T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ZAMANA BAĞLI İŞBİRLİKÇİ FİLTRELEME

ÇİĞDEM BAKIR

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ.DR.SONGÜL ALBAYRAK

İSTANBUL, 2014

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ZAMANA BAĞLI İŞBİRLİKÇİ FİLTRELEME

Çiğdem BAKIR tarafından hazırlanan tez çalışması 20.06.2014 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi _____

Doç. Dr. Şule Gündüz ÖĞÜDÜCÜ

İstanbul Teknik Üniversitesi _____

Yrd. Doç. Dr. Mehmet AKTAŞ

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Tavsiye sistemlerini inceleyen ve tavsiye sistemlerinde kullanıcılara özgü en doğru ürün önerisinde bulunmak için öneri metotlarının uygulanma sürecini içeren bu tez çalışması Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü Yüksek Lisans programında yürütülmüştür.

Bu tezin hazırlanması süresi boyunca, bilgisi, deneyimi ve önerileriyle araştırma ve geliştirmelerimi yönlendiren tez danışmanım Sayın Doç. Dr. Songül ALBAYRAK'a sonsuz teşekkürü borç bilirim.

Bilgisayar Mühendisliği Bölümündeki bütün değerli hocalarıma ve arkadaşlarıma katkılarından ve desteklerinden dolayı teşekkür ederim.

Bu tezi, maddi ve manevi hiçbir fedakârlıktan kaçınmayıp desteklerini hiçbir zaman esirgemeyen sevgili aileme ithaf ediyorum.

Haziran, 2014

Çiğdem BAKIR

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

ÇİZELGE LİSTESİ x

ÖZET xiii

ABSTRACT xv

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti 1
1.2 Tezin Amacı 4
1.3 Hipotez/Orijinal Katkı 5

BÖLÜM 2

TAVSİYE SİSTEMLERİ 5

2.1 İşbirlikçi Filtreleme 6
2.1.1 Hafıza Tabanlı İşbirlikçi Filtreleme 8
2.1.1.1 Kullanıcı Tabanlı İF Tekniği 8
2.1.1.2 Nesne Tabanlı İF Tekniği 11
2.1.1.3 Hibrit Yaklaşımlı İF Algoritması 12
2.1.2 Model Tabanlı İF Tekniği 15
2.1.2.1 Simple Bayesian İF Tekniği 15
2.1.2.2 Kümeleme Yöntemi ile İF Tekniği 16
2.1.2.3 Latent Semantic Yöntemiyle İF Tekniği 16
2.2 İçerik Bazlı Filtreleme 16

2.3 Hibrit Yöntemler 17

BÖLÜM 3

İF TEKNİKLERİNDE OLASI PROBLEMLER VE DEĞERLENDİRME ÖLÇÜTLERİ 18

3.1 Yeni Kullanıcılar(Cold User) 18
3.2 Yeni Nesneler(Cold item) 18
3.3 Ölçeklenebilirlik 19
3.4 Eşanlamlılık 19
3.5 Değerlendirme Ölçütleri 20
3.5.1 Mean Absolute Error(MAE) 20
3.5.2 Normalized Mean Absolute Error(NMAE) 20
3.5.3 Root Mean Squared Error(RMSE) 20

BÖLÜM 4

BOYUT İNDİRGEME YÖNTEMLERİ 22

4.1 Singular Value Decomposition(SVD) 22
4.2 Tavsiye Sistemlerinde SVD Kullanımı 23
4.2.1 Nesne Tabanlı İF Tekniklerinde SVD Kullanımı 23

BÖLÜM 5

ZAMANA BAĞLI YÖNTEM 25

5.1 Puanların Yaşını Kullanarak Yapılan Zamana Bağlı İF Tekniği 25

BÖLÜM 6

İF TEKNİKLERİNDE KULLANILAN VERİ KÜMELERİ 31

6.1 Netflix Veri Kümesinin Analiz Edilmesi 31
6.2 MovieLens Veri Kümesinin Analiz Edilmesi 33

BÖLÜM 7

DENEYSEL ÇALIŞMA 36

7.1 Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği 36
7.1.1 Az Sayıda Filme Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği 36
7.1.2 Çok Sayıda Filme Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği 39
7.2 Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği 42
7.2.1 Az Sayıda Puan Verilmiş Filmler için Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği 43
7.2.2 Çok Sayıda Puan Verilmiş Filmler için Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği 47
7.3 Klasik ve Zamana Bağlı Hibrit İF Tekniği 49
7.3.1 Az Sayıda Puan Verilmiş Filmler ve Az Sayıda Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Hibrit Tekniği 50

7.3.2 Çok Sayıda Puan Verilmiş Filmler ve Çok Sayıda Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Hibrit Tekniği 53
7.4 MovieLens Verisine SVD Uygulanması 56
7.5 Netflix Veri Kümesi İçin Uygulama Sonuçları 65

BÖLÜM 8

SONUÇ VE ÖNERİLER 72

KAYNAKLAR 74

ÖZGEÇMİŞ 81

SİMGE LİSTESİ

Age_{nm}	n kullanıcısının m ürününe vermiş olduğu puanın yaşı
Age_r	Mevcut puanın yaşı
Age_{w}	Mevcut puanın ağırlıklandırılmış yaşı
$p_{a,i}$	a kullanıcısının i nesnesine verdiği tahmini puan
r_{aged}	Zamana bağlı olarak ağırlıklandırılmış puan
$\overline{r_a}$	a aktif kullanıcısının tüm nesnelere verdiği puanların ortalaması
$r_{a,i}$	a aktif kullanıcının i nesnesine verdiği puan
$\overline{R_\iota}$	i nesnesine verilen puanların ortalaması
r_{org}	Mevcut puan
$\overline{r_u}$	u kullanıcısının tüm nesnelere verdiği puanların ortalaması
σ_a	a kullanıcısının nesnelere verdiği puanların standart sapması
$w_{a,u}$	a ve u kullanıcısı arasındaki benzerlik
w_{ij}	i ve j nesnesi arasındaki benzerlik

KISALTMA LİSTESİ

CF	Collaborative Filtering
İF	İşbirlikçi Filtreleme
LSI	Latent Semantic Indexing
MAE	Mean Absolute Error
NB	Naive Bayes
PCA	Principial Component Analysis
RMSE	Root Mean Squared Error
SVD	Singular Value Decomposition

ŞEKİL LİSTESİ

Sayfa

Şekil 5. 1 Puanların yaşını ağırlıklandırmak için kullanılan dönüşüm fonksiyonları.... 28
Şekil 6. 1 Netflix veri kümesinde müşterilerin yıllara göre verdikleri puan sayıları.... 32
Şekil 6. 2 Netflix veri kümesinde müşterilerin puan verdikleri film sayıları 33
Şekil 6. 3 MovieLens veri kümesinde film sayılarına göre puan veren kullanıcı sayısı.35
Şekil 7. 1 Netflix veri kümesinde zamana bağlı kullanıcı tabanlı İF tekniğinin sonuçları (az oylama yapan kullanıcılar için)……39
Şekil 7. 2 Netflix veri kümesinde zamana bağlı kullanıcı tabanlı İF tekniğinin sonuçları (çok oylama yapan kullanıcılar için)……42
Şekil 7. 3 Netflix veri kümesinde zamana bağlı nesne tabanlı İF tekniğinin sonuçları (az oylanmış filmler için)……46
Şekil 7. 4 Netflix veri kümesinde zamana bağlı nesne tabanlı İF tekniğinin sonuçları (çok oylanmış filmler için)……49
Şekil 7. 5 Netflix veri kümesinde zamana bağlı hibrit yönteminin sonuçları (az puan veren kullanıcılar ve az oylanmış filmler için)……52
Şekil 7. 6 Netflix veri kümesinde zamana bağlı hibrit yönteminin sonuçları (çok puan veren kullanıcılar ve çok oylanmış filmler için)……56
Şekil 7. 7 MovieLens veri kümesinde rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımı ……58
Şekil 7. 8 MovieLens veri kümesinde çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımı……60
Şekil 7. 9 MovieLens veri kümesinde tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımı ……62
Şekil 7.10 MovieLens veri kümesinde klasik nesne tabanlı İF ve k=12 optimum değerinin komşulukları alınarak elde edilen MAE sonuçları……63
Şekil 7.11 Zamana bağlı nesne tabanlı İF tekniğinin sonuçları (netflix verisi için)……..69
Şekil 7.12 Netflix veri kümesinde klasik nesne tabanlı ve k=35 optimum değerinin komşulukları alınarak elde edilen RMSE sonuçları ……71

