THE ANALYSIS OF PRODUCT CATEGORIES AND SALES RELATIONSHIPS AMONG VALUABLE CUSTOMERS THROUGH DATA MINING AND ITS APPLICATION TO A NATIONAL RETAILER THROUGH ASSOCIATION RULES AND CLUSTER ANALYSIS

Prof. Dr. Mahmut Tekin[1], Yunus Köse[2] and Özdal Koyuncuoğlu[3], Ertuğrul Tekin[4]

Selçuk University, Turkey
[1]mahtekin@selcuk.edu.tr

Necmettin Erbakan University, Turkey
[3]okoyuncuoglu@konya.edu.tr

The Institute for Graduate Studies in Social Sciences
Selçuk University, Turkey
[2]konya.yunuskose@adese.com.tr
[4]ertugrultekin42@gmail.com

Abstract

Many businesses today have a large amount and various kinds of data related to their customers. Processing this data through various techniques and analyses and making it meaningful and useful for the enterprises will help them make healthier decisions and also increase their competitive capacity in today’s globalizing and increasingly competitive world. The data generated by computer systems seems worthless alone because it does not mean anything to the naked eye. These data make a sense when processed in line with a certain purpose. In this regard, it is important to use techniques that can process large volumes of data. It is the task of data mining to turn raw data into useful knowledge. Data mining is the process of discovering the patterns and trends hidden in data sets. The purpose of this study is to determine the sales relationships between product categories and classes. To this end, data mining methods such as association rules analysis and hierarchical clustering were performed to analyze the receipts collected by a retailing company throughout a year. An overall evaluation was made after the implementation of methods and analyses and some recommendations were offered.

Keywords: Retailer, Data Mining, Database

Introduction

Retailing is a big service sector that brings together customers and the products of national and international manufacturers. By observing the retailing sector of a country, it is possible to make interpretations on its economic boom, competitiveness and standards also on the character and modernity of its social structure. “Retailing”, or in a more scientific approach “retail marketing”, is among the sectors that might play a crucial role in the transformation and development of societies. Retailing, especially organized retailing, has been one of the economic sectors where productivity, efficiency and effectiveness are best achieved. Moreover, retailing sector is highly sensitive to the developments in the economy.

Data mining was first articulated in the early 1990s and is described as a process that starts with the storage of data in large databases and ends with its presentation following analyses. A significant amount, such as 80 percent, of this process involves pre-operations like data storage, organization, normalization and variable selection. The remaining time is allocated to the analysis and presentation of the results. The analysis part of this process refers to the data mining while the whole process is called "knowledge discovery in databases".

With the growth of databases, enquiries and the techniques implemented on the databases like OLAP (Online Analytical Process) have remained inadequate. For that reason, the data on the databases updated frequently and even synchronously are cleaned and arranged on a regular basis and then transferred to high capacity computers called data warehouses that are not updated as often as databases. On data warehouses are performed analyses developed by statistics, machine learning and computer science. Unlike statistics, data mining uses a deductive approach. Hidden and unpredictable information is extracted from the data. Rather than building models through hypothesis, it tries to discover structures called patterns which are not explicit yet predictable within a priori probability.

Recently, data mining has also gained popularity in retailing sector. Looking into the studies, it is seen that methods such as decision tree, classification analysis, logistic regression, the analysis of association rule and cluster analysis are often used in these studies. A literature review indicated that there are no studies extensively examining the relationships between product categories and product classes by analyzing long-term data. Consequently, examining the data from one year's receipts of an enterprise, the present study identified the associative sales relationships between product categories and classes and also the clusters formed by the product classes when sold together.

Methodology

The purpose of this research is to examine, using data mining methods (association rule: Apriori), the sales data (shopping details: receipt no, receipt date, time, products purchased etc.) that a supermarket collected via sales terminals and to reveal the relationships between categories. Thus, it is aimed that the analysis of association rules is performed using data mining methods to find out the rules for associative selling regarding product types and in the light of these rules various types of information are acquired about category management and arrangement-display (Silahtaroğlu, 2008).

At the end of the study, it is intended that stored data is converted into meaningful data sets and these data sets are exploited to boost the sales. In this way, an enterprise is likely to carry out studies using more systematic and thorough analyses.

