Seán Kelly

is CEO of Comhra, a software company specialising in applications for customer dialogue interaction and customer segmentation solutions. Seán was previously the CEO and founder of the Data Warehouse Network. He authored the best-selling Data Warehousing — The Route to Mass Customization, subsequently published Data Warehousing in Action in 1997 and is currently finalising The Information Wars — a book for business leaders which draws on his considerable experience of bringing organisations into the age of information exploitation.

Keywords: data mining, customer segmentation, CRM, mass customisation, clustering, direct marketing

Seán Kelly
Comhra Ltd
PO Box 7
Skibbereen
Co Cork
Ireland
Tel: +353 2838483
Fax: +353 2838485
E-mail: Sean.Kelly@comhra.com

Mining data to discover customer segments

Seán Kelly

Received (in revised form): 30 October 2002

Abstract

Increasingly companies realise that consumers differ in their needs, preferences, sensitivities, opinions and behaviours. The transition from mass marketing to target marketing that is currently in progress creates demands for more sophisticated customer segmentation techniques. The key enabler of any segmentation strategy is customer data. Customer data are the raw material that must be captured, integrated and effectively analysed in order to achieve the goal of profiling customers. Effective segmentation requires both static customer profile data (eg demographic, lifestyle, preferences) and customer behaviour data (eg usage, loyalty, profitability). Once these data are integrated they can be interrogated to discover groups of customers sharing the same characteristics and needs. Market segmentation is the process of partitioning the heterogeneous market into separate and distinct homogeneous segments. A segment consists of a group of consumers who react in a similar way to a given set of marketing stimuli. Traditionally, the enterprise would define a segmentation matrix and then, based on the data, would allocate customers to segments. This a priori approach to segmentation defines, in advance, a framework or system that describes characteristics of customers or prospects based on information that is known about those individuals. What data mining makes possible is a different approach to segmentation — namely cluster segmentation. The cluster segmentation approach, in direct contrast to the a priori method, seeks to discover naturally occurring clusters of customers who share common characteristics or behave in the same way.

Introduction

Businesses have become increasingly sophisticated in their efforts to capture consumer information, but the process of exploiting consumer information remains relatively immature. Data mining is a process that applies the techniques of artificial intelligence to the task of discovering useful patterns in data, and is proving particularly powerful in the identification of customers sharing the same characteristics. This segmentation of customers into affinity clusters presents new possibilities for customer segmentation, and it is this particular application of data-mining techniques that is the subject of this paper.

Data mining is a term that describes the automatic, or semi-automatic, analysis of large datasets in order to discover meaningful patterns and rules. Data mining differs from conventional data analysis techniques in

Role of reporting tools

so far as it is the software, rather than an individual knowledge worker, which seeks to discover patterns. The human business expert is not made redundant by data-mining tools, as there will always be a place for conventional query tools in the hands of business experts following a definite line of enquiry. In circumstances where you know exactly what you are looking for, the conventional query and reporting tools will be perfectly satisfactory. But in some cases we do not know exactly what we are looking for, and it is in such cases that data mining offers a solution.

Routine business reports fulfil an important role in business management and these are generally referred to as demand-driven. The generation of *ad hoc* queries by business knowledge workers following a hunch is described as hypothesis-driven. The use of data-mining tools to uncover hidden patterns is described as discovery-driven. These are separate and complementary techniques for data exploitation and should not be perceived as competing one with the other. The current upsurge of interest in data mining is explained by the confluence of business drivers (the need to understand customers) with enablers (the technology to achieve that understanding).¹

Acquiring customer data

It is a necessary prerequisite of a data-mining project that the required data to describe the phenomena under observation actually exist. The data required for customer segmentation generally arise from two sources — observation and dialogue. Observation describes the data that are generated through customer transactions and reflect the behaviour of the customer. Dialogue describes the attributes of customers that make up the profile of the customer, and these data are generally acquired directly from the customer.² Ideally the observation and dialogue elements are combined in a data warehouse or marketing database. But even in an integrated database environment it can be difficult to distinguish the wood from the trees. Unfortunately, the nature and evolution of information has meant that the focus on the particular (individual customer applications for ordering, billing, faults, fulfilment, promotions) has prevented a focus on the whole (an integrated perspective of the customer). This is why information systems can readily generate a profusion of detail without always being able to deliver any reliable insight into the overall situation.

