ANALYSIS OF THE IMPACT OF CLUSTERING ON APRIORI DATA MINING ALGORITHM

(KÜMELEMENİN APRİORİ VERİ MADENCİLİĞİ ALGORİTMASINA ETKİSİNİN İNCELENMESİ)

by

Nergis YILMAZ, B.S

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

INSTITUTE OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

September 2013

ANALYSIS OF THE IMPACT OF CLUSTERING ON APRIORI DATA MINING ALGORITHM

(KÜMELEMENİN APRİORİ VERİ MADENCİLİĞİ ALGORİTMASINA ETKİSİNİN İNCELENMESİ)

by

Nergis YILMAZ, B.S

Thesis

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE

Date of Submission : Sep 13, 2013
Date of Defense Examination : Oct 03, 2013

Supervisor : Assist. Prof. Dr. Gülfem IŞIKLAR ALPTEKİN
Committee Members: Assoc. Prof. Dr. S. Murat EGİ
Assoc. Prof. Dr. Temel ÖNCAN

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere thanks to my advisor Asst. Prof. Dr. Gülfem IŞIKLAR ALPTEKİN for the continuous support of my master study and research, for her patience, motivation, enthusiasm, and immense knowledge.

I would like to thank to my family and my friends for their encouragement and support. Especially my deepest thanks to my dear sister Çiğdem YILMAZ.

Nergis YILMAZ
September, 2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Mobile computing and communication devices are widely utilized by users from different occupations and their usage is steadily increasing. Mobile communication characterizes our current information era. This rapid diffusion has had a direct effect on the applications proposed for smart devices in recent years.

Many organizations collect and store data about their customers, suppliers and business partners. However, much of the useful marketing insights are hidden in that enormous amount of data. Data mining is the process of searching and analyzing data in order to find potentially useful information.

In this study, as a preliminary approach, we have proposed an ontology-based methodology. Although data mining consists of a broad family of computational methods and algorithms, for this study, we have chosen the Apriori algorithm in order to determine association rules. Next, we have aimed at examining the effectiveness of different clustering algorithms when determining the association rules. Hence, we have compared the results of five different approaches. Three clustering algorithms are used: K-means, Expectation Maximization and Hierarchical Clustering. Most predictive association rules with best values are obtained by 'K-means' and 'Hierarchy-based' data mining methodologies. Therefore, we have proposed a new algorithm that combines these two algorithms and we have called it the 'Hierarchical K-Means algorithm'.

The data analysis framework is applied to the data of mobile operating systems' users. By extracting most important information from consumer data, we claim that this framework may direct providers/application developers offer the right product/ advertisement to the right consumer.

RESUME

Appareils informatiques et de communication mobiles sont largement utilisés par les utilisateurs de différents métiers. La communication mobile caractérise notre ère de l'information actuelle. Cette diffusion rapide a eu un effet direct sur les applications proposées pour les appareils intelligents au cours des dernières années. De nombreuses organisations recueillent et stockent les données de leurs clients, leurs fournisseurs et leurs partenaires d'affaires. Cependant, une grande partie des idées de marketing utiles sont cachés dans cette énorme quantité de données.

Dans cette thèse, comme une approche préliminaire, nous avons proposé une méthodologie basée sur l'ontologie. Bien que l'extraction de données se compose d'une grande famille de méthodes et algorithmes de calcul, pour cette étude, nous avons choisi l'algorithme Apriori afin de déterminer les règles d'association. Ensuite, nous avons examiné l'efficacité des différents algorithmes de classification pour déterminer les règles d'association. Par conséquent, nous avons comparé les résultats de cinq approches différentes. Trois algorithmes de classification sont utilisés: K-means, la maximisation de l'espérance et de classification hiérarchique. Règles d'association les plus prédictifs avec les meilleures valeurs sont obtenues par K-means et la méthodologie d'exploration en hiérarchie de données. Par conséquent, nous avons proposé un nouvel algorithme qui combine ces deux algorithmes et nous l'avons appelé le 'K-means hiérarchique'. Le cadre de l'analyse des données est appliqué sur les données des utilisateurs de systèmes d'exploitation de téléphonie mobile. En extrayant information la plus importante à partir des données de consommation, nous prétendons que ce cadre peut ordonner fournisseurs/ développeurs d'applications en offrant le bon produit/publicité pour leurs consommateurs.

ÖZET

Son yıllarda, mobil iletişim, akıllı telefonlar ve tabletler yaygınlaşmalarını gitgide arttırdılar. Birkaç sene öncesine kadar Iphone veya Android cihazları ve bu cihazların uygulamalarına yabancıydık. Mobil tüketici raporuna[1] göre, 2012 yılında Android ve iOS başta olmak üzere, akıllı telefonların popülerliği büyük artış gösterdi. 2011 yılında mobil servisler nüfusun tamamına ulaşacak şekilde gelişmiştir. Avrupa'da, 2011 yılında mobil servis kullanımı bir önceki yıla göre %128 oranında artmıştır (Japonya'da %100 ve Amerika Birleşik Devletleri'nde %104). Bu verilere göre Avrupa'nın yaklaşık 456 milyon nüfusunun 656 milyon aboneliği bulunmaktadır. İstatistikler gösteriyor ki[2], 2012 yılında mobil abone sayısı 741 milyar bireye yükselmiştir. Bu mobil abone sayısı mobil servis endüstrisini, özellikle de uygulama endüstrisini doğrudan etkilemektedir.

Akıllı telefon uygulama pazarındaki hızlı büyüme insanların uygulamalara ulaşım ve uygulama tüketim şeklini değiştirmiştir. Bu değişiklik uygulama pazarları için uygulama geliştirenleri ve telefon üreticilerini de etkileyerek, rekabet dinamikleri üzerinde de değişiklikler yaratmıştır. Gartner araştırma firmasına göre, ücretsiz indirilen uygulamalar 2012 toplam indirilen uygulamaların %89'unu oluşturmaktadır[3]. Konuyla ilgili bir rapora göre, akıllı telefon kullanıcıları Internet'te harcadıkları zamandan daha çok süreyi mobil uygulamalarla geçirmektedir. Haziran 2012'de yapılan bu inceleme, akıllı telefon ve tablet kullanıcılarının günlerinin önemli bir kısmını uygulama kullanarak (özellikle oyun ve sosyal medya uygulamalarını) geçirdiklerini göstermektedir. Bunun sonucunda, ABI araştırmalarına göre, 2016'da toplam global mobil uygulama gelirinin $46 milyara ulaşması beklenmektedir[4]. Bu kapsamda, tüketicinin akıllı telefona eğilimi ve dolayısıyla mobil işletim sistemi pazarı

[1] European Mobile Industry Observatory, www.gsma.com

[2] Global Mobile Statistics 2012 Part A, http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/a#subscribers

[3] Gartner Mobil Application Statistics 2012, http://www.gartner.com/newsroom/id/2153215

[4] ABI Research, https://www.abiresearch.com/

gittikçe önemli bir hale gelmektedir. Mobil işletim sistemi ve uygulamaları pazarı tamamen ticaret üzerine kurulu olduğu için, müşteri segmentasyonu hem tüketici memnuniyetini sağlamak, hem de servis sağlama profillerini arttırmak için önemlidir.

Birçok organizasyon kendi müşterileri, potansiyel müşteriler, üreticiler ve iş ortakları hakkında zengin bir veri tabanına sahiptir, fakat bu veri içinde saklı duran değerli bilgilere erişemezler. Veri madenciliği araçları, şirketleri geniş veriler içinde gizli bilgilerin keşfedilmesi konusunda yönlendirir.

Bu çalışmada, tüketicilerin mobil işletim sistemi tercihleri, veri madenciliği tabanlı bir sistem kullanılarak araştırılmıştır. Mobil işletim sistemi kullanımı ile ilgili olarak bir anket hazırlanmış ve toplam 209 kişiye yaptırılmıştır. Amaç bu veri topluluğu içinden, birliktelik kurallarını bulmak olmuştur. Tezin ilk kısmı, veri madenciliği ve veri madenciliğinde kullanılan kümeleme analizi yöntemleri hakkında teorik bilgiler içermektedir. Başlıca üç farklı algoritma sınıfından seçilen algoritmalar karşılaştırılmıştır. İlk algoritma sınıfı olan model tabanlı algoritmalar arasından *Expected Maximization* (EM) seçilmiştir. 2. algoritma sınıfı olan hiyerarşik kümeleme algoritmaları içinden *yukarıdan aşağı* ve *aşağıdan yukarı* yaklaşımları ve 3. sınıf olan bölümleyici kümeleme algoritmaları arasından ise *K-means* algoritması ayrıntılı ve karşılaştırmalı olarak incelenmiştir. Algoritmalar, WEKA adlı veri madenciliği aracı üzerinde çalıştırılmıştır[5]. Veri madenciliğinde çok kullanılan yöntemlerden biri olan Apriori algoritmasının sonuçlarının iyileştirilebilmesi yönünde araştırmalar yapılmıştır. İyileştirme yöntemlerinden biri olan kümeleme üzerinde durulmuş ve halihazırda akademik yazında bulunan başlıca algoritmalar, tüketicilerden elde edilen veriler üzerinde çalıştırılıp performansları gözlenmiştir. Kümeleme işlemi veri analizinden örüntü oluşturma aşamasında, veri kaynağındaki tüm verileri kullanmak yerine, benzer özellik gösteren verileri temsil eden kümeleri kullanır. Farklı algoritma uygulamaları sonuçlarına göre, Apriori algoritmasından önce, verilere herhangi bir kümeleme algoritmasının uygulanmasının, sonuçların değerlerini arttırdığı görülmüştür. En iyi sonuçların K-Means ve hiyerarşi esaslı algoritmalar ile elde edildiği görülmüştür. K-means algoritmasının zayıf noktası olan, ilk merkezî nokta belirleme işlemi, hiyerarşik

[5] WEKA , http://www.cs.waikato.ac.nz/ml/weka/

algoritmadan elde edilen ilk noktalar kullanılarak bertaraf edilmiştir. Önerilen bu yeni yaklaşımın, bu vakaya ait güven değerlerini K-means algoritmasına göre %70, destek değerlerini ise %60 oranında daha yüksek verdiği gözlemlenmiştir.

1. INTRODUCTION

Mobile communications have received considerable attention in recent years, as well as smart phones and tablets. Couple of years ago, there was not any iPhone, Android or any application. Currently, smart phones become more and more popular and according to the mobile consumer report of 2012, Android and iOS are leading the way[6]. Mobile communication is a key European industry, comparable in size to aerospace and larger than pharmaceuticals, with total revenues amounting to €174 billion in 2010. In 2011, mobile services have achieved a population coverage rate of nearly 100%, and a mobile penetration rate of 128% in Europe (versus 100% in Japan and 104% in the USA). This represents 656 million individual subscriptions held by an estimated 456 million Europeans[7]. The statistics have shown that the mobile subscriptions have increased to 741 million individuals in 2012[8]. These mobile subscribers have directly influenced the mobile services, especially the mobile applications industry.

The rapid growth of the smart phone applications market has fundamentally changed the way in which people access and consume content. This has contributed to a shift in competitive dynamics that impact network operators, operating system (OS)/ application store developers and handset manufacturers. According to Gartner Inc., free applications will account for 89% of total downloads in 2012. Worldwide mobile application store downloads will surpass 45.6 billion in 2012; with paid-for downloads totaling $5 billion. A report has showed that an average smart phone user spends more time in his mobile applications than he does browsing the web[9]. This analysis which has been conducted in June 2011 has showed that smart phone and tablet users now

[6] Courting Today's Mobile Consumer, http://www.nielsen.com/content/dam/ corporate/us/en/reports-downloads/2012-Webinars/courting-the-mobile-consumer-7-18-final.pdf
[7] European Mobile Industry Observatory, www.gsma.com
[8] Global Mobile Statistics 2012 Part A, http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/a#subscribers
[9] Mobile Apps Put the Web in Their Rear-view Mirror, http://blog.flurry.com/bid/63907/Mobile-Apps-Put-the-Web-in-Their-Rear-view-Mirror

spend over an hour and half of their day using applications and games and social networking applications capture the significant majority of consumers' time. As a result of that, in 2016, total global mobile application revenue is expected to reach $46 billion according to ABI Research. In this context, consumer's tendency to smart phone and accordingly to mobile OS market becomes more and more important. Since mobile OS and their applications' marketing is a commercial activity, precise customer segmentation must be investigated in order both to satisfy customer and to increase service provider's profits. This research proposes a data mining based framework to determine consumers' mobile OS preferences. Such knowledge would not only direct firms in the developing process of their own applications, but would also contribute to mobile application marketing.