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2. 1 Örnek bir kullanıcı-nesne matrisi 7
Çizelge 2. 2 İF'de kullanılan kullanıcı-nesne matrisi örneği[1] 8
Çizelge 2. 3 Kullanıcı tabanlı İF tekniği için kullanıcı-ürün matrisine bir örnek 9
Çizelge 2. 4 Kullanıcılar tarafından 1-5 arasında puan verilmiş kullanıcı-nesne matrisine bir örnek[3]……13
Çizelge 2. 5 Kullanıcılar tarafından 1-5 arasında puan verilmiş kullanıcı-nesne matrisine bir örnek ……..…14
Çizelge 5. 1 Kullanıcı-ürün matrisindeki puanların test örneğine göre yaşlarını gün olarak gösteren tablo ……27
Çizelge 5. 2 n=0.25 için bulunan dönüşüm katsayıları 29
Çizelge 5. 3 n=0.5 için bulunan dönüşüm katsayıları 29
Çizelge 5. 4 n=0.75 için bulunan dönüşüm katsayıları 30
Çizelge 6. 1 Netflix veri kümesinde yıllara göre verilen toplam puan sayısı 31
Çizelge 6. 2 Netflix veri kümesinde kullanıcıların puan verdikleri film sayıları..... 34
Çizelge 7. 1 Netflix veri kümesinde az sayıda filme puan veren kullanıcılar grubu için klasik kullanıcı tabanlı İF'nin puan tahmin hatası ……36
Çizelge 7. 2 Netflix veri kümesinde az sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……..37
Çizelge 7. 3 Netflix veri kümesinde az sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……..38
Çizelge 7. 4 Netflix veri kümesinde az sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……..38
Çizelge 7. 5 Netflix veri kümesinde çok sayıda filme puan veren kullanıcılar için klasik kullanıcı tabanlı İF'nin tahmin hatası ……40
Çizelge 7. 6 Netflix veri kümesinde çok sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……..40
Çizelge 7. 7 Netflix veri kümesinde çok sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası …….41
Çizelge 7. 8 Netflix veri kümesinde çok sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası …….41
Çizelge 7. 9 Netflix veri kümesinde az sayıda puan verilen filmler için klasik nesne tabanlı İF'nin tahmin hatası ……..43

Çizelge 7. 10 Netflix veri kümesinde az sayıda puan verilmiş filmler için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası …….44
Çizelge 7. 11 Netflix veri kümesinde az sayıda puan verilmiş filmler için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası ……45
Çizelge 7. 12 Netflix veri kümesinde az sayıda puan verilmiş filmler için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası …….45
Çizelge 7. 13 Netflix veri kümesinde çok sayıda puan verilmiş filmler için klasik nesne tabanlı İF'nin puan tahmin hatası ……..47
Çizelge 7.14 Netflix veri kümesinde çok sayıda puan verilmiş filmler için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası …….47
Çizelge 7.15 Netflix veri kümesinde çok sayıda puan verilmiş filmler için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası …….48
Çizelge 7.16 Netflix veri kümesinde çok sayıda puan verilmiş filmler için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen paun tahmin hatası …….48
Çizelge 7. 17 Netflix veri kümesinde az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar için klasik hibrit İF tekniğinin puan tahmini……..50
Çizelge 7.18 Netflix veri kümesinde az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……51
Çizelge 7.19 Netflix veri kümesinde az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……51
Çizelge 7.20 Netflix veri kümesinde az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……52
Çizelge 7. 21 Netflix veri kümesinde çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar için hibrit İF tekniğinin puan tahmin hatası……. 53
Çizelge 7 .22 Netflix veri kümesinde çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……54
Çizelge 7.23 Netflix veri kümesinde çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……54
Çizelge 7.24 Netflix veri kümesinde çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası ……55
Çizelge 7.25 MovieLens veri kümesinde rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(düşük k için)…….. 57
Çizelge 7.26 MovieLens veri kümesinde rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(yüksek k için)…….57
Çizelge 7.27 MovieLens veri kümesinde çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(düşük k değerleri için)……..59
Çizelge 7.28 MovieLens veri kümesinde çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(yüksek k değerleri için).……59
Çizelge 7.29 MovieLens veri kümesinde tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(düşük k değerleri için).……..61

Çizelge 7.30 MovieLens veri kümesinde tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımının MAE sonuçları(yüksek k değerleri için).....61
Çizelge 7.31 MovieLens veri kümesinde k=12 optimum değeri için nesne tabanlı İF tekniğinin tüm komşulukları alınarak elde edilen MAE sonuçları.....62
Çizelge 7.32 MovieLens veri kümesinde tüm test verisi için klasik nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen MAE sonuçları 63
Çizelge 7.33 MovieLens veri kümesinde tüm test verisi için k değerlerine bağlı olarak puan tahmininde bulunurken geçen süre(düşük k değerleri için).....64
Çizelge 7.34 MovieLens veri kümesinde tüm test verisi için k değerlerine bağlı olarak puan tahmininde bulunurken geçen süre(yüksek k değerleri için).....64
Çizelge 7.35 MovieLens veri kümesinde tüm test verisi için klasik nesne tabanlı İF tekniklerinde puan tahmininde bulunurken geçen süre 65
Çizelge 7.36 Netflix veri kümesinde klasik nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen RMSE sonuçları 66
Çizelge 7.37 Netflix veri kümesinde klasik nesne tabanlı İF ile SVD kullanımı..... 66
Çizelge 7.38 Netflix veri kümesinde n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası 67
Çizelge 7.39 Netflix veri kümesinde n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası 68
Çizelge 7.40 Netflix veri kümesinde n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası 68
Çizelge 7.41 Netflix veri kümesinde $\mathrm{Age}_{\mathrm{w}}$ [0.5.....1.1] için SVD sonuçları 70
Çizelge 7.42 k=35 için komşulukları alınarak elde edilen puan tahmin hatası 70

ÖZET

ZAMANA BAĞLI İŞBİRLİKÇİ FİLTRELEME

Çiğdem BAKIR

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç.Dr.Songül ALBAYRAK

Dünyada sanayi ve teknolojinin gelişmesiyle birlikte web üzerinden alışveriş yapan kullanıcı sayısı da her geçen gün artmaktadır. Bu hızlı artıştan dolayı firmalar gelişmelere ayak uydurarak kârını arttırabilecek, rekabet ortamından kazançlı çıkmayı sağlayabilecek yeni fikirlere ihtiyaç duyar. Firmaların sosyal ve ekonomik yönden başarıya ulaşmasını sağlayan yapılar tavsiye sistemleridir.

Tavsiye sistemleri (recommendation systems) ürün ve hizmetlerin etkileşimi sırasında kişiselleştirilmiş öneriler yapmak için bilgiyi keşfeden teknikleri uygulayan sistemlerdir. Günümüzde müşteriye özgü tavsiyelerde bulunmak, müşterilerin ilgisini canlı tutmak, müşterilerin beğenisini tahmin edip onlara bu beğenileri doğrultusunda önerilerde bulunmak önemlidir. İşbirlikçi Filtreleme-İF (Collaborative filtering-CF) tavsiye sistemleri tarafından kullanılan etkili ve kolay bir yöntemdir. Genel anlamda İF veri kaynakları arasında çeşitli teknikler kullanarak bilgiyi filtreleme sürecidir. Ayrıca İF günümüzde çok farklı alanlarda örneğin; çoklu sensörlerde veriyi algılama, izleme, finansal verilerde, e-ticaret ve web uygulamalarında kullanılmaktadır.

İF müşterilere ilgileri doğrultusunda çeşitli tavsiyeler sunmakla beraber müşteri tercihlerinin ve ürünlerin sürekli değişmesi sonucunda birtakım zorluklarla karşı karşıya kalmaktadır. Müşterilerin algısını, ürünlerin popülerliğini zamanla değiştiren yeni ürünler ve hizmetler ortaya çıkabilir. Mevsimsel değişiklikler ya da bölgesel faktörler kişinin algısını değiştirebilir [1]. Örneğin; kullanıcılar film tavsiye sistemlerinde bir aktör ya da bir film türüne göre yeni bakış açısı kazanabilir ya da zaman içinde tercih ettiği film

türünü değiştirebilir. Geleneksel İF teknikleri çok sayıda kullanıcı ve nesnelerden oluşan veri kümeleri için kullanıcıların geçmişte verdikleri puanlara bakarak gelecekteki eğilimini, kullanıcıların birbirine olan benzerliklerinden (user similarity) ya da ürünlerin birbirine olan benzerliklerinden (item similarity) yola çıkarak tahmin etmeye çalışır. Ancak bu İF teknikleri müşterilerin zamanla değişen isteklerine cevap vermekte yetersiz kalır. Ayrıca her ürün ve her müşteri farklı değişimler geçirir. Kullanıcıların davranışlarını belirleyebilecek, zaman değişimlerini gösterecek, her ürün ve her müşteri için ortak bir modele ihtiyaç duyulur. Bu çalışmayla birlikte kullanıcıların zevkine, beğenisine uygun kişiselleştirilmiş tavsiyeler sağlamak için zamana bağlı olarak kullanıcıların ilgisi analiz edilerek kullanıcıya verilen tavsiyelerin doğruluğu ve niteliği arttırılır.

Bu çalışmada özgün olarak kullanıcıların sadece puanlarına göre değil de puan verdiği tarihlere de bakılarak müşterinin o ürüne vereceği puan tahmin edilmeye çalışılmıştır. Bir ürün ne kadar eskiyse o kadar fazla kişi tarafında puan verilmiştir. Ancak kişinin alışkanlıkları zamanla değişeceği için geçmişte verdiği puan yerine günümüze yakın zamanda verdiği puan daha önemlidir. Diğer bir deyişle kullanıcılar tarafından her ürüne verilen puanın bir yaşı vardır. Bu nedenle geleneksel İF tekniğindeki eksikleri gidermek adına bu çalışmada mevcut puanlar, bu puanların yaşına göre zayıflatılması (puanlar yaşlı ise) veya kuvvetlendirilmesi (puanlar genç ise) planlanmıştır. Bu çalışmada geliştirilen yöntemlerin başarısını ölçmek amacıyla netflix verisi üzerinde son iki yılda toplanan puanlar üzerinden işlem yapılmıştır [1]. Netflix Amerika'da film ve video kiralama üzerine kurulmuş bir şirkettten alınan veri kümesidir ve tavsiye sistemlerinde sıklıkla kullanılır. Bu çalışmada düşünülen yöntem yaklaşık yarım milyon Netflix müşterisinin 1999'dan 2005'e kadarlık geçen sürelerde 17770 film için verdikleri puanların tarihleri belli olan gerçek veri kümesi üzerinde yapılmıştır. Bu veri kümesinde 100 milyon üzerinde puan verilmiş kayıt vardır. Sistem başarısını ölçmek için geleneksel İF ve zaman katılarak oluşturulacak yöntemle elde edilecek sonuçlar karşılaştırılmıştır. Ayrıca boyut indirgeme metotlarından Singular Value Decomposition (SVD) hem movielens hem de netflix veri kümesine uygulanarak sistemin başarısı ölçülmüştür. MovieLens veri kümesi Minnesota Üniversitesinde GropLens tarafından toplanan verilerden oluşur. 943 kullanıcının 1682 filme verdiği puanlardan oluşur. Kullanıcılara ait yaş, cinsiyet, meslek gibi basit demografik bilgilerin yanı sıra film türlerine ait kategorik bilgiler mevcuttur. Ancak bu veri kümesinde tarihsel bilgi mevcut olmadığı için hem zamana bağlı İF tekniğini hem de SVD tekniğini birleştirmek için netflix veri kümesi kullanılmıştır.

