SOCIAL MENTALITY AND RESEARCHER THINKERS JOURNAL
Open Access Refereed E-Journal & Refereed & Indexed
ISSN: 2630-631X
Social Sciences Indexed
www.smartofjournal.com / editorsmartjournal@gmail.com
December 2018
Article Arrival Date: 21.11.2018 Published Date:24.12.2018 Vol 4 / Issue 14 / pp:1186-1190

FP-GROWTH ALGORİTMASI- WEKA UYGULAMASI

FP-GROWTH ALGORITHM- WEKA APPLICATION

Öğr. Gör. Serpil SEVİMLİ DENİZ
Van Yüzüncü Yıl Üniversitesi, sdeniz@yyu.edu.tr, Van,Türkiye

ÖZET

Veri madenciliği, farklı formatlarda saklanan büyük hacimli verilerdeki bilgi keşfine ilişkin çeşitli yaklaşımları nedeniyle araştırmacılar için geniş bir alandır. Veri madenciliği tekniklerini uygulanarak, bilinmeyen örüntülerin keşfi ve veriler arasındaki ilişkiler keşfedilir. Veri madenciliğinin farklı işlevleri temel alarak sınıflandırma, kümeleme, özellik seçimi ve birliktelik kuralı madenciliği olarak sınıflandırılmıştır. Birbirleriyle ilişkili verilerden özelliklerin çıkarılması ve aralarındaki ilişkilerin büyüklüğünün tespit edilmesine yönelik çalışmalar birliktelik kuralları olarak tanımlanır. En fazla kullanılan birliktelik kuralı algoritmaları Apriori ve FP-Growth algoritmalarıdır. Bu çalışmada amaç genel olarak sepet market analizlerinde kullanılan bu kural çıkarımı yöntemlerinin farklı çalışmalarda da kullanımını göstermektir. Bu amaçla yüksekokul öğrencilerinin sosyal medya kullanım eğilimleri araştırması için yapılan çalışma Fp-Growth algoritması ile modellenmiştir. Model sonucunda öğrencilerin mahremiyet sorunları olmasına rağmen, sosyal medya araçlarında paylaşım yapmaktan kaçınmadıkları görülmektedir. Öğrenciler üzerinde aile ve toplum baskısının hissedilir bir etkisi olduğu gözlenmiştir.

Anahtar Kelimeler: Birliktelik Kuralları, Veri madenciliği, Fp-Growth Algoritması, Sosyal medya kullanımı

ABSTRACT

Data mining is a wide field for researchers because of their various approaches to information discovery in large volumes of data stored in different formats. By applying data mining techniques, the discovery of unknown patterns and the relationships between the data are discovered. It is classified as classification, clustering, feature selection and association rule mining based on different functions of data mining. The rules of association are aimed at identifying the characteristics of interrelated data and determining the magnitude of the relationships between them. The most commonly used association rule algorithms are Apriori and FP-Growth algorithms. The aim of this study is to show the usage of these rule extraction methods used in basket market analysis in different studies. For this purpose, the study of social media usage trends of high school students was modeled with Fp-Growth algorithm. As a result of the model, it is seen that although students have privacy problems, they do not refrain from sharing in social media tools. It was observed that family and community pressures had a significant effect on students.

Key Words: Association rules, Data mining, Fp-Growth Algorithm, Social media usage

1. GİRİŞ

Verilerin hızla artmasının sonucu olarak bu verilerin yorumlanması veri madenciliği tekniklerinin gelişmesini sağlamıştır (Silahtaroğlu, 2008). Veriler arasındaki bağlantıları arayan denetimsiz veri madenciliği şekli birliktelik kuralları olarak da tanımlanabilir. Geçmiş veriler analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır (Han ve Kamber, 2001). Birliktelik kuralında, müşterilerin alışveriş esnasında satın aldıkları ürünler arasındaki birliktelik-ilişki bağları bulunarak, müşterilerin satın alma alışkanlıklarının tespit edilmesi sağlanır. Keşfedilen bu birliktelik-ilişki bağıntıları sayesinde satıcılar daha etkin ve kazançlı satışlar yapabilme imkanına sahip olmaktadırlar (Li vd., 2007).