Many organizations collect and store a wealth data about their customers, potential customers, suppliers and business partners. However, the inability to discover valuable information hidden in the data prevents the organizations from transforming these data into valuable and useful knowledge (Berson 2000; Ngai 2009). Data mining tools can direct these organizations to discover the hidden knowledge in the enormous amount of data.

Ontology represents the concepts and the relationship between them for specialized domain (Noy & McGuinness, 2001). The term ontology in philosophy refers to the theory about the nature of existence, while in computer science, it is a term referring to all the core concepts, including their terms, attributes, values, and relationships that belong to a specified knowledge domain. Building ontology is a complex work and it requires a domain expert to help you declare all domain concepts and the relationship between them. Ontologies, introduced in data mining for the first time in early 2000, can be used in several ways (Nigro et al., 2007): domain and background knowledge ontologies, ontologies for data mining process, or metadata ontologies. Ontology has become increasingly popular, because it promises a shared and common understanding of knowledge domains that can be communicated between people and application systems.

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst them and dissimilar compared to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters and hence. It models data by its clusters.

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurement, or a point in a multidimensional space) into clusters based on similarity. Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. It is important to understand the difference between clustering (*unsupervised classification*) and discriminate analysis (*supervised classification*) in supervised classification; we are provided with a collection of labeled (*preclassified*) patterns; the problem is to label a new pattern. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are data driven; that is they are obtained solely from the data.

In the first part of this study, we investigate the potential contextual relationships between consumers and mobile OS market activities. Doing so, we have used the ontology concept. In order to collect consumer data, we have designed a questionnaire and constructed a relational database accordingly for further consumer data mining.

The second part of the thesis involves the data analysis framework. The framework is based on Apriori data mining algorithm (Alpaydın, 2010). We first apply Apriori algorithm to the consumer data and obtain related association rules. An association rule expresses an association between items or sets of items. We utilize two metrics, *confidence* and *lift*, for evaluating these association rules. As a second approach, we again apply Apriori algorithm, but after clustering consumers. The segmentation of the consumers is generated using three different algorithms: K-means algorithm, Expected Maximization algorithm (EM) and hierarchical algorithm (Alpaydın, 2010). The application is carried out on the WEKA package (Witten et al., 1999). We compare the final results of these three approaches in order to reveal the impact of consumer clustering on the results of the Apriori algorithm. After we applied all approaches, we

proposed our hybrid approach and we show that our algorithm has had 70% of higher confidence values and 60% of higher lift values when compared to K-means algorithm.

2. LITERATURE REVIEW

Data mining, or knowledge discovery, is used for extracting knowledge from collected data and determines meaningful patterns. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. It is possible to use different types of algorithms to extract most important information from a database. Ten most popular data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) are presented in (Wu et al., 2008). They are listed as: C4.5, K-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART. These top ten algorithms are among the most influential data mining algorithms in the research community. Detailed explanation of most of these algorithms can be found in (Alpaydın, 2009).

The knowledge may be classified into two broad classes: Data mining knowledge and domain knowledge (Kuo et al., 2007). Data mining knowledge focuses on data mining algorithms, their usage, parameters tuning, and formats of input data, whereas domain knowledge includes understanding of a dataset, relationships among variables and so on. The framework in this research has touched both of these two classes.

The tools and technologies of data warehousing, data mining, and other customer relationship management (CRM) techniques afford new opportunities for businesses to act on the concepts of relationship marketing (Rygielski et al., 2002). In (Cheng et al., 2012), the authors propose a comprehensive CRM strategy framework by segmenting the customers. They present a procedure to provide different kinds of usage analysis, including inter-cluster analysis and intra-cluster analysis.

Ontologies are nowadays one of the most popular knowledge representation techniques. An ontology captures the domain concepts and their relations; hence, it provides an alternative knowledge source than domain experts. In one of the recent research

(Boufardea & Garofalakis, 2012), the authors propose an ontology-based framework of a distance learning system which improves a recommendation system with rules generated by data mining techniques. This system is used in order to estimate learners' progress and their final grade. The developed system is a new framework using data mining techniques in metadata derived from an ontology. The main objective is to design an ontology that can store knowledge about the learners' skills in relation to a specific educational purpose. In (Kuo et al., 2007), the authors explore the possibility of utilizing a medical domain ontology as a source of domain knowledge to aid in both extracting knowledge and expressing the extracted knowledge in a useful format. They demonstrate that domain ontology driven data mining can obtain more meaningful results than naïve mining.

In this paper, we have used a tool called Protégé, when dealing with domain ontology. Protégé is a free open source ontology editor, which is based on Java. In (Yokome & Arantes, 2011), the authors present the development of a domain ontology on Protégé. The proposed Meta-DM ontology presents a common method for data mining and it can be used with many tools applied to this domain, to help the data miner through discovery of knowledge in databases.

There are lots of applications of data mining in marketing. They all have the objective of satisfying the customers to make more profit from them. In one of the works (Liao et al., 2009), the authors indicate that marketing segmentation must be investigated frequently and it would help to know the market after a specific customer profile, segmentation, or pattern come with marketing activities has found. In the paper, the Apriori algorithm is used to find association rules, and clustering analysis based on an ontology-based data mining approach, for mining customer knowledge from the database. Knowledge which is taken from data mining results is presented as knowledge patterns, rules, and maps in order to offer suggestions and solutions to the case firm. This paper is the most similar framework to the one in our thesis. Another efficient marketing method is direct marketing. In another recent research (Chen et al., 2011), the authors have proposed a novel forecasting method that integrates the union sequential pattern with classification algorithms to facilitate the construction of

customer response models. Based on the use of a union sequential pattern, the potential customer size is established by identifying attributes with a high level of association. Our findings are useful to determine market trend and build up new strategies. This new strategies also could include direct marketing. Apriori algorithm is more proper than classification algorithms to determine customer behavior with our data set.

Hierarchical structure is a very useful and widely adopted technique in information processing. There are lots of types of hybrid clustering algorithms. In (Adigun et al., 2012), the authors combine the two clustering algorithms to come up with a hybrid algorithm for better clustering. The initial clusters' centers are found using K-means algorithm. These centers are widely spread within the data. Then, EM takes these centers as its initial variables and iterates to find the local maxima. Hence, better distributed clusters may be obtained. EM algorithm considers the noise data. Our data set do not have much noise data. Therefore, we did not prefer EM algorithm to set K-means initial centroids. In our case, hierarchical algorithm is more appropriate than EM algorithm.

In another paper, the core of the proposed method is as follows: averaging is used as a sampling method so as to reduce data hierarchically, then clustering is performed on the reduced data, meanwhile, the clustering problem is treated as a weighted clustering one (Lu et al., 2008). Based on the above idea, better initial cluster centers can be found, and the efficiency can be improved by hierarchical clustering due to clustering on the reduced data. This methodology is not appropriate for our data, since it proposes to reduce the data. We already have a small data set so reducing data is an unnecessary step for us. We use all data set to calculate clustering. Our proposed approach's objective is increasing the accuracy of the association rules for small data set.

3. QUESTIONNAIRE

For the case study, a field survey questionnaire approach is applied to collect information from consumers having mobile smart phones. The survey is answered by 209 consumers, where the number of male responders exceeds the number of female responders. The rate of female accounts for 43%, whereas the rate of male accounts for 57%. Most of the respondents, 41.5%, are between 23–32 years old. The questionnaire is given in Appendix.

It has been generated a relational database with questionnaire questions. It consist of seven tables such as, customer, development, device, operating system, applications, most popular application type, application find method. The ER diagram for tables and the relationships are illustrated in Fig. 3.1.

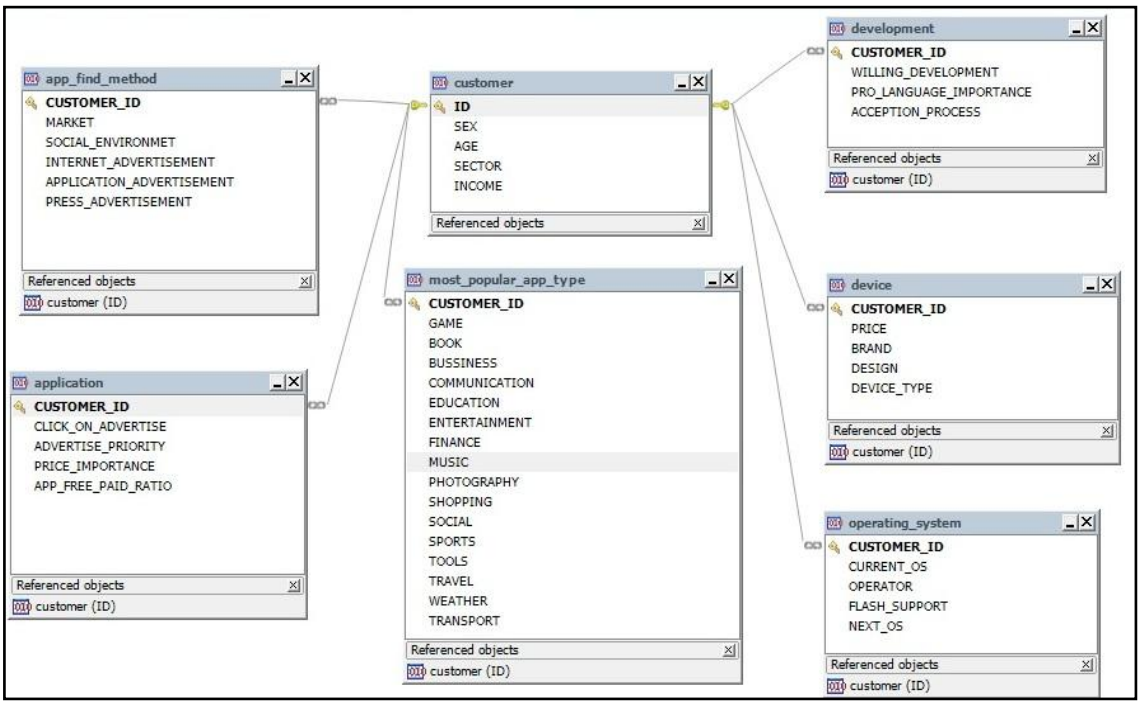

Figure 3.1. E-R Diagram

4. DATA MINING METHODOLOGIES

In this section, the ontology based data mining framework and model results are presented.

4.1. Ontology-Based Data Mining

4.1.1. What is an Ontology?

An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. Ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases. For this reason, ontologies are said to be at the "semantic" level, whereas database schema are models of data at the "logical" or "physical" level. Due to their independence from lower level data models, ontologies are used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services.

Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema. In this study, RDF and OWL have been chosen. An ontology formally represents knowledge as a set of concepts within a domain, and

the relationships among those concepts. Ontology diagram is used to discover consumer behavior pattern on different types of mobile OS (Fig. 4.1).

The OWL ontology design is derived by mapping the fields of the DB tables in classes and it is automatically filled up with about 209 instances. The ontology design consists of two sections: Consumer and mobile OS. The consumer of ontology is classified according to age, income, sector and gender, while the mobile OS of ontology is classified according to software development, device type, most popular application type which are chosen, way of finding applications and the preferences of mobile OS.