Geliştirilen sistem sayesinde kullanıcıların zamanla değişen ilgi ve zevklerine bağlı olarak sonuç üretebilen bir ürün öneri sistemi tasarlanmıştır. Böylece kullanıcının bu sisteme olan güvenini arttırmak hedeflenmiştir. Bu bir e-ticaret sitesi ile entegre edilerek müşterinin ilgisi canlı tutulabilir ve böylelikle firma kârlılığı arttırılabilir.

Anahtar Kelimeler: Veri Madenciliği, Tavsiye Sistemleri, İşbirlikçi Filtreleme

ABSTRACT

COLLABORATIVE FILTERING WITH TEMPORAL DYNAMICS

Çiğdem BAKIR

Department of Computer Engineering

MSc. Thesis

Adviser: Assoc.Prof.Dr.Songül ALBAYRAK

Recommendation systems are the systems implementing techniques discovering information in order to make personalized recommendations during product and service interaction. Today, it is important to make custom tailored recommendations, keep customer interest, predict customer preferences and make recommendations accordingly. Collaborative filtering-CF is an effective and easy method used by recommendation systems. Generally CF is the process of filtering information using various techniques among the information sources. In addition, CF is used in a differential area for example; such as identification of data in multiple sensors, data detection, financial data, e-commerce and web applications.

Together with offering various recommendations to its customers in accordance with their interests CF meets several difficulties upon the continuous change of the preferences of the customers and the products. New product and services may emerge that change the perception of the customers and the popularity of the products. Seasonal changes or the regional factors may change the perception of the person [1]. For example; user may gain a new point of view in movie recommendation systems according to the type of actor or movie and they can also change the type of movie they prefer in time. Traditional CF techniques tries to estimate the tendency of the users by considering the points they gave in the past for the data sets consisting of numerous user and elements starting from the similarity of the users to each others (user

similarity) or the similarity of the products to each others (item similarity). However, these CF techniques remain incapable to respond to the demands of the customers that change in time. In addition every product and customer undergoes different changes. A common model is required that will determine the behaviors of the users, show their time changes for every product and customer. Together with this study the accuracy and the attribute of the recommendations given to the user by analyzing the interest of the users depending on time in order to provide suitable personalized recommendations to the pleasure and appreciation of the users are increased.

In this study, the point that customer given the point to that product had tried to be estimated not only according to the points of the users originally, but considering the dates they gave the points. The older a product the more it was given points by that many persons. However, since the habits of the person would change in time, the point they give close to today is more important than the point they gave in the past. In other words, there is the age of the point given for each product by the users. Therefore, on behalf of correcting the deficiencies in traditional CF technique the existing points in this study had planned to decreased according to the age of the points (if the points are old) or strengthened (if the points are fresh). In this study, action had taken over the data of Netflix through the points collected in the last two years in order to measure the success of the methods developed. Netflix is a data set taken from a company established based on movie and video rental in America and used frequently in recommendation systems. The method considered in this study will be performed over the natural data set whose point dates are certain that were given by half a million Netflix customers from 1999 to 2005 for 17770 movies. In addition, in this data set there are records over 100 million that were given points. Moreover, SVD of size reduction methods was measured system performance by applying both netflix and movielens to the data sets. MovieLens data set consists of data collected by GroupLens at the Minnesota University. It is composed 1682 film's of scores given 943 users. It is available user's age, gender, occupation as well as demografic information such simple categorical information on the film types. In order to measure the system success, the results that obtained by the method to be formed by traditional CF and adding time had compared.

Due to the system to be developed, a product recommendation system had designed that could produce result depending on the interest and the pleasures of the users that could change in time. Thus, the confidence of the user to the system will increase. If this is an e-commerce site, the company profitability will be increased by keeping the interest of the customer live because of recommending the proper product.

Keywords: Data Mining, Recommendation Systems, Collaborative Filtering

BÖLÜM 1

GİRİŞ

Geleneksel İF tekniklerinde kullanıcılara önerilerde bulunurken veri kümesinde kullanıcıların ürünlere vermiş olduğu puanlar kullanılmıştır. Ancak önerilen zamana bağlı İF tekniğinde mevcut puanların yaşı sisteme eklenmiştir. Puanların yaşı sisteme çeşitli dönüşüm fonksiyonlarıyla ağırlıklandırılarak kullanılmıştır. Bu ağırlıklandırmada mevcut puanlar gençse kuvvetlendirilmiş; yaşlıysa zayıflatılmıştır. Ayrıca boyut indirgeme yöntemleriyle zamana bağlı İF teknikleri birleştirilerek kullanıcılara daha hızlı öneriler sunulmuştur.

Yapılan bu çalışmada kullanıcıların zamana bağlı isteklerine ve zevklerine uygun öneriler verilmiştir. Kullanıcıların zamana bağlı olarak değişen ihtiyaçlarına uygun öneri sistemi geliştirilmiştir.

1.1 Literatür Özeti

Tavsiye sistemleri kullanıcıların nesnelere verdiği puanı tahmin edip kullanıcılara özgü nesne önerisinde bulunan bilgi filtreleme sistemlerinin bir alt sınıfıdır. Günümüzde film, müzik, video, haber, kitap, sosyal ağlar, web sayfaları gibi birçok alanda yaygın bir şekilde kullanılmaktadır.

Tavsiye sistemleri İşbirlikçi Filtreleme-İF ve İçerik Tabanlı Filtreleme (content-based filtering) olmak üzere iki şekilde tavsiye üretir. İF kullanıcıların geçmiş davranışlarına, diğer kullanıcılarla olan etkileşimine ya da ürün benzerliğine bakarak tavsiyelerde bulunur. İçerik tabanlı filtreleme ise kullanıcıların yaş, ırk, meslek ya da yaşadığı yer gibi demografik bilgilerine ya da nesneleri çeşitli kategorilerine ayıran karakteristik özelliklerine bakarak tavsiyelerde bulunur.

Tavsiye sistemleriyle ilgili birçok çalışma yapılmıştır. Goldberg ve Nichols tarafından yapılan Tapestry tavsiye sistemleri alanındaki en eski çalışmadır [2]. Tapestry kural tabanlı oluşan, kullanıcılara özgü bir mail sistemidir. Kullanıcıların okudukları dokümanlarla ilgili tepkilerini kaydederek mail sistemindeki bilgiyi filtreler. Hem İF 'yi hem de içerik tabanlı filtrelemeyi destekler. Ancak kullanıcıların çok fazla olması ve bütünleşik sorguların çalıştırılmamasından dolayı çeşitli zorluklarla karşı karşıya kalır.

Resnick, Suchak ise sosyal ağlarla ilgili Tapesty'e benzer bir çalışma gerçekleştirmiştir. Elektronik bilgileri filtreleyerek kullanıcılara sevdikleri yazıları sunmuşlardır [3].

Çeşitli benzerlik ölçütleriyle yapılan bellek tabanlı İF algoritmalarının eksik yönlerini çözmek amacıyla Bayesian, kümeleme, sınıflandırma ve Singular Value Decomposition (SVD) gibi boyut indirgeme yöntemlerine dayalı model tabanlı İF algoritmaları geliştirilmiştir[4]. Miyoharo ve Pazzani kullanıcıların nesnelere vermiş oldukları puanları negatif ve pozitif olarak ayırmışlardır ve Jester ve EachMovie verisi üzerinde Simple Bayesian tekniği uygulayarak elde ettikleri sonuçları değerlendirmişlerdir[5]. Ungar ve Foster , Gibbs örnekleme ve k-means kümeleme tekniklerini çeşitli veri kümelerine uygulayarak sonuçları karşılaştırmışlardır[6]. Rongfe ve Maozhogn , DBSCAN kullanarak kümeleme işlemi gerçekleştirmiştir[7]. Su ve Khosgoftar gerçek veri kümeleri üzerinde yaptıkları çalışmada hafıza tabanlı İF (memory based collaborative filtering) yaklaşımlarının puan tahmininde model tabanlı İF yaklaşımından daha başarılı olduğunu ancak ölçekleme problemlerinin çözümünde ise model tabanlı İF yaklaşımlarının daha başarılı olduğunu göstermişlerdir[5].

Pernock ve Hervitz geleneksel benzerlik hesaplamalarından farklı olarak Personality Diagnosis (kişilik tanıma) algoritmasını geliştirmişlerdir. Personality Diagnosis (PD) algoritması bellek ve model tabanlı İF tekniklerinin birleştirilmesiyle oluşturulmuştur [8]. PD algoritmasında kullanıcılardan rastgele seçilerek aktif kullanıcı üretilmiştir. PD melez İF yaklaşımıdır ve bu yaklaşımla kullanıcıya daha doğru önerilerde bulunulur [9]. Hem yeni kullanıcı ve yeni nesne problemini aşmak hem de ölçekleme problemini çözmek amacıyla hibrit yaklaşım kullanılmıştır.