A store in Konya owned by a Konya-based retailer that operates in nine different cities was included in the scope of the study. The data covers the whole year of 2013. The retailer has a customer loyalty card system which he runs using his own software. TR identity number is essential to update the card info. Any identity numbers entered into the system is instantly checked on ID Check screens of the Turkish Republic General Directorate of Population and Citizenship Affairs and the age and gender information for each customer is verified. This way, the reliability of the data is ensured.

Data Collection

962.000 receipts were collected and examined for the data analysis. First, the total of the receipts were examined and customers using a card were selected from these receipts. Then, it was checked if the information of the card users was up-to-date and those up-to-date ones were selected.

Data from the collected receipts were grouped based on the category management system used by the enterprise. The card-based shopping rate was 62.93% for the store's turnover in 2013. Of this, 5.70% were those whose card information was not up-to-date. In sum, 57.10% of the shopping in 2013 was done by the customers having up-to-date card information. In relation to the number of shoppers, it was seen that a total of 109.000 people did shopping. Of these people, 73.33% were males and 26.67% were females. Data analysis also showed that 78.14% of them were married and 21.86% was made up of the singles.

The table below presents the product category table used by the enterprise. The descriptions in this category table were abbreviated to be used during later data analysis. The abbreviations were provided in this table.

Table 1: Food and Non-Food Group Abbreviations

NON-FOOD GROUPS	**Abb.**
SHOES	AYK
GARDEN AND ANIMAL SUPPLIES	BAH
COMPUTER SUPPLIES	COM
LEATHER	DRI
DURABLE GOODS	DTM
ELECTRICAL EQUİPMENT	ELT
HOME TEXTILE	EVT
CARPET&FURNITURE	HMB
HARDWARE	HRD
READY-TO-WEAR	KNF
BOOK& STATIONERY	KRT
AUTO ACCESSORY	OTO
TOYS	OYN
AUDIO AND VIDEO EQUIPMENT	SGU
SPORTS EQUIPMENT	SPR
GLASSWARE	ZUC
FOOD GROUPS	
FOOD	GDA
Food – Chewing Gum	GDA.CK
Food – Crisps	GDA.CI
Food – Spices	GDA.BA
Food - Flour- Semolina and Starch	GDA.UN
Food – Dried Nuts	GDA.KU
Food - Medicinal Plants	GDA.SB
Food – Dough Additives	GDA.HA
Food – Ready-Made Food	GDA.HG

Food - Baked Foods	GDA.HU
Food – Drinks	GDA.IC
Food – Sweeteners	GDA.TT
Food – Frozen Foods	GDA.DO
Food – Dessert and Sweets	GDA.TA
Food - Biscuits and Chocolate	GDA.BI
Food – Pulses	GDA.BK
Food – Oil	GDA.YG
Food – Canned Food	GDA.KO
Food – Milk and Dairy Products	GDA.ST
Food – Olive	GDA.ZE
Food – Egg	GDA.YU
MEAT PRODUCTS	ETU
BAKERY PRODUCTS	FRN
PAPER	KGT
PERSONAL CARE PRODUCT	KTU
COSMETICS	KZM
GREENROCERY	MNV
CLEANING PRODUCTS	TMZ
RESTAURANT	RES

Modelling

In the modeling stage, the analyses were performed using appropriate models. Apriori and Web models included in the SPSS Clementine software were used for the analysis of association. Hierarchical clustering in the SPSS 22.0 was performed for the cluster analysis.

Thanks to the created models, the results of the analysis were obtained for the association rules between the category types and within the category groups.

Initially, Web Analysis was applied to the data in order to determine the relationships between the categories. Web Analysis is used to show the strength of the relationship between two or more areas. The shape of the lines depends on the strength of the relationship between two areas. Thick lines represent strong relationships. A thick line shows that two areas are strongly related and need to be examined. Moderate relationships are represented by standard lines and weak relationships by interrupted lines. The lack of line between two areas means either this pair does not occur in the same records or the number of occurrences remains below the threshold value. The relationships between the categories were represented visually by the graph generated. The thin lines between the categories indicate a low relationship of associative sales. Yet, Apriori model is needed to determine these relationships more explicitly and accurately and to identify the sales relationships between categories.