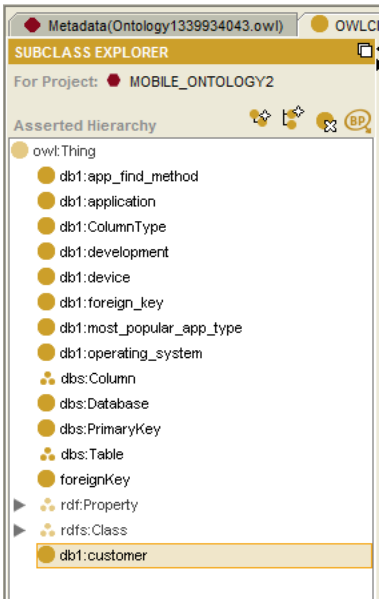

Figure 4.1. Mobile Operating System Ontology

4.1.2. An Ontology Design and Development Tool: Protégé

Protégé is one of the ontology design tools (Noy & McGuinness, 2001). It is an open source ontology editor and a knowledge-based framework. Protégé allows users to construct domain ontologies by entering data and storing them in format such as XML, RDF or OWL. RDF and its extensions such as OWL have been developed to define metadata schemas, domain ontologies and resource descriptions. RDF is a World Wide Web Consortium (W3C) standard developed in 1997. Web ontology language (OWL) built on RDF is the new W3C recommendation for ontology construction with

facilitates for effective reasoning capabilities by consistency checking through inference rules such as transitivity, symmetry, etc. Protégé, which is based on Java, is extensible and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development. An ontology formally represents knowledge as a set of concepts and the relationships among those concepts within a domain. We have used an ontology diagram to discover customer behavior on different types of mobile OSs.

4.1.3. A Data Mining Tool: WEKA

The data mining module uses a well-known data mining algorithms to extract association rules from the given data. The WEKA package (Witten et al., 1999) supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. From the WEKA package, the Apriori algorithm for data mining was used.

4.1.4. Ontology Research Framework

The objective of this preliminary ontology-based research framework is to understand consumer behavior and preferences when using mobile OSs on smart phones (Fig. 4.2). The first step involves developing the domain ontology. Hence, we have observed the trends in mobile OS industry and have interviewed a marketing manager of Android OS during Google I/O. Using this knowledge, we have built the mobile OS domain ontology. This ontology has enabled us to prepare the consumer questionnaire and the relational database. The answers coming from the consumers have been the input of the data mining process. First, the data coming from database is inserted into the Protégé tool. Protégé transforms data to an OWL file. Then, this OWL file has utilized into the WEKA in order to determine related association rules.

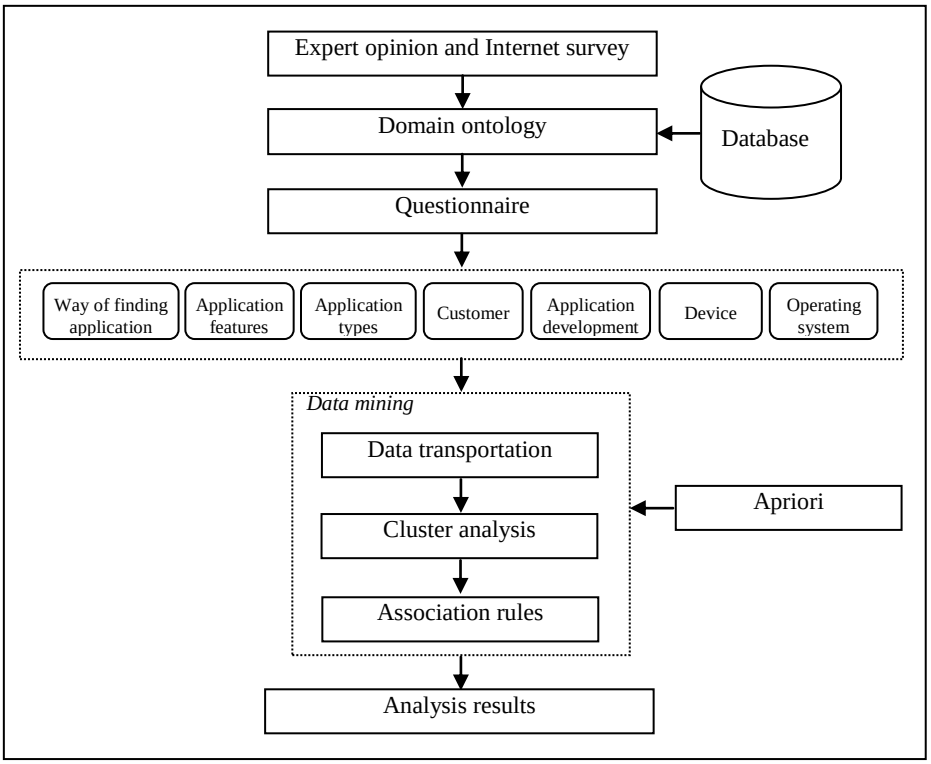

Figure 4.2. Ontology-Based Research Framework

The information flow of the proposed framework is depicted in Fig. 4.3

1. The input receiving from the consumers' results of questionnaire is formatted to be inserted into the database.

2. A database (MySQL), whose E-R diagram is presented in Fig. 3.1, is built. The tables of the database reflect seven main information categories in the questionnaire (Fig. 4.2). The relational database is used as a means of meta-data representation, and accordingly the ontology is created.

3. The OWL file is generated by Protégé. As it is represented using RDF data format, it needs its own RDF-specific query language and facilities.

4. An *arff* file is generated from database and domain ontology, as a means of meta-data. The *arff* file is used in WEKA to build association rules. Doing so, we have used the Apriori algorithm. As a conclusion, we aim at analyzing consumer preferences in terms of mobile OS usage. The knowledge can then be integrated into the decision support process of telecommunication companies' marketing departments.

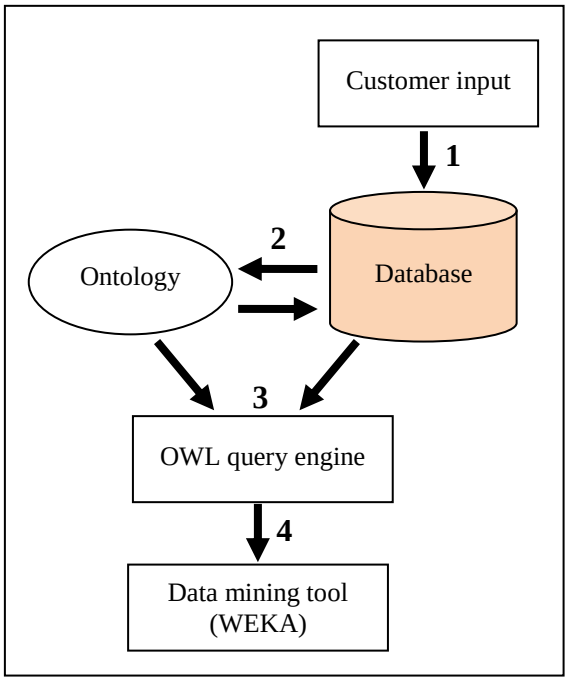

Figure 4.3. System Architecture

4.1.5. Apriori Algorithm to Determine Association Rules

The main objective of the algorithms that determine association rules is to reveal relationships among variables that occur synchronously in large databases. An example of this type of algorithm is the market basket analysis[10]. If people who buy item *X* also buy item *Y* as well, we can state that there exists a relationship between item *X* and item *Y*. An association rule is an implication of the form $X \rightarrow Y$, where *X* is the antecedent and *Y* is the consequent of the rule. The association rules are defined as follows (Alpaydın, 2010):

Confidence is the conditional probability, $P(Y|X)$, which is what we normally calculate. To be able to say that the rule holds with enough confidence, this value should be close to 1 and significantly larger than $P(Y)$, the overall probability of people buying Y. It is interested in maximizing the support of the rule, because even if there is a dependency with a strong confidence value, if the number of such customers is small, the rule is worthless.

[10] Mobile Apps Put the Web in Their Rear-view Mirror, http://blog.flurry.com/bid/63907/Mobile-Apps-Put-the-Web-in-Their-Rear-view-Mirror

Confidence of association rule $X \rightarrow Y$:

$$Confidence\ (X \rightarrow Y) \equiv P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{Customers who bought } X \text{ and } Y\}}{\#\{\text{Customers who bought } X\}} \quad (4.1)$$

Support shows the statistical significance of the rule, whereas confidence shows the strength of the rule. The minimum support and confidence values are set by company, and all rules with higher support and confidence are searched for in the database.

Support of the association rule $X \rightarrow Y$:

$$Support\ (X,Y) = P(X,Y) = \frac{\#\{\text{Customers who bought } X \text{ and } Y\}}{\#\{\text{Customers}\}} \quad (4.2)$$

If X and Y are independent, then it is expected that the lift to be close to 1. If the ratio differs – if $P(Y|X)$ and $P(Y)$ are different- it is expected to be a dependency between two items: If the lift is greater than 1, X makes Y more likely, and if the lift is less than 1, having X makes Y less likely.

Lift, also known as interest of association rule $X \rightarrow Y$:

$$Lift\ (X \rightarrow Y) = \frac{P(X,Y)}{P(X)\,P(Y)} = \frac{P(Y|X)}{P(XY)} \quad (4.3)$$

Apriori algorithm (Agrawal & Srikant, 1994) is interested in finding all such rules having high enough support and confidence, which has two steps:

(1) Finding frequent item sets (those which have enough support), and
(2) Converting them to rules with enough confidence by splitting the items into two, as items in the antecedent and items in the consequent.

4.1.6. Ontology-Based Data Mining Framework Results

At the time the questionnaire is given to consumers, the most common mobile OSs were Android and iOS, which is also visible from Fig. 4.4. According to all consumers' answers, both Android and iOS are more desirable by male consumers.

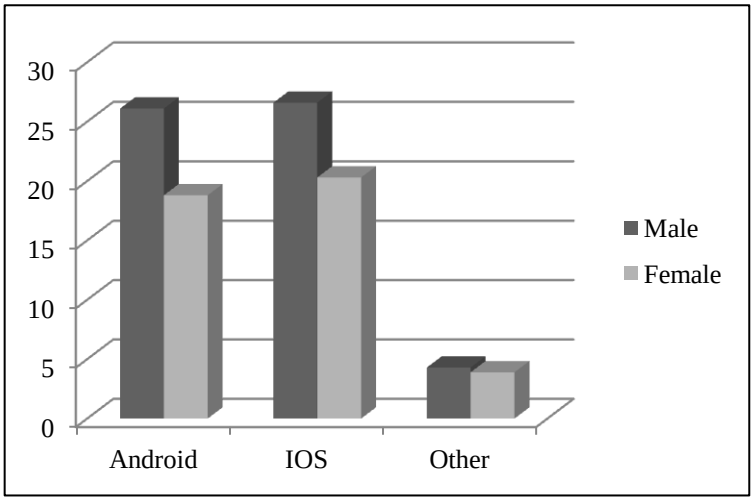

Figure 4.4. Next Desired Operating System

The results in Table 4.1 reveal that Android is found more attractive among male consumers, whereas iOS is more preferable by female consumers. Moreover, Android seems to be more desirable by the young consumers of age 23-32 years.

Table 4.1. Association Rules From Demographic Information (min sup: 20%; min conf: 45%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.51	1.14	NEXT_OS=Android	AGE=23 - 32
R2	0.26	0.46	1.02	NEXT_OS=Android	Male
R3	0.20	0.47	1.01	NEXT_OS=iOS	Female

The minimum threshold of support and confidence is 25% and 60%, respectively. The results in Table 4.2 show that two leaders in mobile OS market are Android and iOS, and their desirability is very close to each other.