The conventional approach: A priori segmentation

Justifying segmentation

Many segmentation approaches have been devised and each of these has merit. Experience suggests that most enterprises use a combination of approaches to deliver maximum benefit. The ability to deliver sophisticated segmentation techniques to the business enhances the ability of the enterprise to deliver more tailored marketing programmes, to identify segments that are more important to the business, to identify segments that have been neglected and to become more attentive to previously unrecognised consumer needs.

Increasingly companies realise that consumers differ in their needs, preferences, sensitivities, opinions and behaviours. The transition from mass marketing to target marketing that is currently in progress creates demands for more sophisticated customer segmentation techniques. The key enabler of any segmentation strategy is customer data. Customer data

Data quality

are the raw material that must be captured, integrated and effectively analysed in order to achieve the goal of profiling customers. Before customer data can be integrated they must first be assessed for quality. Inconsistencies in semantics (what the data mean) and the occurrence of null fields are encountered as well as incorrect data. These so-called 'noisy' data must be conditioned and cleansed before proceeding to uncover meaningful patterns.

Defining segments

Once the data are cleansed and integrated they can be interrogated to discover groups of customers sharing the same characteristics and needs. Market segmentation is the process of partitioning the heterogeneous market into separate and distinct homogeneous segments. A segment consists of a group of consumers who react in a similar way to a given set of marketing stimuli. Usually the enterprise defines a segmentation matrix and then, based on the data, would allocate customers to segments. This a priori approach to segmentation defines, in advance, a framework or system that describes characteristics of customers or prospects based on information that is known about those individuals. Some common a priori approaches to looking at segments include loyalty, profitability, sensitivity, usage, demographics, psychographics and attitude. Table 1 provides an overview of the common a priori approaches to segmenting customers.

Discovering segments

In addition to the a priori approach, data-mining techniques make possible a different approach to segmentation — namely *cluster segmentation*. The cluster segmentation approach, in direct contrast to the a priori method, seeks to discover naturally occurring clusters of customers who share common characteristics or behave in the same way. Regardless of the segmentation technique used, the starting point is the collection of the data that provide the variables to construct the segments.

An overview of data-mining techniques

Data mining is the term that is now applied to the use of artificial intelligence (AI) in the discovery of valuable data patterns that provide insights into the behaviour of the subject of the mining activity. Usually the subject of data mining is the entity 'customer', which reflects the early use of this software in marketing/sales and customer care applications.

Pattern recognition

Data mining is about finding hidden information — the golden nuggets of information that are disguised in the mass of data and are unlikely to be uncovered by knowledge workers with query tools, regardless of their degree of knowledge, creativity or intuition. Such patterns, particularly ones that are counter-intuitive, are not always capable of being uncovered using standard query tools. The discovery of such patterns using standard query tools becomes even more difficult where there are multiple variables. It is in this type of analysis that data mining has the potential to be an extremely powerful adjunct to conventional tools for data analysis and exploitation. All AI systems share a few characteristics: they are designed to work with datasets that are large, complex or both; to search through them and find relevant data; and to look for patterns and deduce rules. Table 2 identifies the more common data-mining techniques and their applications.³