Tavsiye sistemlerinde çok büyük veri kümeleri kullanılmaktadır. İF tekniklerinde oluşturulan kullanıcı-ürün matrisinde kullanıcılar tarafından puan verilmemiş çok fazla

ürün olabilir ya da birçok ürün çok az kullanıcı tarafından puanlanmış olabilir. Bu veri seyrekliği problemine sebep olmaktadır. Su ve Khoshgoftar veri seyrekliği problemini aşmak için model tabanlı İF algoritmaları kullanmışlardır [10]. Ayrıca Greinemr ve Su sistemdeki herbir örneğin sınıfını doğru tahmin edilmesini sağlamak için Bayesian belief net sınıflandırıcıları kullanmışlardır [11].

Landaver, Litmand SVD temeline dayanan kullanıcılar arasındaki gizli etkileşimleri görmek ve bilgi çıkarımında bulunmak için LSI(Latent Semantic Indexing) tekniğini kullanmışlardır[12,13]. Ayrıca Goldberg, Roeder, Gupta ve Perkins online sistemlerde işlem zamanını azaltıp kullanıcılara daha hızlı ve doğru önerilerde bulunmak için SVD'ye benzer olarak temeli PCA'a dayanan "eigentaste" metodunu geliştirmişlerdir. Bu çalışmada sistemdeki belli kullanıcılar ve belli nesneler atılarak boyut indirgenmiştir.

Group Lens ilk İF sistemleriyle ilgili çalışmalar yapan araştırma laboratuvarıdır. Çeşitli benzerlik ölçütleriyle kullanıcı benzerliği ya da nesne benzerliği kullanarak kullanıcılara tahminlerde (prediction) bulunmuştur [3]. Linden, Smith ve York Amazon.com üzerinde tekrarlamalı ürün benzerliği kullanmışlardır [14]. Ziegler, Lausen Last.fm verisi üzerinde yaptıkları çalışmada şarkılara verilen etiket verisinden yararlanılarak kullanıcılara dinleyecekleri şarkılar hakkında önerilerde bulunmuşlardır [15].

Pavlov ve Pennock veri seyrekliği ve yüksek veri boyutuyla başa çıkabilmek için dokümanları kümeleme yoluna gitmişlerdir [16]. Ayrıca sisteme yeni giren kullanıcılar ve nesneler ile ilgili tahminlerde bulunmak için de çeşitli çalışmalar yapmışlardır. Nhat Lam, Vu Duc Le yeni kullanıcı ve yeni nesne problemine çözüm bulabilmek için İF teknikleriyle içerik tabanlı filtreleme yaklaşımını birleştirmiştir. Burke yaptığı çalışmada içerik, demografik ve İF yaklaşımlarını kullanmıştır [17].

Goldberg, Roeder, Gupta ve Perkins online sistemlerde işlem zamanını azaltıp kullanıcılara daha hızlı ve doğru önerilerde bulunmak için Jester verisi üzerinde eigentaste algoritmasını geliştirmiştir[18]. Eigentaste algoritması bir grup nesne üzerinde kullanıcı puanları üzerinde PCA uygulayan bir İF algoritmasıdır. Bu çalışmalarla birlikte yeni kullanıcı ve yeni nesne problemine kısmen çözüm sağlanmıştır.

Yapılan çalışmaların hiçbirinde sistemdeki zamana bağlı değişikleri gösterecek çözümler bulunmamaktadır. Yapılan bu çalışmaların hepsi tavsiye sistemlerindeki çeşitli

problemleri kısmen çözümlemiştir. Ancak yapılan çalışmaların hiçbirinde kullanıcıların zamana bağlı değişen isteklerine , ihtiyaçlarına , beğenilerine, zevklerine uygun öneriler verilmemiştir. Kullanıcının zevkleri, ürünlerin değişikleri göz önüne alınarak yapılmış bir çalışma mevcut değildir. Bu da her kullanıcı için önceki verdiği puanların günümüze olan uzaklığına bakılmaksızın aynı şeylerin önerilmesine neden olacaktır. Mevcut çözümler zaman parametresini dikkate almadığı için kişinin isteklerine cevap vermekte yetersiz kalmaktadır. Bu eksiklikleri giderebilecek akılcı, kişiye doğru öneride bulunacak bir sistem gereklidir. Önerilen projeyle birlikte mevcut çalışmalardaki eksiklikler giderilecek ve daha doğru öneri sistemi geliştirilecektir. Bu çalışmayla birlikte hem hızlı hem de daha doğru öneri sistemi geliştirilmiştir.

1.2 Tezin Amacı

Günümüzde tavsiye sistemleri önemli bir yer edinmektedir. Her ürünün bir alıcısı vardır. Ancak doğru alıcı ile doğru ürünü buluşturmak zor bir uğraştır. Özellikle internet üzerinden yapılan satışlarda bu önemli bir sorun olarak karşımıza çıkmaktadır. Çünkü müşterilerin istekleri, beğenileri ve zevkleri sürekli değişmektedir. Geleneksel İF teknikleri hızla değişen ürünlere ve müşterilerin zamanla değişen beğenilerine, ilgilerine cevap vermekte yetersiz kalmaktadır. Müşteriler internet üzerinden alışveriş yaptıklarında kendi isteklerine uygun öneriler yapılmasını ister.

Tezin amacı kullanıcıların gelişen isteklerine, zamana bağlı olarak değişen zevk ve beğenilerine göre akılcı çözümler geliştirip her kullanıcının ihtiyacına uygun zaman parametresini de kullanan ortak bir tavsiye sistemi geliştirmektir. Bu sistem e-ticaret, telekom, reklamcılık, medya, müzik siteleri gibi çok geniş alanlarda müşterinin zevkine, ihtiyacına uygun tavsiyeler vererek daha fazla kişiye hitap edecek alanlar oluşturacaktır.

İstenilen hedefler aşağıda maddelenmiştir:

- Bu sistemde milyonlarca müşteri ve ürünün bulunduğu çok büyük veri kümeleriyle çalışılmaktadır. Yapılan uygulama ile terabayta kadar varan çok büyük veri kümeleri üzerinden müşteriye özgü ve müşterinin **zamanla değişen** zevk ve beğenilerine göre ürün önerisi yapılacaktır.

- Yeni müşteriler(cold user) ve yeni ürünler (cold item) ile ilgili sistemde geçmişe dönük veri olmadığı halde bunların sisteme entegrasyonu sağlanacaktır.
- Gerekli bellek büyüklüğü ve işlemci zamanı minimuma indirilecektir.
- Klasik İşbirlikçi Filtreleme-İF algoritmaları ile bu projede önerilen zamana bağlı İF çözümleri arasında karşılaştırmalar yapılacaktır.

1.3 Hipotez/Orijinal Katkı

Xiang , Yuan ve Zhao kullanıcıların tercihlerini belirlemek için CiteUlike ve Delicious verisi üzerinde grafik yöntemi geliştirmişlerdir [19]. Ancak yapılan çalışmada kullanıcılar arasındaki etkileşimi belirlemek oldukça zordur.

Luo, Cai ve Chowdhury tarafından matris MovieLens verisi üzerinde kullanıcıların tercihlerini kendi kendine eğitmek için matris faktorizasyonuna dayalı bir yöntem geliştirmişlerdir [20]. Ancak tavsiye sistemlerinde kullanılan veriler genelde eksik ve azdır. Bu sebeple istenilen sonuca ulaşılamamışlardır.

Shrot, Rosenfeld ve Golbeck sınıflandırma yöntemine dayalı CRISP algoritmasını geliştirmişlerdir [21]. Kullanıcılar arasındaki benzerliği bulmak için bilinen kullanıcıları kullanmışlardır. Choi ve Suh hedef nesnenin komşuluklarını seçen yeni bir benzerlik fonksiyonu kullanmışlardır [22]. Ancak yapılan bu çalışmaların hiçbirinde kullanıcıların puan verdikleri tarihlerin etkisi sistemde kullanılmamıştır. Yapılan çalışmayla kullanıcıların puan verdiği tarihe uygun olarak kullanıcının puanı tahmin edilip kullanıcıya özgü öneriler sunulması sağlanmıştır.

Bu çalışmada geleneksel ve yapılan diğer zamana bağlı İF çalışmalarından farklı bir çalışma yapılmıştır. Mevcut puanların yaşları çeşitli dönüşüm fonksiyonları kullanarak sisteme eklenmiştir. Bu dönüşüm fonksiyonlarında mevcut puanın yaşı yaşlıysa zayıflatılarak; gençse kuvvetlendirilerek puanlar ağırlıklandırılmıştır. Ağılıklandırılmış puan sisteme eklenerek daha doğru ve daha güvenilir bir öneri sistemi geliştirilmiştir. SVD ile zamana bağlı İF tekniği birleştirilerek kullanıcılara daha hızlı öneriler verilmiştir. Kullanıcıların ihtiyaçlarını, isteklerini belirleyebilecek etkili bir yöntem geliştirilmiştir. Ayrıca büyük veri kümelerinde karşılaşılan veri seyrekliği ve ölçekleme problemlerine de çözüm bulunmuştur.