Table 4.2. Association Rules Related to Brand (min sup: 20%; min conf: 60%)

Rules	Sup.	Conf.	Lift	Consequent	Antecedent
R1	0.21	0.60	1.76	CURRENT_OS=Android	Application Price Next_OS=Android
R2	0.27	0.60	1.71	CURRENT_OS=Android	Next_OS=Android
R3	0.27	0.77	1.71	Next_OS=Android	CURRENT_OS=Android
R4	0.30	0.78	1.67	Next_OS=iOS	CURRENT_OS=iOS

It seems that Android users want to continue to use Android as their next mobile OS. Similarly, iOS users choose iOS as their next OS. Therefore, brand loyalty has shown to be an important criterion that has influence on consumer choices. Telecommunication companies need to focus on attracting first time smart phone users.

Table 4.3 shows that the number of applications in the store and application prices have great effects on consumer's mobile OS preferences. Consumers who prefer non-paid applications will choose Android for their next OS.

Table 4.3. Association Rules Related to Applications (min sup: 20%; min conf: 60%)

Rules	Sup.	Conf.	Lift	Consequent	Antecedent
R1	0.24	0.69	1.36	Free Application	Next_OS=Android Application Price
R2	0.27	0.93	1.38	Free Application	Male Application Price
R3	0.25	0.71	1.40	Free Application	Current_OS=Android
R4	0.25	0.60	1.23	App_Type=Social	Female
R5	0.22	0.70	1.33	Male	App_Type=Transport
R6	0.28	0.64	1.17	App_Type=Comm.	Female
R7	0.28	0.61	1.06	Male	App_Type=Game
R8	0.23	0.60	1.04	App_Type=Comm.	AGE = 23 – 32

The results illustrate that application choices are shaped in respect to the gender and the age of the consumer. For instance, female users like social and communicational applications, whereas male users prefer transportation and game applications.

Besides, users between ages of 23 and 32 like communicational applications. Since the lift values are equal or greater than 1, we can state that the entire consumer group enjoys social and communicational applications.

We have given the huge expected revenues from the application advertising in the introduction section; hence, consumer behavior towards advertising is critical information for application providers. The results in Table 4.4 reveals from which channels, consumers access to applications in applications store.

Table 4.4. Association Rules Related to Advertisements (min sup: 20%; min conf: 75%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.20	0.79	1.15	Interesting Ads	Clicks on Ads Application Price
R2	0.38	0.89	1.45	get App from Social Env.	Female
R3	0.43	0.75	1.16	get App from Market	Male
R4	0.36	0.86	1.05	get App from Social Env.	AGE = 23 – 32
R5	0.36	0.85	1.04	get App from Social Env.	get App from Market Male

Similar to application preferences, the chosen channel depends on the gender of the user. Female users reach the applications from social network advertisements, while male users find applications by searching the store. In general, most popular mobile applications are the social media applications, especially among young consumers. It would be profitable if the advertisement for young target group is placed into social media applications.

4.2. Clustering-Based Data Mining

4.2.1 K-Means Algorithm

K-means is one of the simplest unsupervised learning and partitional clustering algorithms (Alpaydın, 2010). This algorithm classifies a given data set by finding a certain number of clusters (*K*). The clusters are differentiated by their centers. The best choice is to place them as much as possible far away from each other. The algorithm is highly sensitive to initial placement of the cluster centers. A disadvantage of K-means algorithm is that it can only detect compact, hyperspherical clusters that are well separated (Witten et al., 1999). Another disadvantage is that due to its gradient descent nature, it often converges to a local minimum of the criterion function (Selim & Ismail, 1984).

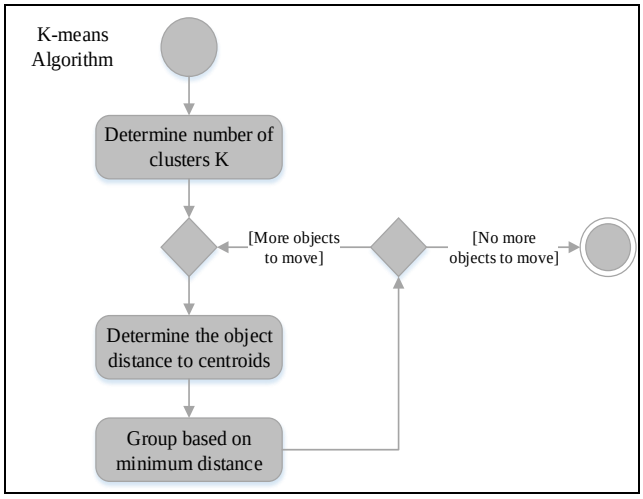

Figure 4.5. Data and Control Flow of K-Means Algorithm

The algorithm is composed of the following steps (Fig. 4.5):

1. Initial value of centroïds: Deciding *K* points randomly into the space which represent the clustered objects. These *K* points constitute the group of initial centroïds.

2. Objects-centroïd distance: Calculating the distance of each object to each centroïd, and assigning them to the closest cluster that is determined by a minimum

distance measure. (A simple distance measure that is commonly used is Euclidean distance.)

3. Determine centroids: After all objects are assigned to a cluster, recalculating the positions of the *K* centroïds.

4. Object-centroid distance: Calculating the distances of each object to the new centroïds and generating a distance matrix.

The whole process is carried out iteratively until the centroïd values become constant.

4.2.2 Expected Maximization (EM) Algorithm

Expected maximization (EM) algorithm is used in maximum likelihood estimation where the problem involves two sets of random variables of which one, X, is observable and the other, Z, is hidden. The goal of the algorithm is to find the parameter vector Φ that maximizes the likelihood of observed values of X, $L(\Phi|X)$. However, in cases where this is not feasible, we associate the extra hidden variables Z and express the underlying model using both, to maximize the likelihood of the joint distribution of X and Z, the complete likelihood $Lc(\Phi|X,Z)$ (Alpaydın, 2010).

Since the Z values are not observed, we cannot work directly with the complete data likelihood Lc; instead, we work with its expectation, Q, given X and the current parameter values Φ^l, where l indexes iteration. This is the expectation (E) step of the algorithm. Then in the maximization (M) step, we look for the new parameter values, Φ^l+1, that maximize this. Thus:

$$\text{E-step}: \ Q\left(\Phi\middle|\Phi^l\right) = E[Lc(\Phi|X,Z)|X,\Phi^l] \tag{4.4}$$

$$\text{M-step}: \Phi^l + 1 = arg\ max\Phi(\Phi|\Phi^l) \tag{4.5}$$

In the case of mixtures, the hidden variables are the sources of observations, namely, which observation belongs to which component. If these were given, for example, as class labels in a supervised setting, we would know which parameters to adjust to fit that data point. The EM algorithm works as follows: in the E-step we estimate these

labels given our current knowledge of components, and in the M-step we update our component knowledge given the labels estimated in the E-step. These two steps are the same as the two steps of K-means; calculation of b_i^t(E-step) and reestimation of m_i(M-step) (Alpaydın, 2010).

Soft clustering gives probabilities that an instance belongs to each set of clusters. Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1) (Roung-Shiunn & Po-Hsuan, 2011).

4.2.3 Hierarchical Algorithm

Hierarchical clustering is a method of cluster analysis which follows to build a hierarchy of clusters. Hierarchical cluster analysis (or hierarchical clustering) is a general approach to cluster analysis, in which the object is to group together objects or records that are ‘close’ to one another.

A key component of the analysis is repeated calculation of distance measures between objects, and between clusters once objects begin to be grouped into clusters. The outcome is represented graphically as a dendrogram (the dendrogram is a graphical representation of the results of hierarchical cluster analysis).

The initial data for the hierarchical cluster analysis of N objects is a set of N x (N – 1)/ 2 object-to-object distances and a linkage function for computation of the cluster-to-cluster distances. A linkage function is an essential feature for hierarchical cluster analysis. Its value is a measure of the "distance" between two groups of objects (i.e. between two clusters).

4.2.4 Proposed Method: The Hybrid Hierarchical K-Means Algorithm

In this study, we propose an even further improvement to the clustering algorithm used. We generate a hybrid approach that is composed of hierarchical clustering algorithm and K-means clustering algorithm. The aim was to get rid of some of the shortcomings of K-means algorithm by finding better initial cluster centroïds.

Hierarchical and K-means clustering are two major analytical tools. However, both have their own disadvantages. Hierarchical clustering cannot represent distinct clusters with similar expression patterns. Besides, as clusters grow in size, the actual expression patterns become less relevant. On the other hand, K-means clustering requires a specified number of clusters in advance and chooses initial centroïds randomly, and in addition to it, it is sensitive to outliers. The difference of the proposed approach is that it uses hierarchical clustering to decide inititial centroïds location and number of clusters as the first step. Then, it runs the K-means clustering as the second step.

The pseudocode of the proposed hierarchical-K-means algorithm is presented in Fig. 4.6.

1. Set K as the predefined number of clusters.
2. Determine p as numbers of computation
3. Set i=1 as initial counter
4. Apply hierarchical algorithm
5. Record the centroïds of clustering results as $C_i = \{c_{ij} \mid j = 1,\ldots, K\}$
6. Increment i=i+1
7. Repeat from step 4 while $i<p$
8. Assume $C = \{C_i \mid i=1,\ldots, p\}$ as new data set, with K as predefined number of clusters
9. Apply K-means algorithm
10. Record the centroïds of clustering result as $D = \{d_i \mid i=1,\ldots, K\}$

Figure 4.6. Pseudocode of the Proposed Methodology

Then, we can apply $D = \{d_i \mid i=1,\ldots, K\}$ as initial cluster centers for *K*-means clustering. The results of the proposed approach present the accuracy of our proposed method.

4.2.5 Clustering-Based Data Mining Research Framework

At the end of the case study, we aim at understanding consumer behavior and preferences when using mobile OSs on smart phones. Figure 4.7 illustrates the research framework. The first step involves developing the database. Doing so, we have observed the trends in mobile OS industry and have interviewed a marketing manager of Android OS during Google I/O. Using these knowledge, we have built the consumer questionnaire and accordingly, the tables in the relational database. The database is built on MySQL. The tables of the database reflect seven main information categories in the questionnaire. The answers coming from the consumers have been the input of the data mining process.

The main objective in this research is to show the effect of clustering into the association rules extracting process. Therefore, we use four different approaches while mining the consumer data. In Approach 1, data are not been clustered and only the Apriori algorithm is applied to obtain related association rules. In Approach 2, data are clustered using K-means algorithm, and then the Apriori algorithm is applied. In Approach 3, data are clustered using EM algorithm, and then the Apriori algorithm is applied and finally, data are clustered using hierarchical algorithm, and then the Apriori algorithm is applied in our proposed approach. The association rules are obtained from WEKA tool. The knowledge can then be integrated into the decision support process of telecommunication companies' marketing departments.