Birliktelik Kurallarının Matematik modeli birliktelik kuralının matematiksel modeli 1993 yılında Agrawal, Imielinski ve Swami tarafından tanımlanmıştır. Bu modele göre;

$I = (i_1, i_2, ..., i_m)$ nesnelerin kümesi ve D işlemler kümesi olarak ifade edilir. Her *i*, bir nesne olarak adlandırılır. *D* veri tabanında her işlem *T*, $T \subseteq I$ olacak şekilde tanımlanan nesnelerin kümesi olsun. Her işlem bir tanımlayıcı alan olan TID ile temsil edilir. A ve B nesnelerin kümeleri olsun. Bir T işlemler kümesi ancak ve ancak $A \subseteq T$ ise yani *A, T*'nin alt kümesi ise A'yı kapsıyor denir. Birliktelik kuralı $A \Rightarrow B$ şeklinde ifade edilir. A önce ve B sonuç olarak ifade edilir. $A \subseteq I \quad ve \quad A \cap B = \varnothing$ dir.

$X \Rightarrow Y$ eşleştirme kuralı kullanıcı tarafından minimum değeri belirlenmiş güvenirlik (c) ve destek (s) eşik değerlerini sağlayacak biçimde üretilir. c güvenirlik ölçütü ve s destek ölçütü iliştirilir ve biçimsel olarak $\theta(D) = (X \Rightarrow Y, c, s)$ ile gösterilir. Burada D örneklemi; $X \Rightarrow Y$ birliktelik-ilişki kuralını; c eşik değeri, ilgili kuralın minimum güvenirliğini (X ürünlerini içeren hareketlerin en az %c oranında Y içeren hareketler kümesinde yer aldığını); s ilgili kuralın, minimum desteğini (X ve Y ürünlerini içeren hareket tutanaklarının toplam hareket tutanakları içinde en az % s oranında var olduğunu) gösterir (Agrawal vd., 1993).

Güven ve destek kavramları

Kuralın destek ve güven değerleri, kuralın ilginçliğini ve ilgililiğini ifade eden iki ölçüdür. Bu değerler sırasıyla keşfedilen kuralların yararlılığını (kullanışlılığını) ve kesinliğini (doğruluğunu) ifade eder (Osmar vd., 2001).

Birliktelik Kuralları, belli bir destek (Support) değerinin üstündeki öğeleri bulur ve bundan sonra kalan öğeler arasından belli bir güven (Confidence) üstündeki istenilen kuralları üretir. Lift ise X ve Y'nin istatistiksel olarak bağımsız olması durumda ne kadar birlikte geçtiklerini tanımlar.

$$KURAL: X \Rightarrow Y \left\{ \begin{array}{c} \sup port = \dfrac{frq(X,Y)}{N} \\ c = \dfrac{frq(X,Y)}{frq(X)} \\ lift = \dfrac{\sup port}{\sup(X) \times \sup(Y)} \end{array} \right\}$$

Bu ölçüm değerleri basit bir örnekle anlatalım. Bir markette, 1000 satış olsun. X ürünü 80 satışta, Y ürünü 100 satışta ve X-Y ürünleri beraber 20 satışta satılmış olsun. Bu durumda Support:

Support (X) = 80/1000 = 8%, Support(Y) = 100/1000 = 10%, Support(X, Y) = 20/1000 = 2%

Confidence: Y ürünü 100 satış olduğuna göre X ürününe bu grup içinde ne kadar rastlayabiliriz. En fazla 20 olduğu için Confidence(X) = 20/100 = 10%

Lift = Support(X,Y) / Support(X) * Support(Y) = (20/1000) / ((80/1000)x(100/1000)) = 2.5