Apriori model in Clementine software is used to determine the association rules. Apriori model extracts from a large dataset the rules that contain the highest information content. Support-Generality and Confidence-Accuracy values are computed while discovering the information content. In general literature, the support value demonstrates the probability of the item combinations and the confidence value indicates how many of the records that contain the first item will also contain the second item. In SPSS Clementine program, however, the support value indicates the total frequency of an antecedent in the dataset. The confidence value indicates the probability of both the antecedent and the consequent appearing in the same record and the rule support indicates how frequently these both components occur together. In other words, the term “support” in the literature is used as the rule support here. Definitions by the SPSS Clementine program were taken into consideration when analyzing the tables (Ergün, 2008).

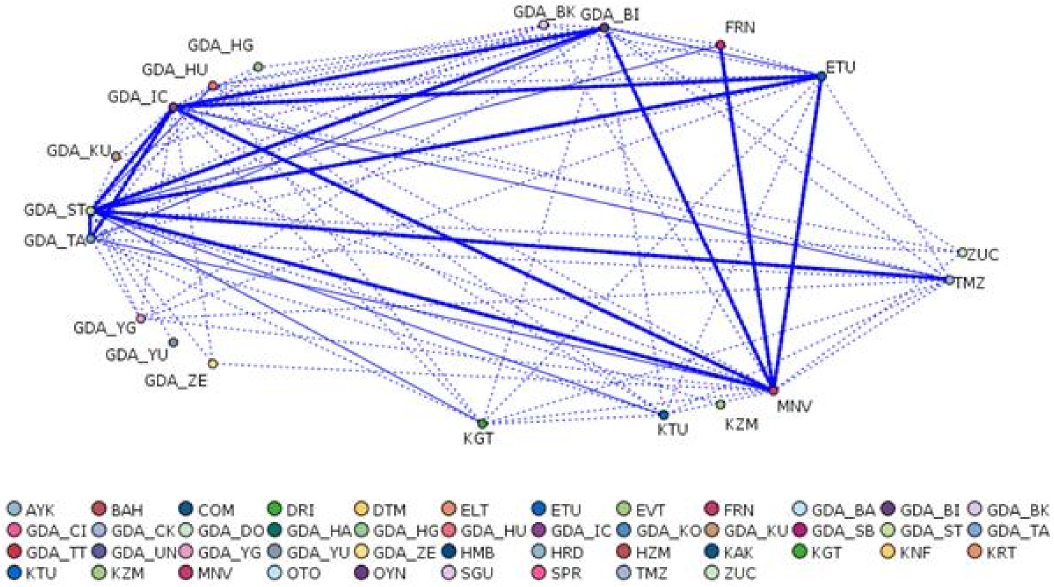
In the analyses, firstly all the categories were examined together and then they were analyzed one by one after obtaining an overall view.

Results

As can be seen in the table below, greengrocery transactions (43.43%) were calculated as having the highest frequency of occurrence among 962.141 transactions, followed by Food-Dairy Products (38.78%)

Table 2: Associative Selling Relationships between Categories

Cluster	**%**	**Cluster**	**%**	**Cluster**	**%**
Greengrocery	43,43	Food – Pulses	12,39	Shoes	3,49
Food –Milk & Dairy Products	38,78	Ready-to-Wear	12,33	Toys	3,32
Food – Drinks	36,01	Cosmetics	12,19	Home Textile	3,15
Food – Biscuits & Chocolate	32,80	Food – Olive	10,98	Hardware	2,32
Meat Products	28,10	Food – Ready-Made Food	10,56	Auto Accessories	2,03
Bakery Products	24,81	Food – Crisps	8,75	Durable Goods	1,71
Food – Desserts and Sweets	21,82	Food – Egg	8,44	Leather	1,30
Cleaning Products	19,63	Food – Chewing Gum	7,50	Personal Accessories	1,06
Personal Care Products	18,57	Food – Sweeteners	7,49	Sports Equipment	0,87
Paper Group	18,08	Food – Dough Additives	7,46	Services	0,77
Food – Baked Foods	16,87	Food – Spices	6,51	Carpet - Furniture	0,60
Glassware	16,56	Food – Canned Food	6,09	Audio Video Furniture	0,51
Food – Dried Nuts	13,21	Electrical Equipment	5,78	Computer Supplies	0,44
Food – Oil	13,15	Food – Flour Semolina	5,75	Garden Animal Supplies.	0,41
Book and Stationery	12,86	Food – Frozen Foods	5,61	Food – Medicinal Plants	0,41



The levels of the relationships in Table 2 are presented in the table below.