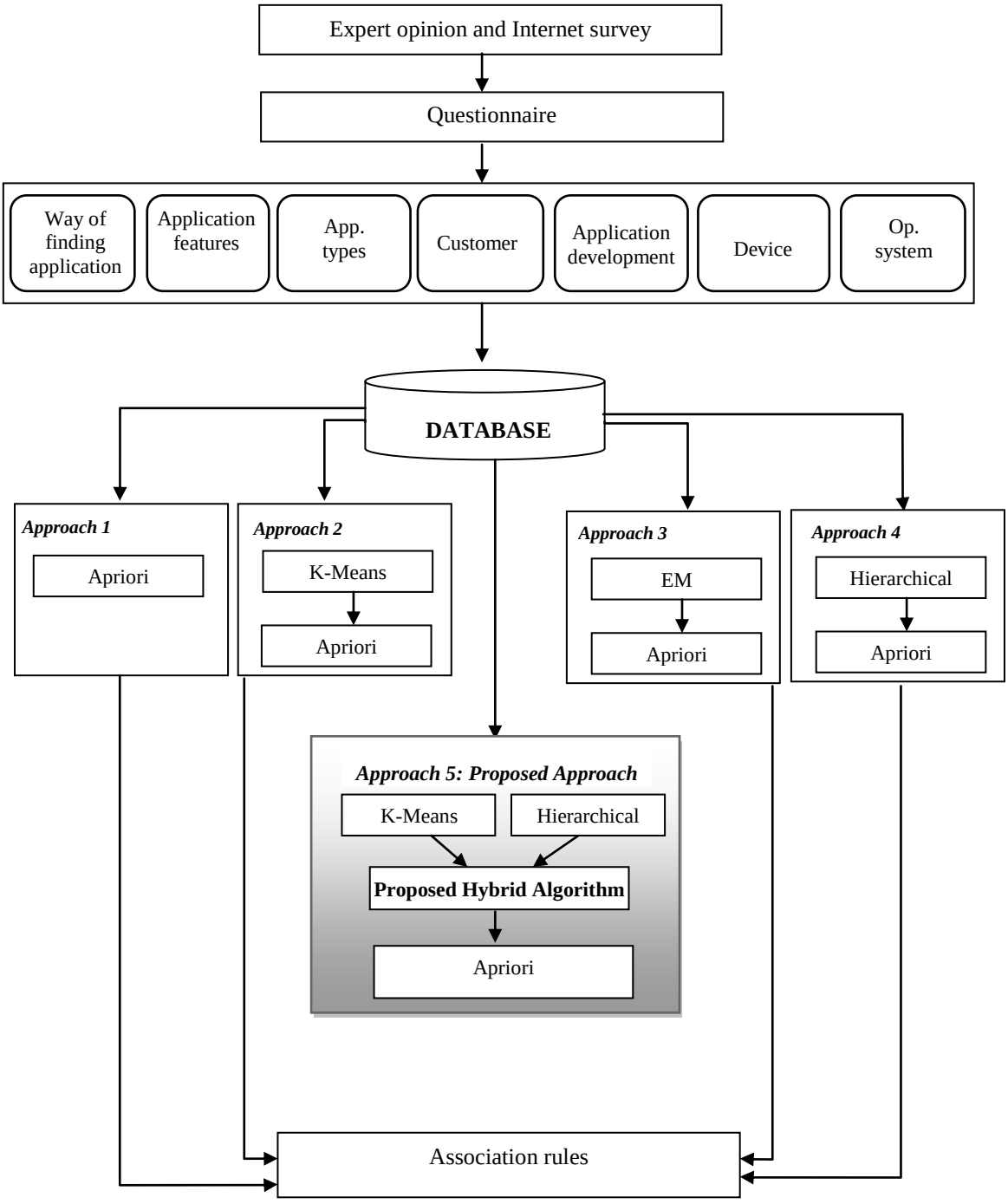

Figure 4.7. Clustering-Based Framework

4.2.6 Comparison of Clustering Algorithms

At the time the questionnaire is given to consumers, the most common mobile OSs were Android and iOS in our country. According to all consumers' answers, both Android

and iOS are found as more desirable by male consumers. Let us analyze in detail the results of five approaches separately.

4.2.6.1 Approach 1: Apriori Algorithm Without Consumer Clustering

Three rules are extracted from the demographic information, referred as R1, R2, and R3. The analysis results in Table 4.5 reveal that Android is found more attractive among male consumers, whereas iOS is more preferable by female consumers. Moreover, Android seems to be more desirable by the young consumers of age 23-32 years.

Table 4.5. Association Rules from Demographic Information (min sup: 20%; min conf: 45%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.51	1.14	Next_OS=Android	AGE=23 - 32
R2	0.26	0.46	1.02	Nextt_OS=Android	Male
R3	0.20	0.47	1.01	Next_OS=iOS	Female

The minimum thresholds of support and confidence are set to 20% and 60%, respectively. The data analysis results in Table 4.6. show that two leaders in mobile OS market are Android and iOS, and their desirability is very close to each other. It seems that Android users want to continue to use Android as their next mobile OS. Similarly, iOS users choose iOS as their next OS. Therefore, brand loyalty has proven to be an important criterion that has influence on consumer choices, even in this small customer dataset. Telecommunication companies need to focus on attracting first time smart phone users.

Table 4.6. Association Rules Related to Brand (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.60	1.76	Current_OS =Android	Application Price Next_OS=Android
R2	0.27	0.60	1.71	Current_OS =Android	Next_OS=Android
R3	0.27	0.77	1.71	Next_OS=Android	Current_OS=Android
R4	0.30	0.78	1.67	Next_OS=iOS	Current_OS=iOS

Table 4.7 shows that the numbers of applications in the store and application prices have great effects on consumer's mobile OS preferences. Consumers who prefer non-paid applications will choose Android for their next OS.

Table 4.7. Association Rules Related to Applications (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.24	0.69	1.36	Free Application	Next_OS=Android Application Price
R2	0.27	0.93	1.38	Free Application	Male Application Price
R3	0.25	0.71	1.40	Free Application	Current_OS=Android
R4	0.25	0.60	1.23	Application Type =Social	Female
R5	0.22	0.70	1.33	Male	App. Type = Transport
R6	0.28	0.64	1.17	App. Type= Communications	Female
R7	0.28	0.61	1.06	Male	App Type Game
R8	0.23	0.60	1.04	App. Type= Communications	AGE = 23 – 32

The data analysis results illustrate that application choices are shaped in respect to the gender and the age of the consumer. For instance, female users like social and communicational applications, whereas male users prefer transportation and game applications. Besides, users between ages of 23 and 32 like communicational applications. Since all the lift values are equal or greater than 1, we can state that the entire consumer group enjoys social and communicational applications.

The application advertising has huge expected revenues; hence, consumer behavior towards advertising is a critical information source for application providers. The results in Table 4.8 reveal from which channels consumers access to applications in applications store.

Similar to the application preferences, the chosen channel depends on the gender of the user. Female users reach the applications from social network advertisements, while male users find applications by searching the store. In general, most popular mobile applications are the social media applications, especially among young consumers. It would be profitable if the advertisement for young target group is placed into social media applications.

Table 4.8. Association Rules Related to Advertisement (min sup: 20%; min conf: 75%)

Rules	Support	Conf.	Lift	Consequent	Antecedent
R1	0.20	0.79	1.15	Interesting Ads	Clicks on Ads Application Price
R2	0.38	0.89	1.45	Get App from Social Env.	Female
R3	0.43	0.75	1.16	Get App from Market	Male
R4	0.36	0.86	1.05	Get App from Social Env.	AGE = 23 – 32
R5	0.36	0.85	1.04	Get App from Social Env.	Get App from Market Male

4.2.6.2 Approach 2: Apriori Algorithm With Consumer Clustering Using K-Means

The clustering process which is generated by K-means algorithm reveals five meaningful groups of data. Obtained clusters are summarized in Table 4.9. The number of instances is the number of consumer data in the given cluster.

Let us interpret Table 4.10 that represents the general view of clustering results. $Cluster_0$ is the most crowded group; therefore the highest cross selling opportunity is obtained. The most popular mobile OS is Android in this cluster. One of the most significant results is that, young population with relatively lower income prefers Android and plans to continue with Android. They are found as more sensitive to price, and they prefer Samsung as the mobile phone brand. $Cluster_0$ reveals cross selling opportunities in telecommunication market in Turkey. For instance Avea, a Turkish mobile operator, has a campaign for Samsung smart phones with Android, for the consumers of 23 to 32 years old.

In $Cluster_1$, female consumers generally use iPhone (iOS) and specify iPhone as their next choices. This cluster's mobile phone choices are influenced primarily from the brand, not from the price range. Hence, it would be a good idea to create social media campaigns for female consumers. While consumers of ages of 23 to 32 in $Cluster_0$ prefer Android, the same age group in $Cluster_1$ prefers iPhone (iOS) as their smart phone.

Table 4.9. Consumer Clusters and Their Number of Instances

	# of instances	Ratio (%)
$Cluster_0$	65	0.31
$Cluster_1$	41	0.20
$Cluster_2$	41	0.20
$Cluster_3$	42	0.20
$Cluster_4$	18	0.09

Table 4.10. General Cluster Model

	Full Data	$Cluster_0$ (65)	$Cluster_1$ (41)	$Cluster_2$ (41)	$Cluster_3$ (42)	$Cluster_4$ (18)
Gender	Male	Male	Female	Male	Female	Male
Age avg.	23-32	23-32	15-22	23-32	15-22	23-32
Income (TL)	0-2000	0-2000	0-2000	2000- 4000	0-2000	2000- 4000
Product Price	Medium	Medium	Low	Low	Medium	Low
Brand	Apple	Samsung	Apple	Apple	Blackberry	HTC
Current_OS	IOS	Android	IOS	IOS	Blackberry	Android
Operator	Turkcell	Avea	Turkcell	Turkcell	Turkcell	Avea
Next_OS	IOS	Android	IOS	IOS	Android	Android

The distribution of instances according to clusters which is found by K-means clustering algorithm (Fig. 4.8).

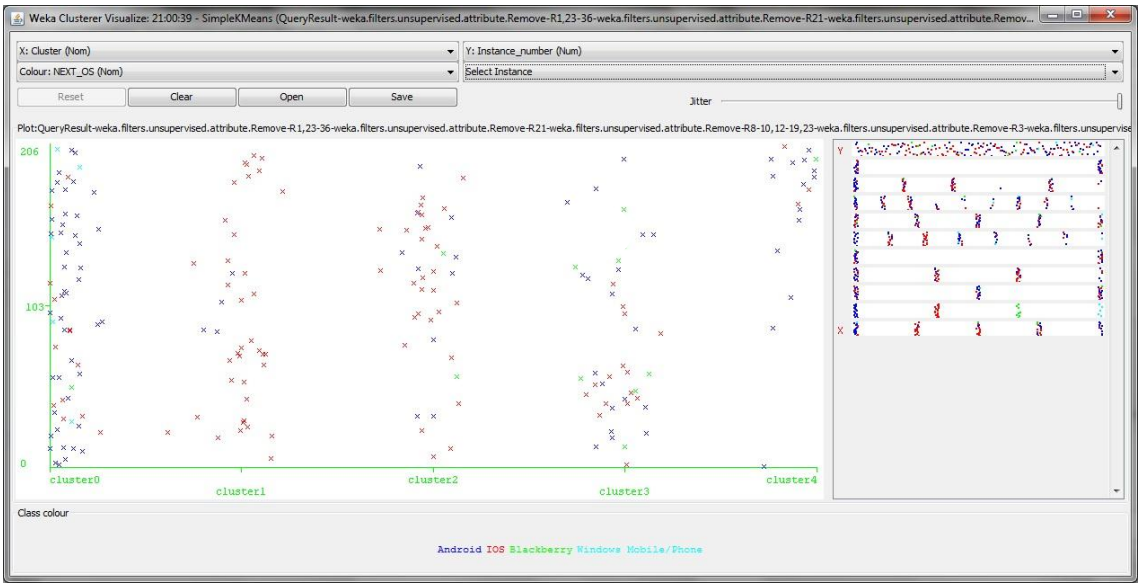

Figure 4.8. K-Means Cluster Analysis with WEKA

In $Cluster_3$, Blackberry users are more crowded compared to iPhone users among university students. Majority of this cluster consists of female consumers who prefer Android as their next choices of mobile OS. Opposite to $Cluster_0$, $Cluster_4$ includes consumers with relatively higher income value, but they are from the same age group. Consumers of $Cluster_4$ prefer HTC as the smart phone brand.

For the comparison purposes, we observe the values of the same association rules found in Approach 1. But actually, we find more rules with higher accuracy values using the Approach 2. Similar to the analysis of the Approach 1, Android and iOS are two mobile OS with the highest utilization rate and they are the most desirable ones (Table 4.11). The confidence and support values are higher than the ones in Approach 1, but the lift values are lower.

The attributes *Next_OS, Current_OS, Brand* and the attributes related to consumer demographic information were dominating the association rules extracting from all data. Hence, we exclude them to observe in more detail the association rules related to applications and advertisements. The *K*-means algorithm is applied to this data subset, and four meaningful clusters are obtained (Table 4.12).

