantages of the neural network technique are that it tends to have a lower misclassification rate than the linear discriminant analysis technique and it tends to provide relatively consistent predictions across task domains [9]. All of these data mining techniques, however, are inadequate for modeling customer data that may belong to more than one classification.

The preliminary work discussed in [2] looks promising for classifying observations in more than one category. Genetic algorithms (i.e. serial or parallel combination) can be used to produce abstract, rank or measurement output level (see [2] for further explanation) for multiple classifiers. Kim et al. [2] did not, however, consider abstract or rank levels. Therefore, the challenge for multiple classification techniques is to develop algorithms for performing the multiple classification process that would take into consideration the type of output that is desired (i.e. abstract, rank, measurement, or a combination of output types).

Finally, the data mining process is based upon the selection of data mining algorithms, hypotheses formation, model evaluation and refinement. Clearly, this process can be time consuming. Therefore, one challenge for the discovery process would be to find a way to make the process more structured and thus more productive [7].

Aggregation

Another challenge for CRM technologies is performance due to the large amount of data stored in data warehouses. The access time must be efficient. One technique for allowing users to efficiently utilize the stored information is to pre-build aggregations and cubes. This may result in a sparse cube and may lead to the database explosion phenomenon. Pre-calculating all aggregates may actually impede the performance because the database size will be much larger, which could result in additional disk access due to memory limitations. The larger database size will occur because each measurable fact can lead to multiple derived values that must be stored that may never be used in queries. Another serious problem in this case is maintenance overhead. The effect is that the data structure can grow exponentially and be difficult to maintain.

So, when building cubes to support CRM, it is best to: (1) avoid pre-building all aggregates and (2) use multicubes instead of hypercubes. Hypercubes are Cartesian space that allow data values to be entered for every combination of dimensions. They typically only handle a single-fact table. Multicubes on the other hand, can handle multiple fact tables, each with different dimensionality. Multicubes are more efficient for storing very sparse data and thus can reduce the database explosion effect. Another technique for building cubes is to build Iceberg cubes. Iceberg cubes only contain aggregates above a certain threshold. This approach tends to also reduce the amount of sparsity in the resulting cube. Additionally, the use of minidimensions can be used to minimize data sparsity in the fact table. While these approaches can be used to minimize the database explosion effect, there is still a need to develop heuristics to decide which aggregates should be pre-built. Furthermore, although there have been a lot of work on how to build the aggregation and cube for ROLAP, very little work has been done on how to efficiently build the aggregation and cube for MOLAP. The work by Li and Srivastava [4] provided a general algorithm for building the aggregation for MOLAP; however, they did not offer any algorithms for either building the MOLAP cube.

Multiple Classification Technique

According to [5, pg. 158], there is little guiding theory or empirical research that meaningfully addresses issues of how firms can capitalize on knowledge resources and evolve from a transactional focus to a market-driven, relational customer focus. Additionally, while customer segmentation (or customer classification) can be a powerful analysis, there are some limitations on using single classification techniques when the customer may belong to multiple segments (or classifications) [2, 8, 9]. So, another potential research area would be to develop better algorithms that can be used efficiently and effectively to analyze customers that belong to multiple segments.

Schema Segmentation Analysis

Another approach to the challenge of multiple segment membership would be to investigate the appropriateness of one schema to another for segmentation analyses. In other words, is the recommended star schema, as was alluded to by [1] and [6], inappropriate for certain types of analyses? If so, which schemas are best suited for which analyses types? Which schemas (e.g. flat files, etc.) would yield better analysis times, better data mining results and better analysis interpretation? It would be an interesting research project to develop a taxonomy for data warehouse schemas and data mart types to specific types of analyses. The implication for the outcome of such an investigation would be on the rules for constructing data structures to support data mining. Different CRM-related analyses could then be selected and performed using different data mart types (i.e. multicube, cubegrades, etc.) to test the heuristics for building the cubes and the appropriateness of the selected data cube type.

6. CONCLUSION

CRM is a strategy that integrates the concepts of Knowledge Management, Data Mining and Data Warehousing in order to support the organization's decision-making process to retain long-term and profitable relationships with its customers. We have discussed segmentation of customers and design requirements fro CRM data warehouses. Additionally, we have contributed a robust starter model for a data warehouse design that can support the types of analyses that are included in CRM. Finally, we summarized challenges and research areas for CRM technologies.

REFERENCES

[1] Claudia Imhoff, Lisa Loftis and Jonathan G. Geiger, *Building the Customer-Centric Enterprise*, Wiley Computer Publishing, New York, 2001.

[2] Eunju Kim, Wooju Kim and Yillbung Lee, Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems*, To Be Published.

[3] Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit* (second edition), Wiley Computer Publishing, New York, 2002.

[4] Jianzhong Li and Jaideep Srivastava, Efficient Aggregatin Algorithms for Compressed Data Warehouses. *IEEE Transactions on Knowledge And Data Engineering,* ***14*** (3) (2002), pp. 515 – 529.

[5] Anne P. Massey, Mitzi M. Montoya-Weiss and Kent Holcom, Re-engineering the customer relationship: leveraging knowledge assets at IBM. *Decision Support Systems,* ***3*** (2) (2001), pp. 155 – 170.

[6] Hamid R. Nemati, David M. Steiger, Lakshmi S. Iyer and Richard T. Herschel, Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems,* 33 (2002), pp. 143 – 161.

[7] T. S. Raghu, P. K. Kannan, H. R. Rao and A. B. Whinston, Dynamic profiling of consumers for customized offerings over the Internet: a model and analysis. *Decision Support Systems,* ***32*** (2) (2001), pp. 117 – 134.

[8] Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan and Michael E. Welge, Knowledge management and data mining for marketing. *Decision Support Systems, 31* (1) (2001), pp. 127 - 137.

[9] William E. Spangler, Jerrold H. May and Luis G. Vargas, Choosing data-mining methods for multiple classification: Representational and performance measurement implications for decision support. *Journal of Management Information Systems,* ***16*** (1) (1999), pp. 37 – 62.

Related Content