Table 3: Levels of Relationships

Connection	Cluster 1	Cluster 2
217,883	Greengrocery	Food – Dairy Products
190,162	Food – Drinks	Food – Dairy Products
177,063	Greengrocery	Food – Drinks
173,728	Food – Biscuits	Food – Drinks
169,384	Food - Biscuits	Food – Dairy Products
166,433	Meat Products	Greengrocery
162,583	Meat Products	Food – Dairy Products
152,932	Greengrocery	Food – Biscuits
134,449	Bakery Products	Greengrocery
131,268	Meat Products	Food – Drinks
125,574	Food – Dairy Products	Food – Desserts and Sweets
122,332	Cleaning Products	Food – Dairy Products
122,237	Food – Drinks	Food – Desserts and Sweets
119,248	Bakery Products	Food – Dairy Products
117,580	Food – Baked Foods	Food – Dairy Products
115,307	Food – Biscuits	Food – Desserts and Sweets
112,348	Meat Products	Food – Biscuits
111,984	Paper Products	Food – Dairy Products

Given the relevant graph and the graph outputs, it is seen that the strongest connection is between Greengrocery and Food-Dairy Products. It is followed by Food-Drinks and Food-Dairy Products.

The results of the analysis of the binary associative selling relationships between the categories show that the greatest rule support is between Food-Dairy Products and Greengrocery Products with a rate of 22.65%. As evidenced by nearly 23% of the receipts, Food-Dairy Products and Greengrocery Products are bought together by the customers who are considered valuable. The difference between the first and second connection is the value of confidence. 58.40% of the customers buying a product from the category of Food-Dairy Products also buy a product from the Greengrocery category. It is evident in the table that there are a lot of categories with a support higher than 10 percent.

When ranked according to the value of confidence, the table items show that 82.43% of the customers buying Food-Eggs buy Food-Dairy Products, as well. This is seen in nearly 7% of the receipts.

Following this stage, the clusters were dealt one by one. Eight clusters with a turnover share of 5% or higher (Meat Products, Food-Dairy Products, Greengrocery, Ready-to-Wear, Cleaning Products, Food-Drinks, Food-Biscuits, Food-Desserts and Sweets) were examined in details.

Initially, the cluster of Meat Products was examined. Greengrocery has the strongest connection with the Meat Products. It is followed by the clusters of Food-Dairy products and Food-Drinks. Taking all these clusters together, it is seen that they constitute the daily needs of customers.

For the cluster of Meat Products, the results of the Apriori Algorithm are as follows: 61.57% of the Meat Products buyers also buy Greengrocery products. This is observed in 17.30% of the receipts. It is followed by Milk and Dairy Products (60.15%), which occurs in 16.90% of the receipts.

Food-Dairy Products are the second cluster analyzed. Greengrocery has the strongest connection with the Food-Dairy Products. It is followed by Food-Milk and Food-Drinks.

For the cluster of Food-Dairy Products, the Apriori Algorithm revealed the following results; 58.40% of the customers buying Food-Dairy Products also buy Greengrocery products. This is observable in 22.65% of the receipts. It is followed by Food-Drink products with a percentage of 50.97%. This appears in 19.77% of the receipts.

The third cluster analyzed is the Greengrocery products. Greengrocery products hold the strongest connection with Food-Dairy Products. This is followed by Food-Drinks and Meat Products.

For the cluster of Greengrocery Products, the following results were obtained from the Apriori Algorithm; 52.15% of the customers buying Greengrocery products also buy Food-Dairy products. This is observable in 22.65% of the receipts. It is followed by Food-Drink products with a percentage of 42.38%. This occurs in 18.41% of the receipts.

Fourthly, the cluster of Ready-to Wear was examined. Ready-to Wear has the strongest connection with Food-Biscuits. It is followed by Food-Drinks and Greengrocery products.

For the cluster of Ready-to Wear Products, the Apriori Algorithm revealed the following results; 39.70% of the customers buying Ready-to Wear products also buy a product from the category of Food-Biscuits. It is seen in 4.90% of the receipts. It is followed by Food-Drink products with a percentage of 39.70%. This occurs in 4.90% of the receipts. Though this cluster has a large share of turnover, its frequency of occurrence in the receipts remains weak.