Table 4.11. Association Rules Using Clustering Related to Brand (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.5	0.70	1.05	Current_OS=Android	Application Price Next_OS= Android
R2	0.6	0.75	1.07	Current_OS=Android	Next_OS=Android
R3	0.6	0.87	1.07	Next_OS=Android	Current_OS=Android
R4	0.6	0.94	1.05	Next_OS=iOS	Current_OS=iOS

In $Cluster_0$, the male consumers of ages 23 to 32 usually prefer to find applications from application market and social media instead of Internet and media. 75% of the applications chosen from consumers of $Cluster_0$ are free, while 25% of them are paid. So, the advertisement in the applications would not reach $Customer_0$ group.

Table 4.12. Cluster Model for Applications and Advertisement Data

Attribute	Full Data (207)	$Cluster_0$ (70)	$Cluster_1$ (51)	$Cluster_2$ (46)	$Cluster_3$ (40)
Age	23-32	23-32	33-50	15-22	23-32
Gender	Male	Male	Female	Male	Female
App. Price	Free	%75 Free – %25 Paid	%50 Free – %50 Paid	Free	Free
Market	Very High	High	Very High	Very High	Very Low
Social Envir.	Very High	Very High	Very High	High	Very High
Internet Ads	Low	Low	Medium	Medium	Very Low
App. Ads	Very Low	Very Low	Very Low	Low	Low
Media Ads	Very Low	Medium	Very Low	Very Low	Low

In $Cluster_1$, the majority of the users of paid applications are female of the age group of 35 to 50. Therefore, it is reasonable to attract more consumers by introducing free downloadable application campaigns. The $Cluster_2$ consists of male consumers of age of 15 to 22, who find their applications from the social environments, in other words from the friends. The advertisements that have this group as the target would be inefficient. The $Cluster_3$ includes female consumers of age of 23 to 32. Unlike other clusters, they do not find the applications by searching the application market. They mostly learn them in the social environment. Again, the media advertisement is not efficient for this group of consumer.

The reason that there are only three association rules in Approach 2 instead of eight is that, we select the same rules as in Approach 1, for the purpose of comparison (Table 4.13). After clustering the data, the confidence, the support and the lift values of R4, R6, and R7 all increase.

Both support and confidence values of three rules are higher than the support and the confidence values of the same rules in Approach 1 (Table 4.14).

Table 4.13. **Association Rules Using Clustering Related to Application (min sup: 20%; min conf: 60%)**

Rules	Support	Confidence	Lift	Consequent	Antecedent
R4	0.44	1	1.64	App type=Social	Female
R6	0.41	0.84	1.38	App type=Comm.	Female
R7	0.45	0.63	1.62	Male	App type= Game

Table 4.14. Association Rules Using Clustering Related to Advertisement (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R2	0.39	0.78	1.15	Get App from Social Env	Female
R3	0.40	0.91	1.12	Get App from Market	Male
R4	0.34	0.75	2.02	Get App from Social Env	AGE=23 – 32

4.2.6.3 Approach 3: Apriori Algorithm with Consumer Clustering Using Expected Maximization (EM)

For comparison purposes, we explore the expectation–maximization (EM) clustering algorithm, implemented in WEKA, to cluster the consumer data. This algorithm can be applied to multiple items as long as the items can be assumed to be independent. Consumers are assigned to different classes in a probabilistic manner. In order to compare the resulting mixed-membership distributions with the 9 micro segments that resulted from the soft clustering, we had the EM algorithm clusters.

As Table 4.15 shows, the influence of the number of iterations is negligible. The number of clusters remains the same and has not an effect on the number of iterations.

The experiments show that the minimum standard deviation has not an effect on the number and the composition of the clusters. A change occurs in the range of 0.000001 to 1, where the number of clusters remains 5.

Table 4.15. EM Algorithm Loglikelihood Tries

	Seed	Cluster	Log likelihood
0.000001	100	5	-26.22402
0.000001	200	5	-26.22192
0.00001	100	5	-26.22192
0.0001	100	5	-26.22192
0.001	100	5	-26.22192
0.005	100	5	-26.22192
0.01	100	5	-26.22192
0.01	**500**	**5**	**-26.22347**
1	100	5	-26.22192
0.1	100	5	-26.22192
0.1	500	5	-26.21107
1	100	5	-26.22192

Table 4.16. Consumer Clusters and Their Number of Instances

	# of instances	Ratio (%)
$Cluster_0$	36	0.36
$Cluster_1$	46	0.46
$Cluster_2$	19	0.19
$Cluster_3$	60	0.60
$Cluster_4$	46	0.46

The clustering process which is generated by EM algorithm reveals five meaningful groups of data. Table 4.16. summarizes obtained clusters. The number of instances represents the number of consumer data in the given cluster. Figure 4.9 is the distribution of instances according to clusters which is determined by EM clustering algorithm in WEKA.

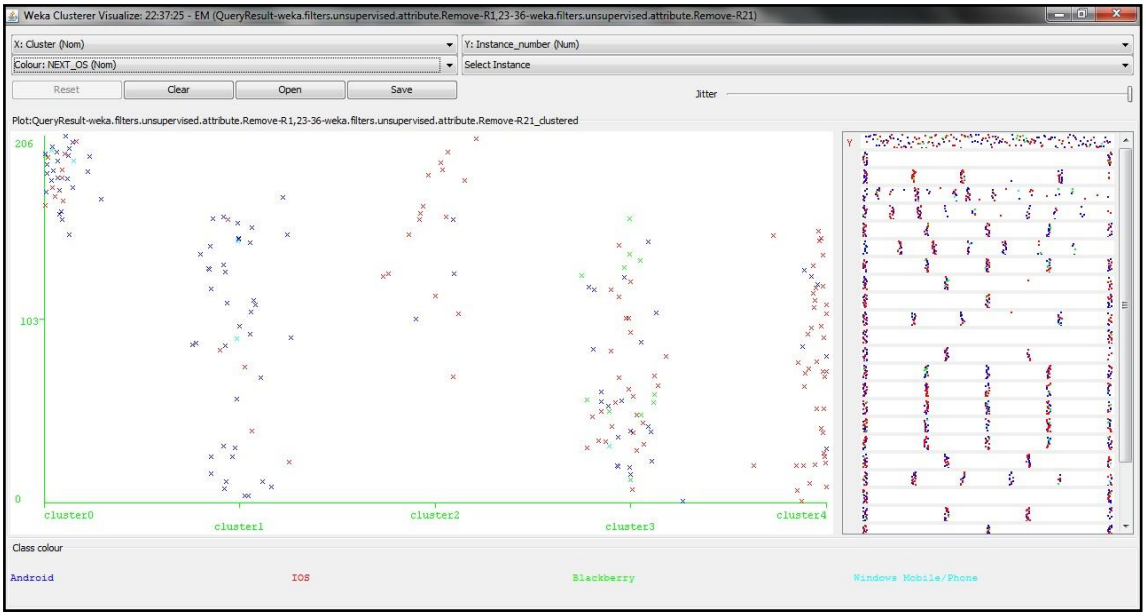

Figure 4.9. EM Cluster Analysis with WEKA

Table 4.17. Association Rules Using Clustering Related to Brand (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.41	0.72	1.08	Current_OS=Android	Application Price Next_OS= Android
R2	0.41	0.75	1.11	Current_OS=Android	Next_OS=Android
R3	0.30	0.71	1.12	Next_OS=Android	Current_OS=Android
R4	0.32	1	1.07	Next_OS=iOS	Current_OS=iOS

The minimum thresholds of support and confidence are set to 20% and 60%, respectively, which is the same as in K-means-based Apriori algorithm approach. There is approximately %5 of raise on the confidence and lift values of the association rules R1 and R2 (Table 4.17). There are not any changes in the confidence and lift values of the rules R2 and R4. Therefore, we may conclude that EM algorithm gives accurate confidence and lift values.

Table 4.18. Association Rules Using Clustering Related to Advertisement (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R2	0.36	0.78	1.43	Get App from Social Env	Female
R3	0.35	0.69	1.10	Get App from Market	Male
R4	0.44	0.56	1.03	Get App from Social Env	AGE=23 – 32

In EM-based Apriori algorithm approach (Approach 3), when application and advertisement related association rules are studied, we can see that both support and confidence values of rules are lower than the ones of the same rules in Approach 2. (Table 4.18 and Table 4.19). Therefore it shows Approach 2 (K-means based) is more appropriate than Approach 3 (EM-based) for our data set.

When the all rules are examined, the results of K-means algorithm are much accurate and reliable than the result of EM algorithm.

Table 4.19. Association Rules Using Clustering Related to Application (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R4	0.38	0.93	1.23	App type=Social	Female
R6	0.34	0.76	1.07	App type= Comm.	Female
R7	0.35	0.61	1.60	Male	App type= Game

4.2.6.4 Approach 4: Apriori Algorithm with Consumer Clustering Using Hierarchical Clustering Methods

In hierarchical clustering-based Apriori algorithm approach, there are five cluster which are the same as K-means and EM algorithms (Table 4.9. and Table 4.16). Figure 4.10 shows the distribution of instances according to clusters which is found by hierarchical clustering algorithm.

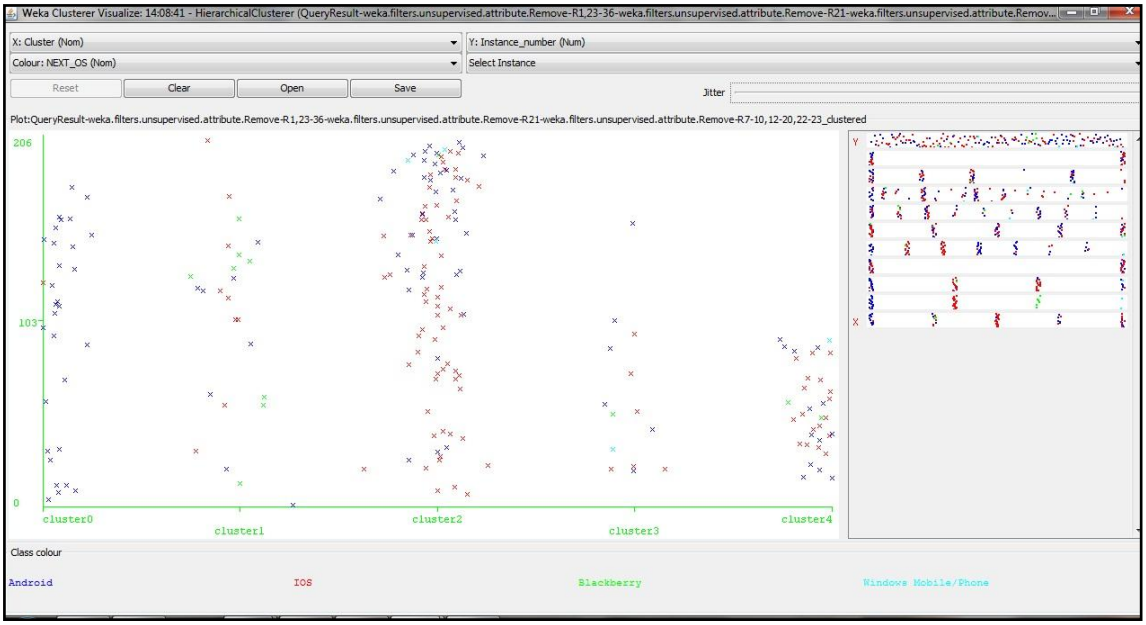

Figure 4.10. Hierarchical Clustering Analysis with WEKA

In hierarchical clustering-based Apriori algorithm approach, in all the rules, both support and confidence values are higher than the ones in all other approaches (Table 4.18 and Table 4.19). As a result, the hierarchical clustering algorithm gives more accurate results compared to the EM algorithm.

The minimum thresholds of support and confidence are set to 20% and 60%, respectively, such as in K-means-based and EM-based Apriori algorithm approach. The data analysis results in Table 4.20 show that there is approximately 10% of raise on the confidence and lift values of all association rules.