BÖLÜM 2

TAVSİYE SİSTEMLERİ

Tavsiye sistemleri çok büyük veri içerisinde kişilere uygun bilgileri getiren bir bilgi çıkarım tekniğidir. Tavsiye sistemlerindeki asıl amaç insanların beğenisini tahmin edip onlara bu beğenileri doğrultusunda tavsiyelerde bulunmaktır. Hem ticari hem de araştırma alanında veri madenciliği ve makine öğrenmesinde aktif olarak kullanılır. Bu alanda en yaygın kullanılan yöntemler;

- İşbirlikçi Filtreleme-İF
- İçerik Bazlı Filtrelereme(content-based filtering)
- Hibrit Yaklaşımlar

2.1 İşbirlikçi Filtreleme

Teknolojinin hızla gelişmesi, internet kullanımının yaygınlaşması, web içeriklerinin sürekli artması kullanıcılara onları ilgilendirecek kişiselleştirilmiş ürün ve servis sunma ihtiyacı doğurmuştur. Özellikle de e-ticaretin gelişmesiyle birlikte tüm şirketler müşterilerine web üzerinden hizmet sunmaya başlamıştır. Ancak ürün ve servislerin artması web üzerinden alışveriş yapan kullanıcıların seçimini zorlaştırır. Örneğin; web sayfası üzerinden kitap siparişinde bulunan bir kullanıcı sipariş ettiği kitaba benzer kitap önerilerinin gelmesini ister. Ya da web üzerinden müzik videolarına bakan bir kişi kendi dinlediği videolara uygun önerilerde bulunmasını bekler[23]. Kendi zevklerine, beğenilerine uygun öneriler gelmediği takdirde o siteden bir daha alışveriş yapamayabilir ya da müzik dinlemeyebilir. Bu da kullanıcı memnuniyetsizliğinin bir göstergesidir. Böyle bir durumda İşbirlikçi Filtreleme-İF kavramına ihtiyaç duyulur. İF kullanıcıların ilgilerine,

beğenilerine uygun otomatik tahminler yapılmasını sağlar ve tavsiye sistemlerinde sıklıkla kullanılır. Büyük veri kümeleri içerisinde istenilen bilgiye erişmek için bilgi çıkarımında bulunur. Kullanıcı-nesne matrisi İF'de kullanılır ve kullanıcıların ilgilendikleri nesnelere verdikleri puanı gösterir. Çizelge 2.1 'de örnek bir kullanıcı-nesne matrisi gösterilmiştir ve İF'nin amacı "?" işaretiyle gösterilen ifadeyi tahmin etmektir.

Çizelge 2.1 Örnek bir kullanıcı-nesne matrisi

Kullanıcılar	**Shrek**	**Snow White**	**Spider-man**	**Super-man**
Alice	Like	Like	-	Dislike
Bob	-	Like	Dislike	Like
Chris	-	Dislike	Like	-
Tony	Like	-	Dislike	?

Çizelge 2.1'de ? işaretiyle gösterilen değeri bulmak için kullanılan İF algoritmalarının genel çalışma prensibi aşağıdaki gibidir:

- Aktif kullanıcının (Tony) bir nesneye vermiş olduğu puanı bulmak için sistemdeki bütün kullanıcıların tarihsel geçmişlerine bakılır.
- Aktif kullanıcının diğer kullanıcılar ile olan benzerliği dikkate alınır.
- Aktif kullanıcının diğer kullanıcılar ya da bulunmak istenen diğer nesneler arasındaki benzerlik korelasyon, kosinüs benzerliği ya da diğer benzerlik ölçütleriyle ölçülür.
- Benzerlik ölçütleri kullanarak aktif kullanıcının nesneye vermiş olduğu puan tahmin edilmeye çalışılır.

İki çeşit İF algoritması vardır:

- Hafıza tabanlı İF (Memory based collaborative filtering)
- Model tabanlı İF (Model based CF)

2.1.1 Hafıza Tabanlı İF

İşbirlikçi Filtreleme-İF algoritmaları bir tahmin üretmek için kullanıcı ve nesneden oluşan veritabanının tümünü ya da belirli bir kısmını kullanır. Her kullanıcı benzer ilgilere sahip kullanıcı grubunun bir parçasını oluşturur [24]. Bir kullanıcının benzer komşularını belirleyerek bu kullanıcının yeni nesnelere olan ilgileri benzer kullanıcı grubunun ilgileri doğrultusunda hesaplanabilir[Çizelge 2.2]. İşbirlikçi filtreleme kullanıcı-nesne matrisinde eksik değerin bulunması için kullanılır. Bu kullanıcı-nesne matrisinde kullanıcıların izledikleri filmlere verdikleri puanlar yer alır ve Nathan kullanıcısının Titanic filmi için vereceği puan tahmin edilmeye çalışılır[25]. Aktif kullanıcı ? işareti ile gösterilen kullanıcıdır. İF kolay uygulanabilen, etkili bir yöntemdir ve temelde üç çeşit hafıza tabanlı İF algoritması vardır:

- Kullanıcı tabanlı İF
- Nesne tabanlı İF
- Hibrit yaklaşımlı İF algoritması

Çizelge 2.2 İF'de kullanılan kullanıcı-nesne matrisi örneği [1]

	Star Wars	**Hoop Dreams**	**Contact**	**Titanic**
Joe	5	2	5	4
John	2	5		3
Al	2	2	4	2
Nathan	5	1	5	?

2.1.1.1 Kullanıcı Tabanlı İF Tekniği

Kullanıcı tabanlı İF tekniğinde kullanıcılar arasındaki benzerlik hesaplanarak ($w_{a,u}$) aktif kullanıcı (a) için tahminler üretilir. Çizelge 2.3 de kullanıcı-ürün matrisi gösterilmiştir. İF tekniklerindeki amaç "?" işaretiyle gösterilen aktif kullanıcının (a) i_3 ürününe vermiş olduğu puanı tahmin etmektir. Kullanıcı tabanlı bu yöntemde aktif kullanıcının tüm

kullanıcılara olan benzerliği hesaplanır. İki kullanıcı arasındaki (a,u) benzerlik , ürünler üzerindeki uzaklık ya da korelasyonu ile ölçülür.

Çizelge 2.3 Kullanıcı tabanlı İF tekniği için kullanıcı-ürün matrisine bir örnek

Kullanıcılar/ Ürünler	$\boldsymbol{i_1}$	$\boldsymbol{i_2}$	$\boldsymbol{i_3}$			$\boldsymbol{i_m}$
$\boldsymbol{u_1}$	$r_{1,1}$		$r_{1,3}$			
$\boldsymbol{u_2}$	$r_{2,1}$		$r_{2,3}$			
$\boldsymbol{u_3}$	$r_{3,1}$		$r_{3,3}$			
a	$r_{a,1}$		?			
$\boldsymbol{u_n}$	$r_{n,1}$					$r_{n,m}$

Kullanıcılar arasındaki benzerliği hesaplayan birçok metot vardır. Bu metotlardan yaygın kullanılan iki tanesi aşağıdaki gibidir:

Korelasyon Tabanlı Benzerlik Hesabı: Pearson korelasyon kullanıcı tabanlı İF metotlarında yaygın olarak kullanılan bir benzerlik hesabıdır [22]. İki kullanıcı arasındaki benzerlik ($w_{a,u}$) ya da iki nesne arasındaki benzerlik Pearson korelasyon ya da diğer korelasyon tabanlı benzerlik metotlarıyla bulunur [25]. İki kullanıcı arasındaki Pearson korelasyon benzerlik hesabı;

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a}) * (r_{u,i}-\overline{r_u})}{\sigma_a \sigma_u} \quad (2.1)$$

ile bulunur.

Bu ifadede **m** toplam nesne sayısı, $r_{a,i}$ **a** aktif kullanıcının i nesnesine verdiği puanı, $r_{u,i}$ ise **u** kullanıcısının **i** nesnesine verdiği puanı, $\overline{r_a}$ **a** aktif kullanıcısının tüm nesnelere verdiği puanların ortalamasını, $\overline{r_u}$ **u** kullanıcısının tüm nesnelere verdiği puanın ortalamasını , $\sigma_a ve \sigma_u$ ise a ve u kullanıcılarının nesnelere verdiği puanların standart sapmasını gösterir [25].

Kosinüs Vektör Tabanlı Benzerlik: İki kullanıcı arasındaki benzerlik hesaplanırken her kullanıcı ve her nesne bir vektör olarak kabul edilir ve bu kullanıcılar arasındaki kosinüs açısı hesaplanır. Kosinüs benzerliği;

$$w_{a,u}=\cos(a, u)= \frac{\vec{a}\vec{u}}{\|a\| * \|u\|} \quad (2.2)$$

şeklinde hesaplanır.

Kullanıcı Tabanlı İF Sistemlerinde Puan Tahmini

Kullanıcı tabanlı İF tekniğinde kullanıcılar arasındaki benzerlik bulunduktan sonra bir puan tahmini ($p_{a,i}$) yapılır. Bu aşama kullanıcı tabanlı İF tekniklerinde en önemli adımı oluşturur. **a** aktif kullanıcısının **i** nesnesine verdiği puan hesaplanırken **a** ve **u** kullanıcıları arasındaki benzerlik $w_{a,u}$ kullanılır. Eşitlik 2.3 'te **n** kullanıcı sayısını gösterir.

$$p_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r_u}) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \quad (2.3)$$

Kullanıcı tabanlı İF uygulamalarında aktif kullanıcının tüm kullanıcılara olan benzerliği bulunurken aktif kullanıcının binlerce komşusu ile benzerliği hesaplanır. Ancak gerçek e-ticaret sistemlerinde aktif kullanıcının milyonlarca komşusu vardır ve aralarında benzerlik hesaplanır. Başka bir ifadeyle, kullanıcı tabanlı algoritmalar sistemde çok fazla puan veren kullanıcılar için performansı yavaşlatır. Çok fazla kullanıcı için benzerlik hesabı yapılacağından benzer kullanıcıları belirlerken arama hızı düşer. Ayrıca var olan kullanıcı ve nesnelerin sayısının çok artması hesaplama karmaşıklığını da beraberinde getirir. Kullanıcıya tavsiyelerde bulunmak için kullanıcı tabanlı İF tekniği tek başına yeterli olmaz. Ölçekleme problemini aşmak ve gerçek zamanlı performans sağlamak için nesne tabanlı İF tekniği geliştirilmiştir.