Cleaning Products are the fifth cluster analyzed and it has the strongest connection with Food-Dairy Products, followed by Food-Drinks and Greengrocery Products.
Regarding the cluster of Cleaning Products, the results of the Apriori Algorithm are as follows; 64.78% of the customers buying Cleaning Products buy Food-Dairy Products, as well. This is observed in 12.72% of the receipts. It is followed by Food-Drink products with a percentage of 55.69% and this occurs in 10.93% of the receipts.

The cluster of Food-Drinks was analyzed in the sixth place. Food-Drink Products has the strongest connection with the Food-Dairy Products and it is followed by Greengrocery Products and Food-Biscuits.
In relation to the cluster of Food-Drinks, the following results were obtained from the Apriori Algorithm; 54.89% of the customers who buy Food-Drink products also buy Food-Dairy products. This is seen in 19.77% of the receipts, followed by Greengrocery products with a percentage of 51.11%. This occurs in 18.41% of the receipts

The cluster of Food-Biscuits was analyzed in the seventh place. Food-Biscuits has the strongest connection with Food-Drinks which is followed by Food-Dairy products and Greengrocery Products
For the cluster of Food-Drinks, the results of the Apriori Algorithm are as follows: 55.06% of the customers buying products from the category of Food-Biscuits also buy Food-Drink products. This is observed in 18.06% of the receipts. It is followed by Food-Dairy Products (53.68%), which occurs in 17.61% of the receipts.

Final analysis was carried out on the cluster of Food-Desserts and Sweets. This group has the strongest connection with Food-Dairy Products and it is followed by Food-Drinks and Food-Biscuits.
Regarding this cluster, the Apriori Algorithm produced the following results: 59.83% of the customers buying Food-Desserts and Sweets also buy Food-Dairy products. This is seen in 13.05% of the receipts. It is followed by Food-Drinks with a percentage of 58.24% and this appears in 12.71% of the receipts.

After the relationships between categories were established, hierarchical cluster analysis was performed using SPSS 22.0 for the purpose of determining the multiple-relationships of the products bought by the customers during their shopping. A total of 43 categories were generated and these categories were associated with each receipt and thus with each shopping record. 4.839.566 shopping transactions were detected in 962.033 receipts.

Greengrocery with a percentage of 8.63% is the most prevalently observed group in the receipts. It is followed by Food-Dairy products (7.71%) and Food-Drinks (7.16%). In non-food products group, this is followed by Glassware and Stationery with a percentage of 3.29% and 2.56% respectively.

The process of clustering was comprised of 44 stages. In the first stage, it is seen that the first cluster was formed by Paper and Cleaning Products. These were accompanied by Personal Care Products in stage 8 and by Meat Products in stage 13. Illustration of a clustering process can be clearly seen in a dendogram. Dendrogram is a graph that displays the strength of relationships by using lines. The dendrogram generated by the cluster analysis is presented below.