2. FP-GROWTH BİRLİKTELİK ALGORİTMASI

FP-Growth algoritması iki aşamadan oluşur: FP Ağacı'nın yapımı ve FP Ağacı'ndan sık kullanılan kalıpların çıkarılması. FP-Ağacı'nın yapısı veritabanı üzerinde iki tarama gerektirir. İlk tarama, F listesinin oluşturulması için daha sonra azalan sıraya göre sıralanan sık öğeleri seçer. İkinci tarama, FP-Ağacı' nı oluşturur. Öncelikle, işlemler sık olmayan öğeler kaldırılarak F listesine göre yeniden sıralanır. Ardından yeniden düzenlenen işlemler FP Ağacı'na eklenir. FP-Growth girişi, FP-Ağacı ve minimum destek sayısıdır. FP-Growth, FP-Ağacı'ndaki düğümleri F listesinde en az bulunan öğeden ayırır. Her bir düğümü ziyaret ederken, FP-Growth, yoldaki öğeleri düğümden ağacın köküne toplar. Bu öğeler, o maddenin koşullu kalıp tabanını oluşturur. Koşullu kalıp tabanı, öğe ile birlikte meydana gelen küçük bir desen veritabanıdır. Sonra FP-Growth oluşturulur. Koşullu desen tabanından küçük FP-Ağacı ve FP Ağacı üzerinde FP-Growth yürütülür. Koşullu desen tabanı oluşturulmadan işlem tekrarlamalı olarak yinelenir (Iko ve Masuri, 2003).