2.1.1.2 Nesne Tabanlı İF Tekniği

Nesne tabanlı yaklaşımda aktif kullanıcının puan verdiği i nesnesine en benzeyen **k** adet nesne için $\{i_1 , i_2 , ... , i_k \}$ benzerlikler $\{w_{i1} , w_{i2} , ...,w_{ik} \}$ hesaplanır. $w_{i,j}$ **i** nesnesinin **j** nesnesine olan benzerliğini gösterir. Nesne tabanlı İF algoritmalarının en önemli adımı benzer öğeleri belirlemek ve en yakın nesneleri gruplamaktır. Benzerliği hesaplamada temel fikir **i** ve **j** nesnelerinin her ikisine de puan veren kullanıcıları belirleyip, bu kullanıcılara puan veren nesneler arasındaki ($w_{i,j}$) benzerliği hesaplamaktır. Nesne tabanlı İF tekniklerinde birçok benzerlik yöntemi kullanılmaktadır:

Korelasyon Tabanlı Benzerlik: i ve j nesneleri arasındaki benzerlik genellikle Pearson korelasyonu ile bulunur. Hem i hem de j nesnesine puan vermiş kullanıcılar belirlenir ve $w_{i,j}$ benzerlik hesabı bu kullanıcılar için aşağıdaki gibi yapılır:

$$w_{i,j}=\frac{\sum_{u\in U}(r_{u,i}-\overline{r_i})\,(r_{u,j}-\overline{r_j})}{\sqrt{\sum_{u\in U}\,(r_{u,i}-\overline{r_i})^2}\,\sqrt{\sum_{u\in U}\,(r_{u,j}-\overline{r_j})^2}} \qquad (2.4)$$

Bu ifade de **U** tüm kullanıcıları , **u** hem **i** hem **j** nesnesine puan veren kullanıcıları , $r_{u,i}$ **u** kullanıcısının **i** nesnesine verdiği puanı , $\overline{r_i}$ **i** nesnesine verilen puanların ortalamasını, $\mathrm{r_{u,j}}$ **u** kullanıcısının **j** nesnesine verdiği puanı, $\overline{r_j}$ ise **j** nesnesine verilen puanların ortalamasını gösterir[26].

Ayarlanmış Kosinüs Benzerliği: Kullanıcı tabanlı algoritmada (İF-K) satırlar arasındaki benzerlik hesaplanırken , nesne tabanlı algoritmada (İF-N) ise sütunlar arasındaki benzerlik hesaplanır [27].

$$w_{i,j}=\frac{\sum_{u\in U}(r_{u,i}-\overline{r_u})\,(r_{u,j}-\overline{r_u})}{\sqrt{\sum_{u\in U}\,(r_{u,i}-\overline{r})^2}\,\sqrt{\sum_{u\in U}\,(r_{u,j}-\overline{r_u})^2}} \qquad (2.5)$$

Nesne Tabanlı İF Sistemlerinde Puan Tahmini

İF sistemlerinde en önemli adım aktif kullanıcının herhangi bir nesne için tahmini puanını hesaplamaktır. Nesne tabanlı İF algoritmasında ise **a** aktif kullanıcısının **i** nesnesine vermiş olduğu puanı hesaplamak için **i** öğesine benzer nesneler $w_{i,j}$ile kullanılan nesneye verdiği puanlar $r_{u,N}$ çarpılıp, $w_{i,j}$ değerlerinin toplamına bölünür[27].

$$p_{a,i} = \frac{\sum_{allsimilaritems,N}(w_{i,N} * r_{a,N})}{\sum_{allsimilaritems,N}(\,I w_{i,N} I)} \quad (2.6)$$

İF-N teknikleri İF-K ile karşılaştırıldığında kullanıcı sayısının fazla olduğu veri kümelerinde daha iyi sonuçlar verir. Genellikle web sitelerinde nesne sayısı kullanıcı sayısından daha fazladır. Kullanıcılar arasındaki benzerlikten yola çıkarak bir tahmin değeri hesaplanmaya çalışıldığında çok fazla işlem süresi ve hafıza gerektirir. Web üzerinden alışveriş yapılan online sistemlerde kullanıcılar isteklerine hemen cevap verilmesini ister. Bu sebeple İF-N tekniğine ihtiyaç duyulur. Linden, Smith ve York Amazon.com'da çeşitli kategorilere ayrılan nesneleri tekrarlamalı nesne tabanlı İF tekniği uygulayarak gerekli bellek büyüklüğünü ve işlemci zamanını minimuma indirmişlerdir [14]. Böylelikle kullanıcılara kısa sürede ve daha doğru tavsiyeler sunmuşlardır.

2.1.1.3 Hibrit Yaklaşımlı İF Algoritması

Hibrit hem kullanıcı hem de nesnelere göre yapılan bilgi çıkarımını kapsadığı için hibrit yaklaşım İF-K (kullanıcı tabanlı) ve İF-N (nesne tabanlı) teknikleri birleştirir [28,29,30].

Çizelge 2.4'deki örnek veri kümelerinde kullanıcı tabanlı İF tekniği kullanarak u_7 kullanıcısının i_6 nesnesi için verdiği puan ($r_{7,6}$) tahmin edilmeye çalışılsaydı u_4 , u_5 ve u_6 kulanıcıları u_7 kullanıcısına en benzer kullanıcılar olacaktı. Bu durumda $r_{7,6}$ puanın değerini u_7 kullanıcısına en benzer kullanıcıların tercihleri olan i_2 , i_4 ve i_5 nesneleri belirleyecekti. Çünkü bu kullanıcılar aktif kullanıcıyla aynı davranışları gösterir. Ancak Şekil 3'de de görüldüğü gibi bu nesneler futbol nesnesi ile ilgilidir ve bulunmak istenen İngilizce nesnesinden (i_6) tamamen farklıdır. Bu durumda sadece kullanıcı tabanlı İF tekniği kullanılarak u_7 kullanıcısı için oluşacak tahmin doğru ve güvenilir olmayacaktır. Diğer taraftan sadece nesne tabanlı İF tekniği kullanıldığında da aşağıda belirtilen sorunlarla karşılaşmak mümkündür.

Çizelge 2.4'deki örnek veri kümesinde nesne tabanlı İF tekniği kullanarak u_7 kullanıcısının i_6 nesnesi için verdiği puan ($r_{7,6}$) tahmin edilmeye çalışılsaydı bulunmak istenen i_6 nesnesine en benzer nesneler i_1 ve i_3 nesneleri olacaktı. Ancak bu durumda da u_4 , u_5 ve u_6 kulanıcıları bu nesnelere puan vermediği için u_7 kullanıcı için bir tahmin üretilmeyecektir. Çünkü İF veri kümelerinde çok az kullanıcının puan verdiği nesneler olabilir ya da bir kullanıcı tarafından puan verilmemiş çok fazla nesne olabilir. Bu İF veri

kümelerinde sıklıkla karşılaşılabilen bir durumdur. Bu durumda sadece kullanıcı tabanlı İF tekniğini ya da sadece nesne tabanlı İF tekniğini kullanmak tek başına yeterli olmadığından hibrit yaklaşıma ihtiyaç duyulur.

Çizelge 2.4 Kullanıcılar tarafından 1-5 arasında puan verilmiş kullanıcı-nesne matrisine bir örnek[3]

Kullanıcı/ Nesne	$\boldsymbol{i_1}$ **(İngilizce)**	$\boldsymbol{i_2}$ **(Futbol)**	$\boldsymbol{i_3}$ **(İngilizce)**	$\boldsymbol{i_4}$ **(Futbol)**	$\boldsymbol{i_5}$ **(Futbol)**	$\boldsymbol{i_6}$ **(İngilizce)**
$\boldsymbol{u_1}$	$r_{1,1}$=3	$r_{1,2}$=1	$r_{1,3}$=2	$r_{1,4}$=3	$r_{1,5}$=5	$r_{1,6}$=5
$\boldsymbol{u_2}$	$r_{2,1}$=3	$r_{2,2}$=1	$r_{2,3}$=2	$r_{2,4}$=3	$r_{2,5}$=5	$r_{2,6}$=5
$\boldsymbol{u_3}$	$r_{3,1}$=3	$r_{3,2}$=1	$r_{3,3}$=2	$r_{3,4}$=3	$r_{3,5}$=5	$r_{3,6}$=5
$\boldsymbol{u_4}$	$r_{4,1}$=1	$r_{4,2}$=5	$r_{4,3}$=3	$r_{4,4}$=3	$r_{4,5}$=1	$r_{4,6}$=1
$\boldsymbol{u_5}$	$r_{5,1}$=2	$r_{5,2}$=5	$r_{5,3}$= 2	$r_{5,4}$=3	$r_{5,5}$=2	$r_{5,6}$=1
$\boldsymbol{u_6}$	$r_{6,1}$=3	$r_{6,2}$=5	$r_{6,3}$= 1	$r_{6,4}$=3	$r_{6,5}$=2	$r_{6,6}$=1
$\boldsymbol{u_7}$	$r_{7,1}$=3	$r_{7,2}$=5	$r_{7,3}$=2	$r_{7,4}$=3	$r_{7,5}$=2	$r_{7,6}$=?