Similarly, there is a raise in the lift and confidence values on application and advertisement related association rules (Table 4.21 and Table 4.22).

Table 4.20. Association Rules Using Clustering Related to Brand (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.41	0.92	1.5	Current_OS=Android	Application Price Next_OS=Android
R2	0.41	0.76	2.04	Current_OS=Android	Next_OS=Android
R3	0.30	0.72	2.04	Next_OS=Android	Current_OS=Android
R4	0.32	0.9	1.41	Next_OS=iOS	Current_OS=iOS

Table 4.21. Association Rules Using Clustering Related to Advertisement (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R2	0.36	1	1	Get App from Social Env	Female
R3	0.35	0.83	1.08	Get App from Market	Male
R4	0.44	0.67	1.44	Get App from Social Env	AGE=23 – 32

Table 4.22. Association Rules Using Clustering Related to Application (min sup: 20%; min conf: 60%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R4	0.35	0.9	1.18	App type=Social	Female
R6	0.39	1	1.18	App type= Comm.	Female
R7	0.45	1	1.02	Male	App type= Game

The confidence and lift values of association rules in Approach 4 are more accurate and reliable than the ones in Approach 2 (K-means based Apriori) (Table 4.23).

We applied to our data set K-means clustering algorithm, Expected Maximization algorithm and hierarchical clustering algorithm (Bottom-up) which are most common clustering algorithm in data mining. The best result is obtained by Approach 2 (K-means based Apriori algorithm) and Approach 4 (hierarchical-based Apriori algorithm).

When examining all the results of the approaches, we have claimed that a hybrid algorithm which composes of K-means and hierarchical clustering algorithm would give most predictive results.

Table 4.23. List of Applied Approaches

Approach 1	Apriori Algorithm
Approach 2	K-Means-based Apriori Algorithm
Approach 3	Expected Maximization-based Apriori Algorithm
Approach 4	Hierachical-based Apriori Algorithm
Approach 5	Proposed Method: The Hybrid Hierachical K-means Algorithm

4.2.6.5 Approach 5: Apriori Algorithm with Consumer Clustering Using Proposed Hybrid Hierarchial K-Means Method

The main purpose in hybrid hierarchical K-means algorithm is to increase the confidence and lift values of the association rules. The increase is presented in Table 4.24. Although there are minor decreases at the lift values of the association rules, they still provide with the general lift condition (lift value > 1). We can conclude that the association rules in our proposed approach are more accurate.

Table 4.24. Association Rules Using Clustering Related to Brand (min sup: 20%; min conf: 75%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R1	0.21	0.88	1.40	Current_OS=Android	Application Price Next_OS=Android
R2	0.27	0.87	1.45	Current_OS=Android	Next_OS=Android
R3	0.27	0.82	1.45	Next_OS=Android	CURRENT_OS=Android
R4	0.30	0.95	1.39	Next_OS=iOS	CURRENT_OS=iOS

The proposed algorithm gets rid of the weak points of K-means algorithm in finding the initial centroïds by applying hierarchical algorithm. The minimum thresholds of support value are set to 20% as in all the other approaches. The minimum threshold of confidence is set to 75%. Table 4.24 shows that there is an increase on confidence and lift values in all the association rules compared to the K-means-based algorithm.

When advertisement and application related association rules are analyzed, it is possible to observe the increase in confidence and lift values. The results of the brand related association rules are also acceptable (Table 4.25 and Table 4.26).

Table 4.25. Association Rules Using Clustering Related to Application (min sup: 20%; min conf: 75%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R4	0.44	0.89	1.20	App type=Social	Female
R6	0.41	0.88	1.14	App type= Comm.	Female
R7	0.45	0.75	1.02	Male	App type= Game

Table 4.26. Association Rules Using Clustering Related to Advertisement (min sup: 20%; min conf: 75%)

Rules	Support	Confidence	Lift	Consequent	Antecedent
R2	0.39	1	1.5	Get App from Social Env	Female
R3	0.40	0.80	1.19	Get App from Market	Male
R4	0.34	0.75	1.19	Get App from Social Env	AGE=23 – 32

The highest confidence value is found in K-means algorithm among five approaches; while the highest lift value is obtained using the hierarchical algorithm. Therefore, we propose the combination of them in order to increase the lift values of K-means algorithm (Fig. 4.12 and Fig. 4.13).

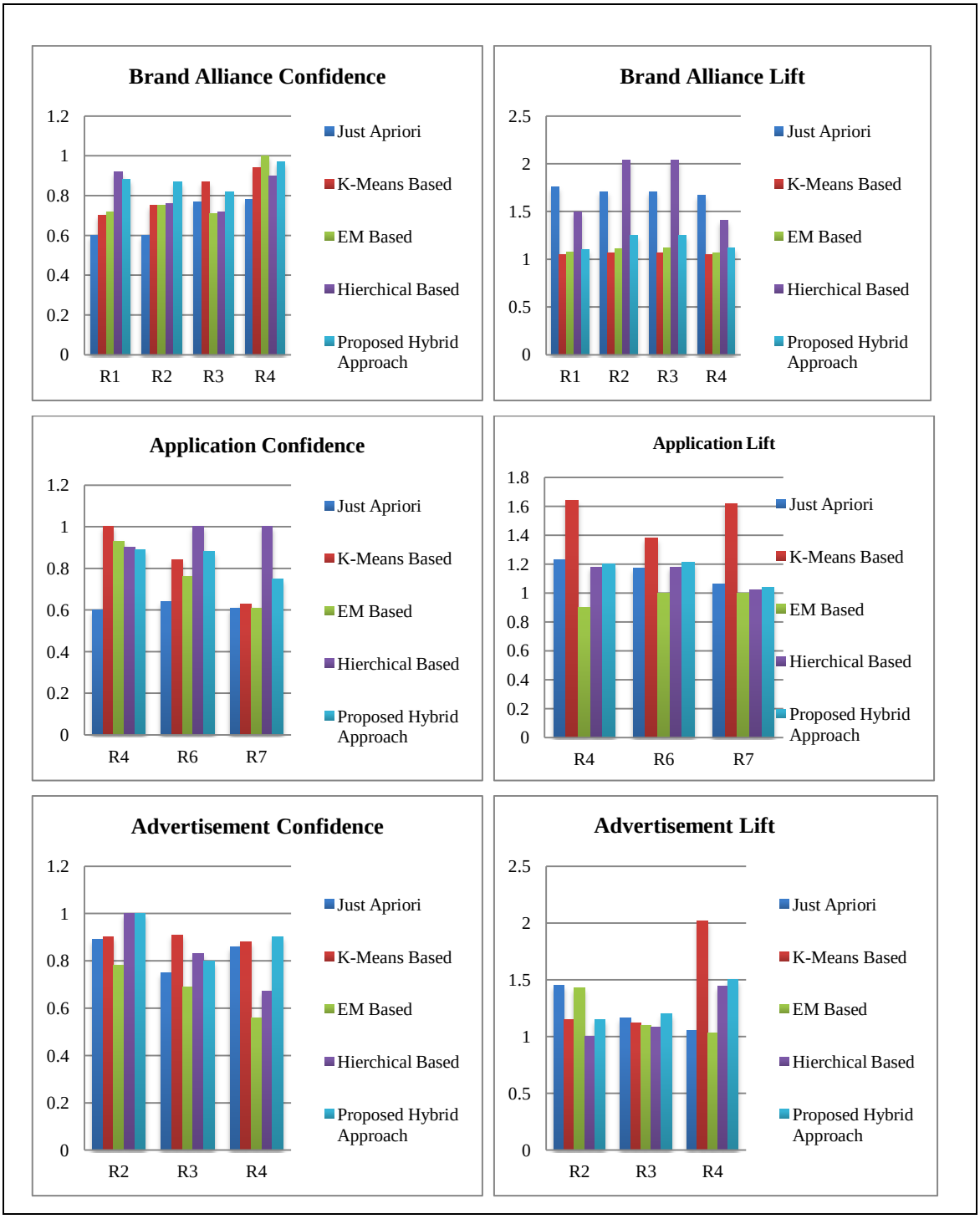

Figure 4.11. General Comparison

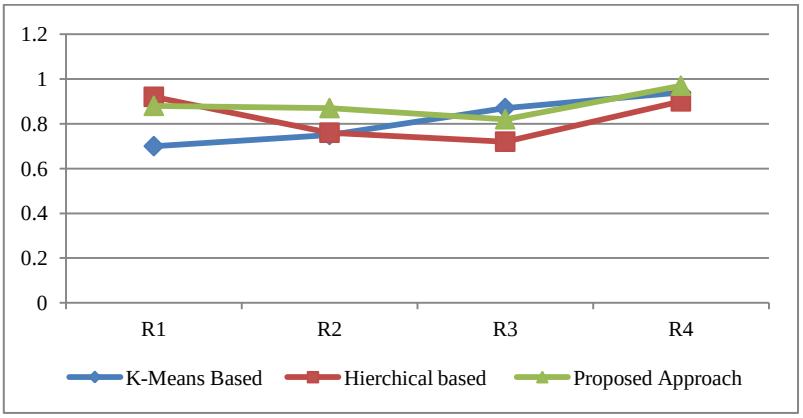

Figure 4.12. Comparison of the Confidence of the Proposed Methodology

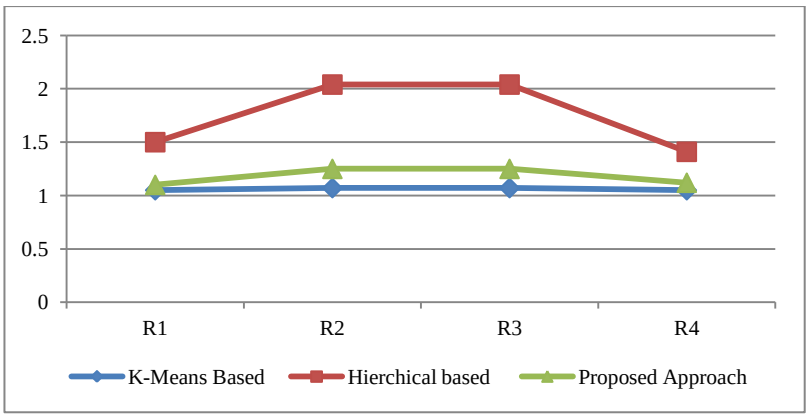

Figure 4.13. Comparison of the Lift of the Proposed Methodology

5. PERFORMANCE ANALYSIS

We have used three performance metrics in order to analyze the performance of the approaches. These metrics are: the number of iterations, the computation time and the system memory usage. Table 5.1 presents the number of iterations of the algorithms to converge.

Table 5.1. Number of Iterations Before Convergence

Algorithm	Number of iterations to converge
K-Means	7
Hierarchical	1
Hybrid HKM	5

Table 5.2. Computation Time of the Algorithms

Algorithm	Computation time
K-Means	0.08 s
Hierarchical	0.5 s
Hybrid HKM	Approx. 0.475 s

Table 5.3. Space Requirements of the Algorithms

Algorithm	Computation time
K-Means	4.60 MB
Hierarchical	4.74 MB
Hybrid HKM	5.90 MB

The results reveal that the hybrid hierarchical K-means has less iterations than K-means.

6. CONCLUSION

Marketing is important in all areas of the organization, and customers are the reason why businesses exist. If companies make sense of customer needs and manage the relationships more intelligently, it is obvious that they will provide crucial competitive differentiation to gain market share and retaining customers. Customer retention marketing is a tactically-driven approach based on customer behavior.

This research first step proposes a data mining approach which can be appropriate for any sector to mine customer knowledge. The data mining in the framework is realized using one of the most known data mining algorithms: Apriori algorithm. Apriori algorithm helps us to find association rules. As a case study, we have chosen the mobile operating system industry. We have aimed at mining the knowledge of the consumers who use smart phones. Since the competition in the mobile OS and their applications' marketing becomes more and more harsh, precise consumer segmentation will help providers to target exact customer profiles. Extracted customer knowledge will direct providers to offer the right product to right customer. It is doubtless that our customer knowledge of 209 users is not enough to extract general strategies for mobile OS industry, but we have aimed at demonstrating the usefulness of our model in a simple case. We have made use of a questionnaire to receive consumer information. The responses to this questionnaire were taken in mid-2012, and now, we have seen that the results are accurate.