3. YÖNTEM

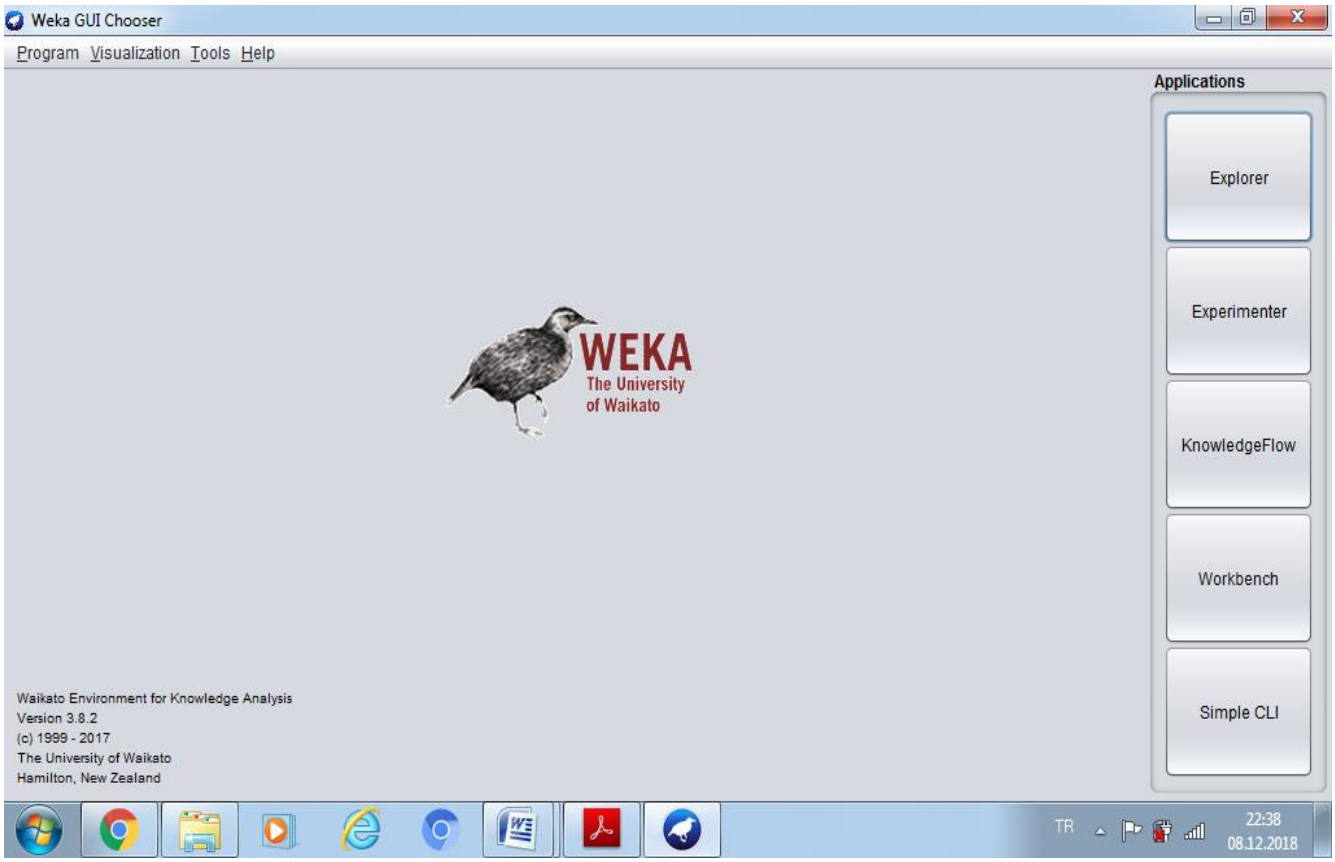
3.1. Veri Seti

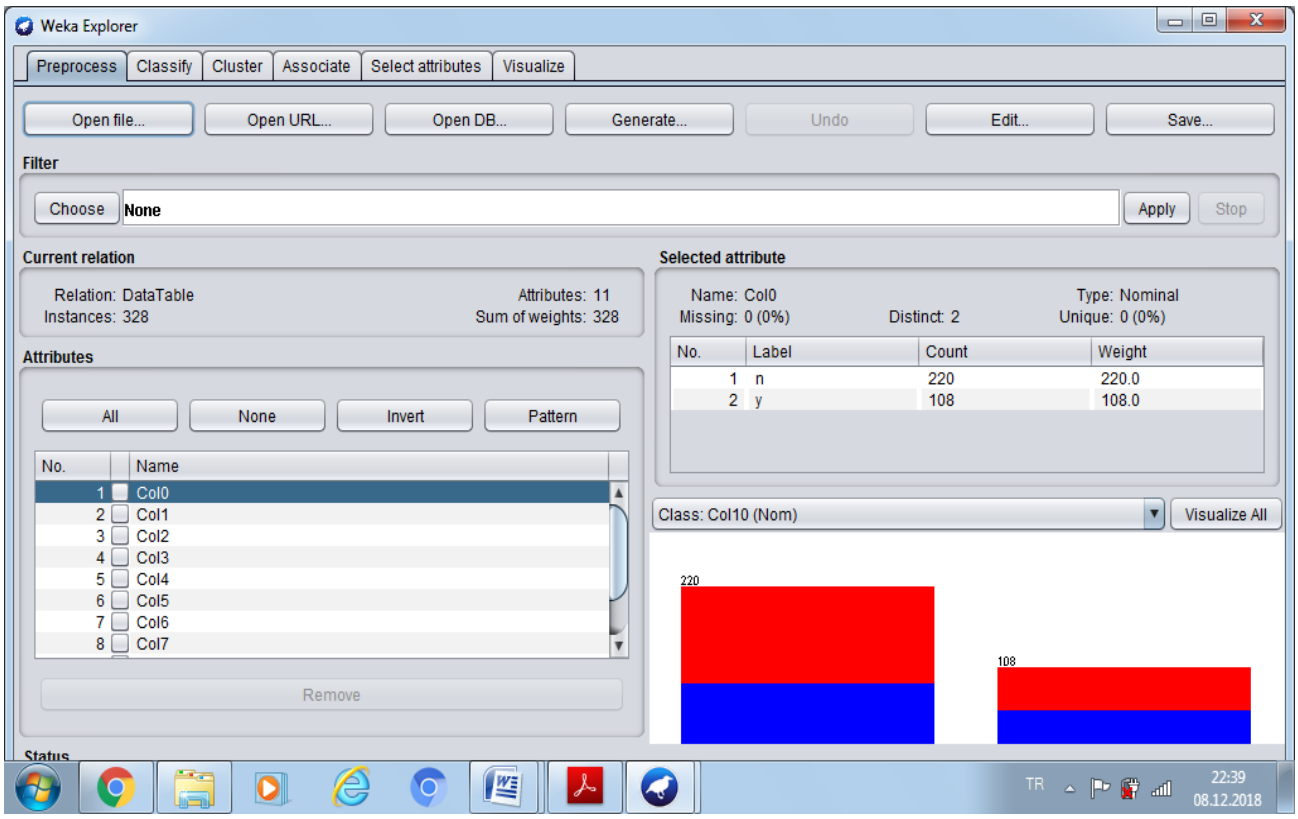
Araştırmanın çalışma grubunu Yüksekokul öğrencileri oluşturmaktadır. On bir sorudan oluşan anketin örneklem sayısı 329 dur. Sorulara verilen cevaplar cinsiyet, kullanılan sosyal medya aracı ve