Hibrit algoritmasının adımları aşağıdaki gibidir:

- Bulunmak istenen hedef **j** nesnesinin diğer nesneler arasındaki benzerliğini gösteren sim (j,q) hesaplanır. Bu benzerlik Pearson korelasyon katsayısı ya da diğer benzerlik yöntemleriyle bulunur.

• sim (j,q) ifadesine göre hedef nesne j -SI_j için benzer nesneler belirlenir. SI_j belirlemek için iki metot vardır. İlk metot belirli bir eşik değer belirlemektir. Bu eşik değer **s** ise hesaplama benzerlik değeri **s** den daha büyük olan nesneler alınır. İkinci yöntemde ise bütün sim(j,q) içerinden belli sayıda en benzer nesneler alınır. Bu değer de SI_jile gösterilir [28].

Çizelge 2.5 Kullanıcılar tarafından 1-5 arasında puan verilmiş kullanıcı-nesne matrisine bir örnek[3]

Kullanıcı/ Nesne	$\boldsymbol{i_1}$ **(İngilizce)**	$\boldsymbol{i_2}$ **(Futbol)**	$\boldsymbol{i_3}$ **(İngilizce)**	$\boldsymbol{i_4}$ **(Futbol)**	$\boldsymbol{i_5}$ **(Futbol)**	$\boldsymbol{i_6}$ **(İngilizce)**
$\boldsymbol{u_1}$	$r_{1,1}$=3	$r_{1,2}$=1	$r_{1,3}$=2	$r_{1,4}$=3	$r_{1,5}$=5	$r_{1,6}$=5
$\boldsymbol{u_2}$	$r_{2,1}$=3	$r_{2,2}$=1	$r_{2,3}$=2	$r_{2,4}$=3	$r_{2,5}$=5	$r_{2,6}$=5
$\boldsymbol{u_3}$	$r_{3,1}$=3	$r_{3,2}$=1	$r_{3,3}$=2	$r_{3,4}$=3	$r_{3,5}$=5	$r_{3,6}$=5
$\boldsymbol{u_4}$	$r_{4,1}$= null	$r_{4,2}$=5	$r_{4,3}$= null	$r_{4,4}$=3	$r_{4,5}$=1	$r_{4,6}$=1
$\boldsymbol{u_5}$	$r_{5,1}$= null	$r_{5,2}$=5	$r_{5,3}$= null	$r_{5,4}$=3	$r_{5,5}$=2	$r_{5,6}$=1
$\boldsymbol{u_6}$	$r_{6,1}$= null	$r_{6,2}$=5	$r_{6,3}$= null	$r_{6,4}$=3	$r_{6,5}$=2	$r_{6,6}$=1
$\boldsymbol{u_7}$	$r_{7,1}$=3	$r_{7,2}$=5	$r_{7,3}$=2	$r_{7,4}$=3	$r_{7,5}$=2	$r_{7,6}$=?

•Aktif kullanıcının SI_j içerisindeki diğer kullanıcılara olan $w_{j(a,i)}$ benzerliği hesaplanır. Burdaki **j** nesne, **a** aktif kullanıcı, **i** SI_j içerisindeki diğer kullanıcılardır. $CSI_{j(a,i)}SI_j$ içerisindeki aktif kullanıcının diğer kullanıcılara olan benzerliklerinin hesaplanmış değeri ifade etmek üzere ortak olarak puan verilmiş nesneleri gösterir.

• Aktif kullanıcıya göre $w_{j(a,i)}$ benzer komşu kullanıcılar belirlenir. $Neighbor_{a,j}$ aktif kullanıcıya benzer komşu kullanıcıları ifade eder. $Neighbor_{a,j}$ hesaplamak için 2 method kullanılır:

İlk yöntem **w** mutlak kullanıcı benzerlik eşik değeri ayarlayarak **w** 'den büyük olanlar seçilir. (**w** kullanıcı korelasyon eşik değeri) İkinci yöntem best-neighbors yöntemidir. Bu yöntem de ise bütün $w_{j(a,i)}$ içerisinden en benzer kullanıcılar alınır.

•Aktif kullanıcıya benzer olan kullanıcılar $Neighbor_{a,j}$ belirlendikten sonra aktif kullanıcının j nesnesine verdiği puan $P_{a,j}$ hesaplanılır.

$$P_{a,j} = \overline{r_a} + k \sum_{i \varepsilon Neighbor_{a,j}} w_{j(a,i)} \, (r_{i,j} - \overline{r_i}) \qquad (2.7)$$

$w_{j(a,i)}$ aktif kullanıcının diğer kullanıcılara olan benzerliği (j nesnesi için) , $(r_{i,j} - \mu_i)$ kullanıcının tahmin edilmesi gereken nesneye verdiği puandan kullanıcının tüm nesnelere verdiği puandan ortalamasının farkına işaret eder.

$$\frac{1}{k} = \sum_{i \varepsilon Neighbor_{a,j}} w_{j(a,i)} \qquad (2.8)$$

• $P_{a,j}$ (j ε I_{Pa}) a aktif kullanıcısı için tüm nesnelere olan tahmin (prediction) değeri hesaplanılır. Aktif kullanıcıya tavsiyelerde (recommendation) bulunulur.

2.1.2 Model Tabanlı İF

Hafıza tabanlı İF yaklaşımlarının eksik yönlerini gidermek için model tabanlı İF algoritmaları geliştirilmiştir. Bayesian modelleri, kümeleme teknikleri, SVD gibi boyut indirgeme teknikleri model tabanlı yaklaşımlardır ve bu algoritmalar makine öğrenmesi, veri madenciliği alanında sıklıkla kullanılır.

2.1.2.1 Simple Bayesian İF

Kullanıcıların nesnelere verdiği puanlar Naive Bayes-NB yöntemi ile hesaplanılır. NB 'de kural sınıfları verilen bütün özelliklerin olasılıklarını hesaplamaktır. En büyük olasılığa sahip olan özellik aktif kullanıcının sınıfı olarak kabul edilir. Miyahara ve Pazzani kullanıcının verdiği puanları pozitif ve negatif olarak ayrı ele almıştır ve Bayes yöntemi kullanarak EachMovie, Jester verileri üzerinde kullanıcılar arasındaki benzerliği hesaplamıştır [5].

2.1.2.2 Kümelenme Yöntemi ile İF

Nesneler arasındaki benzerlik Minkowski ya da Pearson korelasyonu ile hesaplanılır. Üç çeşit kümeleme metodu kullanılır.

- Bölünme metotları (partition methods) içerisinde en yaygın kullanılan k-means kümeleme tekniğidir. Etkili ve kolay uygulanır.
- Yoğunluk tabanlı methodlar (density-based methods) nesnelerin yoğun bir şekilde ayrıldığı kümeleri bulur. DBSCAN ve OPTICS iyi bilinen yoğunluk tabanlı kümeleme methodlarıdır [31].
- Hiyerarşik methodlar verileri ayrıştırarak kümeler. En yaygın hiyerarşik methodu BIRCH 'dir.

Connar ve Herlocker EachMovie verisi üzerinde kullanıcıların vermiş oldukları puanları çeşitli kümelere ayırmışlar ve herbir kümeye k en yakın komşu algoritması kullanarak kendi içerisinde puanını hesaplamışlardır [32]. Böylelikle sistemin tahmin başarısını arttırmaya çalışmışlardır.

2.1.2.3 Latent Semantic Yöntemiyle İF

Kullanıcıların profilini bulmak için gizli sınıf değişkenlerini uygulayan istatiksel bir modeldir. Bu gizli sınıf değişkenleri kullanıcı ve nesneden oluşur ve birbirinden bağımsızdır. Gizli kalmış olan kullanıcıların etkileşimlerini, benzerliklerini bularak sistemin tahmin başarısını arttırır. SVD boyut indirgeme tekniği LSI matris ayrıştırmasını yerine getiren bir tekniktir [33]. Anlamsal olarak aynı nesneyi ifade eden özellikleri bulur.

2.2 İçerik Tabanlı Filtreleme

İçerik tabanlı İF kullanıcıların ve nesnelerin özelliklerini kullanarak kişilere tavsiyelerde bulunur. Bu yöntemler öneriler yapmak için nesnelerin özelliklerinin içeriğini analiz eder [34]. Kullanıcıların geçmişteki tercihlerine, zevklerine göre kullanıcı profili oluşturulur. Bu kullanıcı profili kullanıcıların puan verdiği nesnelerin özelliklerinden üretilir [35]. İçerik bazlı yöntemlerde kullanıcıların mevcut sistemdeki geriye dönük açıkca görülen ya da görülmeyen davranışlarına göre tavsiyeler ürettiğinden yeni kullanıcılar ve yeni ürünler için bilgi çıkarımında bulunulabilir.

İçerik tabanlı İF tekniğinde kullanıcılara ait demografi yöntemler kullanılır. Demografik yöntemler kullanıcıların yaşadığı yer, yaş, cinsiyet, eğitim gibi bilgilerine dayanır [36]. Bir kullanıcının demografik bilgisi diğer oluşturulan kümelerle karşılaştırılıp kullanıcının en uygun olduğu küme belirlenir. Ayrıca nesnelerin karakteristik özelliklerine ve kategorilerine göre de kullanıcının ait olduğu küme belirlenir.