One of the findings of this study is that, we have seen that choosing an ontology-based approach is not the most appropriate decision to deal with this type of domain with quite structured data. The domain ontology could have been constructed with only expert decisions. Hence, the usage of the tool Protégé and obtaining an OWL file have not bring much dimension to our work. We have seen that, an ontology-based data mining approach is more suitable for unstructured data, such as data on WWW.

The intense competition and increased choices available for customers have created new pressures on marketing decision-makers and there has emerged a need to manage customers in a long-term relationship. If companies make sense of customer needs and manage the relationships more intelligently, it is obvious that they will provide crucial competitive differentiation to gain market share and retaining customers. Customer retention marketing is a tactically-driven approach based on customer behavior.

This study uses a research framework which can be appropriate for any sector to mine customer knowledge. The data mining is realized using two of the most known data mining algorithms: Apriori algorithm, K-means algorithm, Expected Maximization Algorithm and hierarchical clustering methods. They help us to find association rules. We have compared the resulting association rules in four different data analysis approaches. In the first analysis, data are not clustered, whereas the other analysis data are clustered.

It is widely reported that the K-means algorithm suffers from initial cluster centers. Our main purpose is to optimize the initial centroïds for K-means algorithm. Therefore, in this research, it is proposed Hierarchical K-means algorithm. It utilizes all the clustering results of K-means in certain times, even though some of them reach the local optima. Then, we transform the all centroïds of clustering result by combining with hierarchical algorithm in order to determine the initial centroïds for K-means. Hierarchical K-means bargains the advantage of K-means algorithm in speed and hierarchical algorithm in precision. Experimental results with random normal data distribution, our data sets performs the accuracy and improved clustering results as compared to proposed other approaches in this study.

REFERENCES

Adigun, A.A., Omidiora, E.O., Olabiyisi, S.O., Adentunji, A., Adedeji, O.T. (2012). Development of a hybrid K-means-expectation maximization clustering algorithm. Journal of Computations & Modelling, 2 (4), p.1-23.

Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In: International Conference on Very Large Databases, p.487-499.

Alpaydın E. (2010). Introduction to Machine Learning. The MIT Press, London.

Berson, A., Smith, S., Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill.

Boufardea, E., Garofalakis, J. (2012). A Predictive System for Distance Learning Based on Ontologies and Data Mining. In: The Fourth International Conference on Advanced Cognitive Technologies and Applications, p.151-158.

Chen, W.C., Hsu, C.C., Hsu, J.N. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response mode. Expert Systems with Applications, 38, p.7451–7461.

Cheng, L.C., Sun, L.M. (2012). Exploring consumer adoption of new services by analyzing the behavior of 3G subscribers: An empirical case study. Electronic Commerce Research and Applications, 11, p.89-100.

Cheng-Hsien, T., An-Ching, H., Meng-Feng, T., Wei-Jen, W. (2010). An Efficient Distributed Hierarchical-Clustering Algorithm for Large Scale Data. In: International Computer Symposium, p.869-874.

Guo Yan, H., Dong Mei, Z., Jia Dong, R., Chang Zhen, H. (2009). Hierarchical Clustering Algorithm based on K-means with Constraints. In: Fourth International Conference on Innovative Computing, Information and Control, p.1479-1482.

Kuo, Y.T., Lonie, A., Sonenberg, L., Paizis, K. (2007). Domain ontology driven data mining: A medical case study. In: International Workshop on Domain Driven Data Mining, p.11-17.

Liao, S.H., Chen, J.L., Hsu, T.Y. (2009). Ontology-based data mining approach implemented for sport marketing. Expert Systems with Applications, 36, p.11045–11056.

Lu, J.F., Tang, J.B., Tang, Z.M., Yang, J.Y. (2008). Hierarchical initialization approach for K-Means clustering. Pattern Recognition Letters, p.787–795.

Ngai, E.W.T., Xiu, L., Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36, p.2592-2602.

Nigro, H.O., Gonzalez Cisaro, S.E., Xodo, D.H. (2007). Data Mining With Ontologies: Implementations, Findings and Frameworks. Idea Group Reference.

Noy, N.F., McGuinness, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology, Stanford University. URL: http://protege.stanford.edu/publications/ontology_development/ontology101.html.

Roung-Shiunn, W., Po-Hsuan, C. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications, 10, p.331–341.

Rygielski, C., Wang, J.C., Yen, D.C. (2002). Data mining techniques for customer relationship management. Technology in Society, 24, p.483-502.

Sameh, H.G. (2008). Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language. In: International Journal of Information Technology 3(3), p.168-172.

Selim, S.Z., Ismail, M.A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6 (1), p.81-87.

Tung-Shou, C., Tzu-Hsin, T., Yi-Tzu, C., Chin-Chiang, L., Rong-Chang, C. (2005). A Combined K-Means and Hierarchical-Clustering Method for Improving the Clustering Effiviency of Microarray. In: International Symposium on Intelligent Signal Processing and Communication Systems, p.405-408.

Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. In: Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, p.192-196.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge of Information Systems, 14, p.1–37.

Yokome, E.A., Arantes, F.L. (2011). Meta-DM: An ontology for the data mining domain. In: Revista de Sistemas de Informacao da FSMA, 8, p.36-45.

APPENDIX

Mobil İşletim Sistemleri Anketi

Cinsiyet

Bayan ☐ Bay ☐

Yaş aralığı

15-22 ☐ 23-32 ☐ 33-50 ☐ 50+ ☐

Gelir aralığınızı seçiniz

0-2000 TL ☐ 2000-4000 TL ☐ 4000-6000 TL ☐ 6000-1000 TL ☐
10000 TL üzeri ☐

Akıllı telefonunuzda hangi mobil işletim sistemini kullanıyorsunuz?

Android ☐ Blackberry ☐ IOS ☐ Windows Phone ☐

Hangi tür mobil cihazlara sahipsiniz?

Telefon ☐ Tablet ☐ Telefon ve tablet ☐

Akıllı telefonunuzda hangi mobil operatörü kullanıyorsunuz?

Avea ☐ Turkcell ☐ Vodafone ☐

Kullanmakta olduğunuz akıllı telefonun markası nedir?

Apple ☐ Blackbery ☐ HTC ☐ LG ☐ Nokia ☐
Samsung ☐ Sony ☐ Diğer ☐

Kullandığınız akıllı telefonu seçerken, pazardaki muadil ürünlerin fiyatları kararınızı etkiledi mi?(1 ile 5 arasında değer veriniz. 1: Çok etkiledi , 5: Hiç etkilemedi)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Kullandığınız akıllı telefonu seçerken görünüş / tasarım özellikleri ne kadar önemli oldu? (1 ile 5 arasında değer veriniz. 1: Seçimimdeki en önemli neden tasarımdı , 5: Tasarım veya görünüşüne bakarak seçmedim.)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Kullandığınız akıllı telefonun Flash desteği olup olmaması, seçimizi etkiledi mi?

Evet ☐ Hayır ☐

Seçtiğiniz işletim sisteminde uygulama geliştiriyor musunuz? Veya ilerde geliştirmeyi düşünür müsünüz?

Evet ☐ Hayır ☐ Belki İleride Olabilir ☐

Mobil işletim sistemlerinde uygulama geliştirme ortamı / dili sizin için ne kadar önemlidir?

Uygulama geliştirmeyle ilgilenmiyorum. ☐
Benim için mobil işletim sistemlerinde uygulama geliştirme dili önemlidir. ☐
Önemli değildir, her ortamda uygulama geliştirebilirim. ☐

Uygulamaların, uygulama marketleri tarafından kabul sürecinin kısa olması sizin için önemli midir?

Evet ☐ Hayır ☐

İlginizi çeken bir reklam olduğunda tıklıyor musunuz?

Evet ☐ Hayır ☐ Bazen ☐

Sizin için bir reklamda hangisi daha önemlidir?

Reklamın yaratıcı ve / veya komik olması. ☐
Reklamda oynayan ünlü kişiler. ☐
İkisi de önemli değil, sadece reklamın içeriğine bakarım. ☐
Ne olursa olsun, reklamlar kesinlikle ilgimi çekmez. ☐

Beğendiğiniz bir uygulamanın ücretli olması uygulamayı indirmenizi engeller mi?

Evet ☐ Hayır ☐

İndireceğiniz uygulamaları genelde nasıl buluyorsunuz/seçiyorsunuz? 1'den 5'e kadar sıralayınız. (1: En çok tercih ettiğiniz, 5: En az tercih ettiğiniz yol) (Her satırda sadece bir seçeneği işaretleyiniz.)

	1	2	3	4	5
Düzenli olarak uygulama marketlerini gezer, popüler uygulama listelerine bakarım.	☐	☐	☐	☐	☐
Arkadaş çevremden duyduklarıma bakarım.	☐	☐	☐	☐	☐
Internet üzerindeki reklamlardan ulaşırım.	☐	☐	☐	☐	☐
İndirdiğim diğer uygulamaların içindeki reklamlardan ulaşırım.	☐	☐	☐	☐	☐
Gazete veya dergilerdeki reklamlardan ulaşırım.	☐	☐	☐	☐	☐

Kullandığınız uygulamaların ne kadarı ücretli, ne kadarı ücretsizdir?

%25 ücretli - %75 ücretsiz ☐ %50 ücretli - %50 ücretsiz ☐
%75 ücretli - %25 ücretsiz ☐ Hepsi ücretsiz ☐
Hepsi ücretli ☐

En çok kullandığınız 4 uygulama tipini seçip, kullanım sıklığına göre 1'den 4'e kadar sıralayınız. (1: En sık kullandığınız) (Önceliklerine göre sıralayınız.)

	1	2	3	4
Oyun	☐	☐	☐	☐
Kitap	☐	☐	☐	☐
İş (Ofis, takvim, e-posta uygulamaları, vb.)	☐	☐	☐	☐
İletişim (What's App, chat uygulamaları, vb.)	☐	☐	☐	☐
Eğitim	☐	☐	☐	☐
Eğlence	☐	☐	☐	☐

Finans	☐	☐	☐	☐
Müzik	☐	☐	☐	☐
Fotoğraf	☐	☐	☐	☐
Alışveriş	☐	☐	☐	☐
Sosyal (Facebook, Twiter vb.)	☐	☐	☐	☐
Spor	☐	☐	☐	☐
Gezi	☐	☐	☐	☐
Hava Durumu	☐	☐	☐	☐
Araçlar (Çeşitli telefon uygulamaları)	☐	☐	☐	☐
Ulaşım	☐	☐	☐	☐

Bir sonraki akıllı telefonunuzda hangi işletim sistemini kullanmak istersiniz?
Android ☐ Blackberry ☐ IOS ☐ Windows Phone

BIOGRAPHICAL SKETCH

Nergis Yılmaz was born in Giresun on July 09, 1987. She received her B.Sc. degree in Computer Science in 2009 from Kocaeli University. She is currently a CRM Application Development Specialist at Akbank. Her research interests include data mining, pattern recognition, machine learning and CRM applications.

PUBLICATIONS

1. N. Yılmaz, G.I. Alptekin, 2013, "The Effect of Clustering in the Apriori Data Mining Algorithm: A Case Study", Proceedings of the 2013 International Conference of Data Mining and Knowledge Engineering (ICDMKE 2013), 3-5 July 2013, London (**Best Student Paper**).

2. N. Yılmaz, G.I. Alptekin, 2013, "An ontology-based data mining approach for strategic marketing decisions", Proceedings of the 9th International Symposium on Management, Engineering and Informatics (MEI 2013), 9-12 July 2013, Orlando.