Table 1: Common a priori customer segmentation categories

Segmentation type	Segment definition
Buyer-readiness segmentation	The division of prospects and customers into groups reflecting the different stages which consumers normally pass through during the purchase process. These usually comprise ignorance, awareness, knowledge, preference and conviction.
Benefits segmentation	Dividing the market into groups according to the different benefits that consumers seek from the product.
Behaviour segmentation	The division of customers into groups based on attitude, usage or response to a product or promotion.
Occasion segmentation	The division of customers into groups which consume a product or service at particular times, in certain situations, in response to particular events or according to seasonal or cyclical times.
Psychographic/lifestyle segmentation	The division of customers into groups based on lifestyle, social behaviour, values, sensitivities and personality characteristics.
Demographic segmentation	The division of customers into different groups based on demographic variables such as age, gender, family size, income, occupation, education, language, religion, race and nationality.
Life-cycle segmentation	The division of customers into different groups that recognise the different needs of consumers at different stages in their life.
Geographic segmentation	The division of customers into different groups based on countries, regions, climate and population density.
Loyalty segmentation	The division of customers into different groups based on different degrees of loyalty to supplier or brand.
Product segmentation	The division of customers into different groups based on levels and type of usage of the product or service.
Profitability segmentation	The division of customers into different groups based on the different levels of value or profitability of the customers.
Interaction segmentation	The division of customers into different groups based on their preferences regarding channels, payment method, promotions and communications.
Satisfaction segmentation	The division of customers into different groups based on their recorded satisfaction levels, complaint history, fault history and upgrade history.

Source: Comhra.com

Table 2: Common data-mining techniques

Data-mining technologies	Use
Neural networks	Prediction
Decision-tree algorithms	Comprehension
Genetic algorithms	Categorisation
Fuzzy logic	Options
k-nearest neighbour	Clustering

Cluster analysis

Customer segmentation is best achieved using clustering algorithms — the best-known and most commonly utilised is the k-nearest neighbour technique. (This approach is also known as k-Means, case-based learning or automatic cluster detection.) k-nearest neighbour refers to records that are close to each other in geometric space where k is equal to the number of neighbouring records investigated. The distance between two points in geometric space is used as the measure of association.⁴

Guiding the analysis

A naturally occurring cluster should have high internal homogeneity (all records in the cluster should be very similar) and high external heterogeneity (the characteristics of records in separate clusters should be very different). Clustering removes any element of human prejudice and does not require any guidance from the knowledge worker as regards the type or number of clusters it is expected to find. Other data-mining techniques require a good deal more guidance from the knowledge worker, and clustering algorithms are an undirected approach to the discovery of affinity that is referred to as ‘unsupervised’ learning. Of course, it is necessary to identify the variables that one wishes to include in the search process and, where necessary, to apply weightings of importance to those variables. Weighting of variables is an acknowledgment that we care about some variables more than others. In addition we need to be mindful of the quality of the data that fuel the analysis. The presence of large numbers of null fields or incorrectly coded fields will give rise to the discovery of clusters that are functionally useless.

Multivariate cluster analysis: The automatic detection of segments

Complexity in data

Clustering allows us to look at the big picture. ‘When human beings try to make sense of complex questions, our natural tendency is to break the subject down into smaller pieces, each of which can be explained more simply.’⁵ The problem that then arises is that we discover interesting patterns in the subsets of data that we have arbitrarily partitioned, while the really interesting patterns may be marbled through the different subsets. This essentially is the problem with relying on a priori segmentation alone.

Multiple variables

SRI International in the USA has developed the popular VALS system (values and lifestyle segmentation) and combined this with demographic data to provide a powerful segmentation system describing distinctive consumer behaviours.⁶ Geodemographics segmentation combines location with demographics. Discounted future value (DFV) segmentation combines life cycle with profitability. The discovery of what variables to combine to provide the best intelligence concerning the affinity behaviours that are actually exhibited in the market is the domain of data mining. Increasingly complex multivariate segmentation requires sophisticated analysis tools.

Clustering allows us to see the real-world affinity groups that elect themselves through their choices, values and behaviours. These real-world clusters may not be of uniform size, because there is no rule which says that the real world is neatly ordered. When human decision makers

Recognising real world behaviours

Refining information from data

order the world around them they tend to select dimensions that reflect their organisation (by product, channel or account), or the fixed boundaries of geography (by country, county and city), or demographics (age, gender, income) or profitability (high, medium, low). These partitions are neat and tend to result in segments of a roughly similar size. But they are also arbitrary, synthetic and potentially misleading. When multivariate cluster analysis is employed we discover all kinds of clusters of different sizes and shapes that are arrived at by combining different variables (Figure 1). We also discover customers who belong to no cluster and who have needs and behaviours that are inconsistent with the overwhelming majority of customers. This is crucial intelligence for the business.