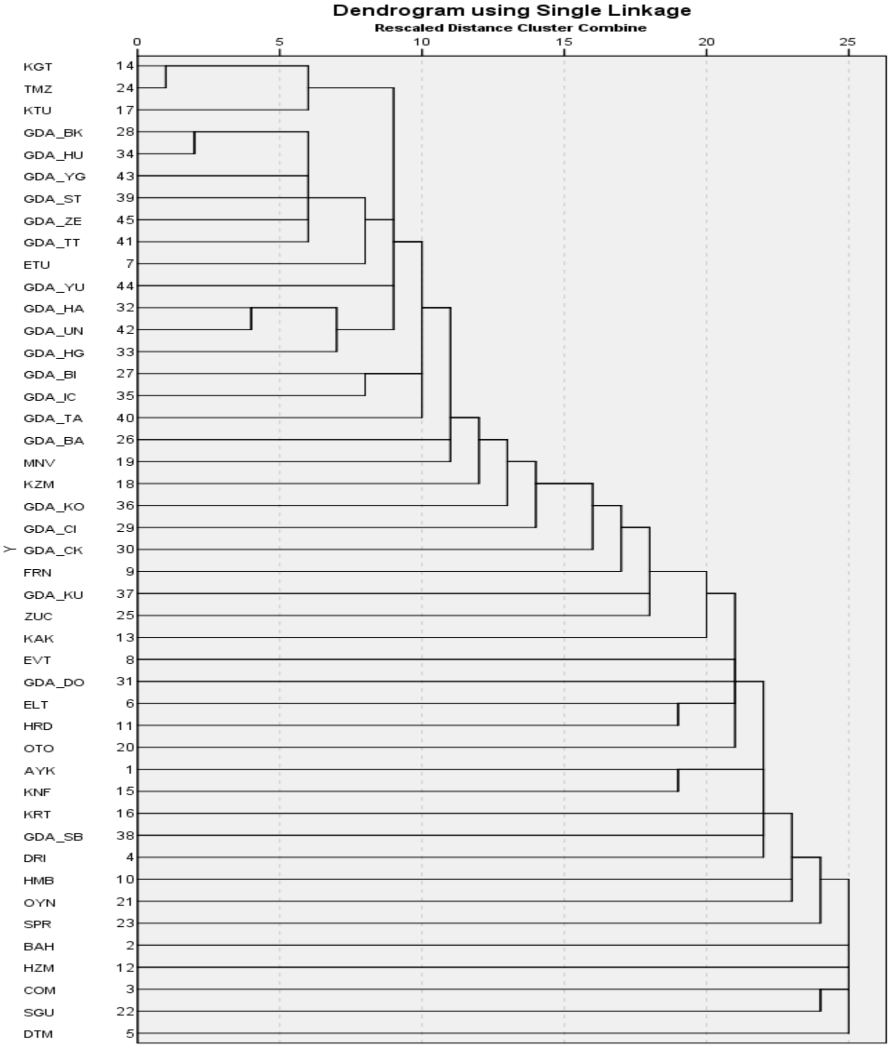


Figure: Dendogram Displaying Relationships Between Product Groups

The figure depicts the relationships between product classes and the clusters they form. The strong relationship between Paper and Cleaning products led to the initial combination of these two classes and thus formation of the first cluster. Personal Care

Products were integrated into this cluster later. It might be suggested that the strength of the relationship within the cluster declines as the linkage point shifts to the right.

Conclusion

The intermediaries in the marketing channel are responsible for transferring or delivering the goods the consumers want to buy based on their needs and demands. Intermediaries act as a bridge between manufacturers and consumers and they, meanwhile, attempt to achieve their own goals in accordance with their modern business sense. Retailers are considered a channel that is responsible for selling goods or services to the consumers who buy from intermediaries. Retailing is a set of business activities in relation to marketing goods and services, without any commercial purpose, directly to the end-consumers to meet their needs (Tek, 1999). Mucuk (1994) defines retailing as a set of business activities associated with marketing goods or services directly to the end-users or consumers.

The results of this study might be used by retailers in offering various promotional campaigns. The product categories and classes used by an enterprise in its category management might be used more efficiently by rearranging them based on the clusters generated through cluster analyses. Also, the patterns discovered might be used to design convincing and guiding activities for indecisive customers. We recommend using the findings of this study for the placement of products on shelves and planograms. By adapting this study to the other stores of a retail company, regional differences could be detected and the future strategies might be developed in the light of this data. The software ERP used by the retail company was developed by the company itself. Necessary software-based modifications could be made in order to get the data from its own systems. This data should be used in activities such as providing leaflets and inserts that are used by the company to reach its customers. The success rate of these activities could be increased thanks to this data.

Certain limitations made it harder to get healthier results in the present study. Though the card infrastructure of the company seems solid, the data beyond the rate of utilization should also be analyzed.

References

Abdullah, F. (2003). Lean Manufacturing Tools and Techniques in the Process Industry with a Focus on Steel. University of Pittsburgh, Pittsburgh.

Ertuğrul, E. (2008). Ürün Kategorileri Arasındaki Satış İlişkisinin Birliktelik Kuralları ve Kümeleme Analizi İle Belirlenmesi ve Perakende Sektöründe Bir Uygulama Afyon Kocatepe Üniversitesi Sosyal Bilimler Enstitüsü İşletme Ana Bilim Dalı Doktora Tezi.

Mucuk, İ. (2004). Pazarlama İlkeleri. İstanbul: Türkmen Kitabevi.

Silahtaroğlu, G. (2008). Veri Madenciliği. Papatya Yayıncılık, İstanbul.

Tek, Ö. B. (1999). Pazarlama İlkeleri: Türkiye Uygulamaları Global Yönetimsel Yaklaşım. Beta Yayımcılık, Ankara.