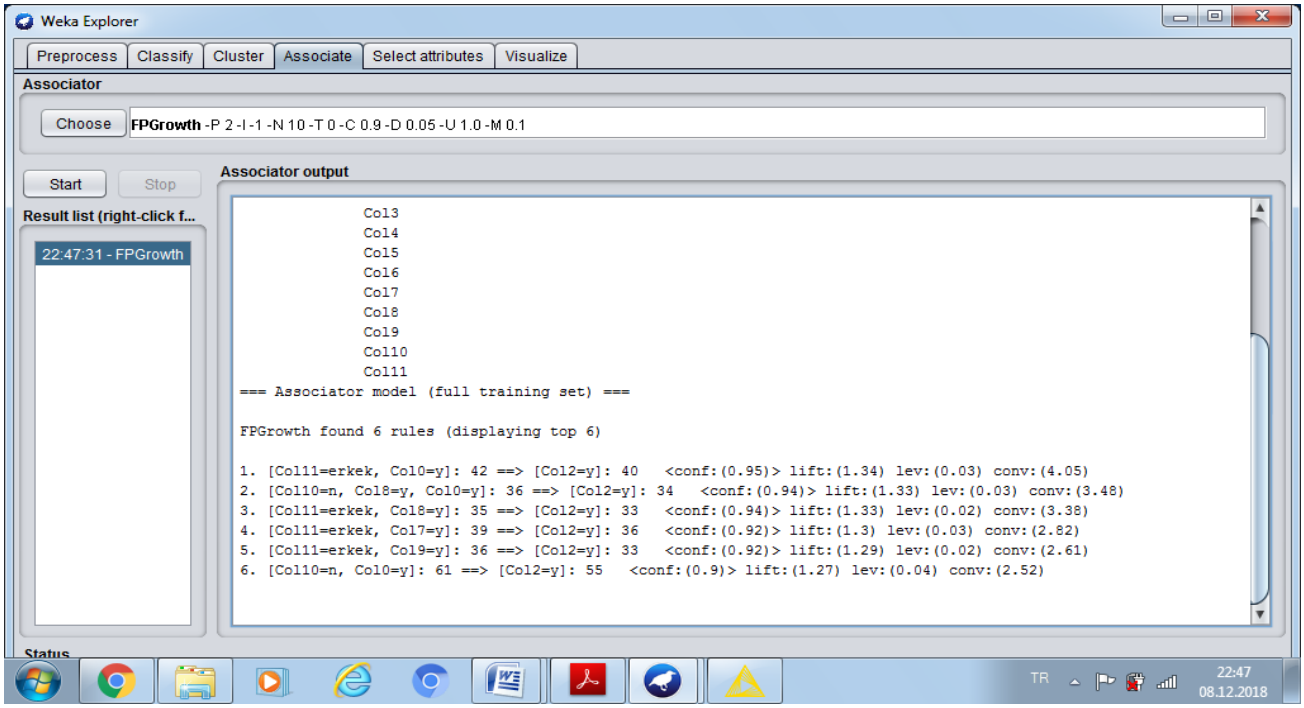
on bir sorudan oluşan sosyal medya kullanım eğilimi anketi ile değerlendirilmiştir. Anket sonuçları uygun formata dönüştürülerek, WEKA üzerinde FP-Growth algoritması kullanılarak analiz edilmiş ve ankete dair kurallar elde edilmiştir. WEKA, Waikato Üniversitesinde geliştirilmiş, Java ile kodlanmış, açık kaynak kodlu, makine öğrenim algoritmalarını ve metotlarını içeren bir yazılımdır. Modüler bir tasarımda olup veri madenciliği ve veri analizi gibi birçok alanda kullanılmaktadır. Üç temel veri madenciliği işlemi; sınıflandırma, kümeleme ve ilişkilendirme WEKA ile yapılabilir (Anonim, 2018).