Conclusion

The technological development of the information economy is happening at an alarming pace, and each influence has been barely analysed before the next wave of innovation sweeps us on. But the tides of innovation have, for the most part, been *physical* as we learn to capture, exchange and store different individual items of information in ever-expanding databases. The result has been an explosion of data, but not necessarily an explosion in information distilled from those data. What we have not been able to do is to keep intellectually abreast of the potentialities of the information resource and truly understand how to exploit it. It has always been the author's conviction that the root cause of this tendency lies in our infatuation with what technology does rather than the necessary curiosity

Figure 1: Cluster analysis sample

Exploiting the information asset

we ought to have concerning what we can most effectively use it for. This is not to suggest that *all* data-mining and AI techniques are sufficiently mature right now, but the application of cluster analysis to the basic problem of market segmentation is compelling.

As businesses amass ever-larger repositories of data, questions are beginning to be asked concerning what alchemy we have extracted from it. If we knew better how to exploit the data that we capture we would be more concerned that we are capturing the *right* data rather than capturing ever-greater volumes of it. It has been a truism of our times that information is power. We need now to consider data as an asset. Like other assets they can create wealth. But they do not do so by their mere existence; they have to be worked to create wealth. And to work them effectively we need to apply more sophisticated tools than we have utilised in the past.

Appendix: Glossary

A priori segmentation: The a priori approach to segmentation defines, in advance, a framework or system that describes characteristics of customers or prospects based on information that is known about those individuals.

Behavioural segmentation: The division of a market into groups based on consumer knowledge, attitude, use of or response to a product.

Cluster: Groups of similar objects that differ significantly from other objects. Clusters, unlike classes, are naturally occurring and are discovered through analysis.

Cluster segmentation: The cluster segmentation approach, in direct contrast to the a priori method, seeks to discover naturally occurring clusters of customers who share common characteristics or behave in the same way.

Data mining: The process of automatically extracting previously unknown and actionable information from large complex databases.

Demography: The study of human populations in terms of size, density, location, age, gender, race, occupation and other reasonably static characteristics.

Market segment: Market segmentation is the process of partitioning the heterogeneous market into separate and distinct homogeneous segments. A segment consists of a group of consumers who react in a similar way to a given set of marketing stimuli.

Mass customisation: A marketing technique whereby products and/or communications are crafted for individuals and executed on a large scale.

Multivariate segmentation: A technique for creating segments by combining two or more variables within a segment category or between different segment categories, eg risk-sensitive older males exhibiting strong brand loyalty and solitary leisure habits.

‘Noisy’ data: Data with missing or invalid values that are not useful or are potentially misleading. The concept of ‘noise’ derives from radio transmission and describes the random disturbance of a signal.

Psychographics: Measures of consumers based on social habits, lifestyle or personality characteristics.

Undifferentiated marketing: A market-coverage strategy whereby a firm decides to ignore the existence of market segments and directs the same marketing message to the entire market.

References

1. Cabena, R., Hadjinian, J., Stadler, P., Verhees, F. and Zanasi, N. O. (1998) *Discovering Data Mining*, Prentice Hall, NJ, p. 5.
2. Boon, C. (2002) 'The Case for an Integrated Customer Dialog and Segmentation System', available at Combra.com.
3. Adriaans, P. and Zantinge, D. (1996) *Data Mining*, Addison Wesley, Harlow, p. 56.
4. Kelly, S. (1996) *Data Warehousing in Action*, Wiley, Chichester, p. 104.
5. Berry, M. J. A. and Linoff, G. (1997) *Data Mining Techniques*, Wiley, Chichester, p. 187.
6. Niles, H. (1981) 'A New Way to View Customers', *Duns Review*, August, pp. 42–46.