Kullanılan Anket:

SOSYAL MEDYA KULLANIM ANKETİ	EVET	HAYIR
1. Sosyal paylaşım sitelerinden (Facebook, tweeter, instagram … gibi) özel bilgilerimi paylaşırım.		
2. Sosyal medyada mahremiyetle (gizli olma durumu, gizlilik) ilgili sorunlar olduğunu düşünüyorum.		
3. Yeni tanıdığım bir kişiyi sosyal medya hesaplarımdan takip ettiğim veya onları hesabıma eklediğim olmuştur.		
4. Sosyal medya üzerinden yaptığınız paylaşımlar sizde beklenti oluşturuyor mu?		
5. Ailenizle ilgili özel anları-fotoğrafları veya durumları sosyal medyada paylaşır mısınız?		
6. Tanımadığın kişilerin, günlük özel hayatınız hakkında sosyal medya üzerinden bilgi sahibi olmasından rahatsız olur musun?		
7. Sosyal medyada yaptığınız paylaşımlarla ilgili dönütler (like, retweet, paylaşım, yorum, beğenme vb.) önemser ve ya kontrol eder misiniz?		
8. Ailem çevrem ve toplum baskısı olmasaydı kendimle ilgili şu an yaptığım paylaşımlardan daha fazlasını yapardım.		
9. Yediğim, içtiğim beğenerek giydiğim vb. şeyleri sosyal medyadan paylaşmaktan rahatsız olmam.		
10. Aile, toplum, çevre vb. olmasaydı daha seküler (dini ya da ruhani olmayan) bir yaşam tarzı benimser ve sosyal medya paylaşımlarımı da o doğrultuda yapardım.		
11. Sosyal medya hesaplarınız üzerinden yaptığınız paylaşımlar, sizi tedirgin ettiği oluyor mu?		







4. BULGULAR VE TARTIŞMA

=== Associator model (full training set) ===

FPGrowth found 6 rules (displaying top 6)

=== Associator model (full training set) ===

FPGrowth found 6 rules (displaying top 6)

1. [Col11=erkek, Col0=y]: 42 ==> [Col2=y]: 40 <conf:(0.95)> lift:(1.34) lev:(0.03) conv:(4.05)

2. [Col10=n, Col8=y, Col0=y]: 36 ==> [Col2=y]: 34 <conf:(0.94)> lift:(1.33) lev:(0.03) conv:(3.48)

3. [Col11=erkek, Col8=y]: 35 ==> [Col2=y]: 33 <conf:(0.94)> lift:(1.33) lev:(0.02) conv:(3.38)

4. [Col11=erkek, Col7=y]: 39 ==> [Col2=y]: 36 <conf:(0.92)> lift:(1.3) lev:(0.03) conv:(2.82)

5. [Col10=n, Col0=y]: 61 ==> [Col2=y]: 55 <conf:(0.9)> lift:(1.27) lev:(0.04) conv:(2.52)

Kuralların yorumları: Önem sırasına göre;

1. Hem sosyal paylaşım sitelerinden özel bilgilerimi paylaşırım hem de sosyal medyada mahremiyetle (gizli olma durumu, gizlilik) ilgili sorunlar olduğunu düşünüyorum diyen erkek öğrencilerin birliktelik oranı en yüksektir.
2. Sosyal medya hesapları üzerinden yaptığı paylaşımlardan tedirgin olmayan ve yediği, içtiği beğenerek giydiği şeyleri sosyal medyadan paylaşmaktan rahatsız olmayan ve yeni tanıdığı kişileri sosyal medya hesaplarından takip ettiğim veya onları hesabıma eklediğim olmuştur diyenlerin oranı %94 bulunmuştur.
3. Ailem çevrem ve toplum baskısı olmasaydı kendimle ilgili şu an yaptığım paylaşımlardan daha fazlasını yapardım ve yeni tanıdığım bir kişiyi sosyal medya hesaplarımdan takip ettiğim veya onları hesabıma eklediğim olmuştur diyen erkek öğrencilerin oranı %94 tür.
4. Aile, çevre ve toplum baskısı olmasaydı kendimle ilgili şu an yaptığım paylaşımlardan daha fazlasını yapardım ve aile, toplum, çevre vb. olmasaydı daha seküler (dini ya da ruhani olmayan) bir yaşam tarzı benimser ve sosyal medya paylaşımlarımı da o doğrultuda yapardım diyen erkek öğrencilerin birlikteliği %92 bulunmuştur.
5. Sosyal medya hesapları üzerinden yaptığı paylaşımlardan tedirgin olmayan ve sosyal paylaşım sitelerinden (Facebook, tweeter, instagram …gibi) özel bilgilerimi paylaşırım ve yeni tanıdığım bir kişiyi sosyal medya hesaplarımdan takip ettiğim veya onları hesabıma eklediğim olmuştur diyenlerin birlikteliği %90 bulunmuştur.

5. SONUÇ

Bu araştırma kapsamında Yüksekokul öğrencilerinin sosyal medya araç kullanımları ile sosyal medya eğilimleri Fp-Growth algoritması ile incelenmiştir. Öğrencilerin mahremiyet sorunları olduğunu bilmelerine rağmen, sosyal medya araçlarında paylaşım yapmaktan kaçınmadıkları görülmektedir. Öğrenciler üzerinde aile ve toplum baskısının hissedilir bir etkisi olduğu, hiç tanımadıkları insanlarla bile günlük yaşamı – özel yaşamı ile ilgili paylaşımlar yapmaktan kaçınmadıkları tespit edilmiştir. Bu çalışma ile birliktelik kuralı algoritmalarının sepet market analizi dışında kullanılabilirliği gösterilmiştir.

KAYNAKLAR

Agrawal, R., Imielinski, T., Swami, A., (1993), "Mining association rules between sets of items in large databases", ACM SIGMOD Conference on Management of Data, Washington.

Anonim, http://www.cs.waikato.ac.nz/ml/weka/ (*Erişim tarihi: 21th of April, 2018*).

Han J., Kamber M., (2001), Data Mining: Concepts and Techniques. Morgan Kaufmann.

Iko Pramudiono and Masaru Kitsuregawa (2003). Parallel FP-Growth on PC Cluster, In PAKDD.

Li Liu, Eric Li, Yimin Zhang, and Zhizhong Tang (2007). Optimization of frequent itemset mining on multiple-core processor.

Osmar R. Zaïane, Mahammad El-Hajj, and Paul Lu.(2001). Fast Parallel Association Rule Mining without Candidacy Generation.

Silahtaroğlu G. (2008), Kavram ve Algoritmalarıyla Temel Veri Madenciliği, 1. Baskı, Papatya Yayıncılık.