Demografik bilgiler işbirlikçi filtreleme yöntemleriyle birleştirilerek kullanılarak verilen tavsiyelerin doğruluğunu ve niteliğini artırır [37]. Özellikle de sistemde daha önce puanları bulunmayan, geçmişe yönelik bilgisi olmayan yeni kullanıcılar ve yeni nesnelere tavsiyelerde bulunmada çok önemli bir yer tutar.

2.3 Hibrit Yöntemler

Hibrit yöntemler içerik bazlı filtreleme ile İF tekniklerinin birleştirilmesi ile oluşur. Ansari kullanıcı tercihleri, kullanıcı ve nesne özellikleri gibi bilgileri tavsiyelerde bulunmak için kullanışlı olan bilgileri birleştirerek Bayes tercih modeli oluşturulmuştur [38].

Balabonovic ve Shoham web sayfalarında içerik bazlı bilgileri kullanarak kullanıcı profiline uygun tavsiyeler geliştirmişlerdir [39].

Martinez ve Mantonegro hibrit sistemle birlikte SVD uygulayarak web üzerinden çeşitli TV programları için tavsiyelerde bulunmuşlardır [40].

BÖLÜM 3

İF TEKNİKLERİNDE OLASI PROBLEMLER VE DEĞERLENDİRME ÖLÇÜTLERİ

İF tekniklerini uygulanırken bir takım zorluklarla karşı karşıya kalınır [41]. Yeni kullanıcı (cold user), yeni nesne (cold item), ölçeklenebilirlik, aynı anlama gelen fakat farklı sözcüklerle ifade edilen kelimeleri gösteren eş anlamlılık, çok büyük veri kümeleriyle çalışılmasına rağmen verinin yetersiz olması İF tekniklerinde sıkça karşılaşılan problemlerdir.

3.1 Yeni kullanıcı (cold user)

Bu problem mevcut veri kümesi için çok az nesneye puan vermiş kullanıcıları ya da sisteme yeni giren kullanıcıları ifade eder [42]. Sistemde bu kullanıcıların davranışlarını gösterecek yeterli bilgi yoktur. Bu sebeple de bu kullanıcıların zevklerine ve isteklerine uygun önerilerde bulunmak oldukça zordur ve internet üzerinden alışveriş yapılan e-ticaret siteleri için ciddi bir problem oluşturur.

3.2 Yeni nesne (cold item)

Bir nesneye hiçbir kullanıcı tarafından puan verilmemişse veya ürün sisteme yeni eklenmişse bu nesne herhangi bir kullanıcıya tavsiye edilemez [42]. Çünkü sistemde yeni nesneleri ifade edecek bir bilgi yoktur.

Yeni kullanıcı ve yeni nesne probleminin üstesinden gelebilmek için çeşitli yöntemler geliştirilmiştir. Nguyen, Denos, Berrut MovieLens verisi üzerinde yaptıkları çalışmayla yeni kullanıcılar için verilen tavsiye niteliğini geliştirerek oldukça başarılı sonuçlar elde etmişlerdir [43]. Çalışmalarında yeni kullanıcıların yaşına, çalışma alanlarına ve yaşadıkları şehre göre bir vektör (α-community space) belirleyip, bu vektörlere bağlı

olarak çeşitli kurallar oluşturup yeni kullanıcıların favori filmini ve puanını belirlemişlerdir. Yeni kullanıcı ve yeni ürün problemini çözmek için hem İF yaklaşımlarını hem de içerik bilgisini (content-based) birleştiren yaklaşımlar da önermişlerdir. Bu sistemle probleme belli oranda çözüm sağlamışlardır.

Ziegler, Lause ve Thine Amazon.com verisi üzerinde kitaplara ait taksonomi sınıf bilgisini kullanan yeni bir yaklaşım önermiştir. Bu çalışmayla birlikte kullanıcının profiline uygun öneriler yapılması sağlanmıştır [15]. Schein ise anahtar bir kelime kullanarak her nesnenin özelliğini yansıtan gizli değişken metodunu (latent variable method) kullanmıştır. Yaptığı çalışmada oldukça başarılı sonuçlar elde etmiştir. Kim ve Li kullanıcı puanları üzerinde Gauss dağılımı kullanarak tahminlerde bulunmuş ve nesneleri sınıflandırarak cold user ve cold item problemini çözmeye çalışmıştır [44].

3.3 Ölçeklenebilirlik

İF tekniklerinde kullanıcıya özgü önerilerde bulunmak için kullanıcıya en yakın binlerce komşusu aranır ve bu çok fazla işlem maliyeti gerektirir. Mevcut algoritmalar çok fazla bilgisi olan bireysel kullanıcılar için performansı yavaşlatır. Örneğin; bir siteyi çok ziyaret eden bir kullanıcı için tavsiyelerde bulunmak oldukça zordur. Çünkü milyonlarca müşteri ve binlerce farklı kategorilerde ürünler vardır. Kullanıcıya benzer milyonlarca komşusunu aramak performansı yavaşlatır. Ayrıca sistemde çok sayıda kullanıcı ve nesnenin olması hesaplama karmaşıklığını da beraberinde getirir.

SVD gibi boyut indirgeme teknikleri ölçekleme problemlerinin üstesinden gelir. SVD yüksek boyutlu veriyi düşük boyutlu uzaya geçirerek ölçekleme probleminin üstesinden gelir. Bellek tabanlı İF tekniklerinde ölçekleme problemini çözmede oldukça başarılıdır. SVD, PCA gibi boyut indirgeme teknikleriyle önemsiz, sistemde az sayıda nesneye puan veren kullanıcılar ya da sistemde az sayıda puan verilmiş nesneler atılır [45].

3.4 Eşanlamlılık

Benzer ya da aynı nesnelerin farklı isimlerle ifade edilmesidir. Örneğin; "film" ve "movie" aynı nesnelerdir. Ancak tavsiye sistemlerinin birçoğu bu nesneleri ayrı nesneler olarak algılar ve bu iki nesneyi birbiriyle eşleştiremez. Bu problemi çözmek için birtakım

çalışmalar yapılmıştır. Deerwester ve Hoffman yaptıkları çalışmada Latent Semantic Indexing (LSI) metodu ile nesneler arasında anlamsal ilişkiler bulmuşlardır [46, 47].

3.5 Değerlendirme Ölçütleri

Tavsiye sisteminin niteliği, doğruluğu sisteme uygulanan yöntemler sonucunda elde edilecek sonuçların değerlendirilmesine bağlı olarak değişir. İF uygulamalarının türüne göre çeşitli değerlendirme ölçütleri kullanılır. Herlocker, Tervear, Rield ve Konstan yaptıkları çalışmada kullanılan yönteme göre değerlendirme ölçütlerini çeşitli kategorilere ayrılmıştır [44]. MAE (Mean Absolute Error) ve RMSE (Root Mean Squared Error) İF tekniklerinin değerlendirilmesinde en yaygın kullanılan değerlendirme ölçütleridir [48].

3.5.1 Mean Absolute Error (MAE)

MAE sınıflandırma hatasını ya da sınıflandırma doğruluğunu ölçmek yerine gerçek puanlar ile elde edilen tahmini puanlar arasındaki ortalama mutlak farkı hesaplar[42].

$$\text{MAE}= \frac{\sum_{\{i,j\}} I p_{ij} - r_{ij} I}{n} \qquad (3.1)$$

Bu eşitlikte **n** tüm test kullanıcıların puan verdiği toplam sayıyı, p_{ij} **i** kullanıcısının **j** nesnesine verdiği tahmini puanı, r_{ij} ise **i** kullanıcının **j** nesnesine verdiği gerçek puanı gösterir. Bulunan MAE değeri ne kadar düşük çıkarsa sistemin başarısı o kadar yüksektir.

3.5.2 Normalized Mean Absolute Error (NMAE)

NMAE Mean Absolute Error değerlendirme ölçütünün hatalarını yüzdelik olarak normalize eder.

$$\text{NMAE}= \frac{MAE}{r_{max} - r_{min}} \qquad (3.2)$$

Bu eşitlikte r_{max} kullanıcılar tarafından verilen puanların en büyüğünü, r_{min} ise kullanıcılar tarafından verilen puanların en küçüğünü ifade eder.

3.5.3 Root Mean Squared Error (RMSE)

RMSE netflix veri kümesinde çok yaygın kullanılan değerlendirme ölçütlerinden biridir. RMSE aşağıdaki şekilde hesaplanılır:

RMSE= $\sqrt{\frac{1}{n}\sum_{\{i,j\}}(p_{ij} - r_{ij})^2}$ (3.3)

Bu eşitlikte **n** test kümesindeki kullanıcıların puanlandığı ürün sayısını diğer bir deyişle tahmin edilen puanların sayısını p_{ij} **i** kullanıcısının **j** nesnesine verdiği tahmini puanı , r_{ij} ise **i** kullanıcının **j** nesnesine verdiği gerçek puanı gösterir [1].

BÖLÜM 4

BOYUT İNDİRGEME METOTLARI

4.1 Singular Value Decomposition (SVD)

SVD bir matris faktorizasyon tekniğidir. *m*×*n* boyutlarında bir matris ve **r** rank olmak üzere A matrisin SVD ile ayrıştırma şekli eşitlik 4.1 de gösterilmiştir [49].

SVD(A)= $U \times S \times V^T$ (4.1)

U ve V sırasıyla *m*×*m* ve *n*×*n* boyutlarında olan ortogonal matrislerdir. S ise tekil değerler(singular matrix) ve köşegen girişleri sıfırdan farklı olan bir köşegen matristir. S tekil matrisinin ilk **r** köşegen girişleri (s_1, s_2 ,....., s_n) s_i>0 ve s_1 >= s_2 >=.....>= s_r kuralına sahiptir. U matrisinin ilk **r** sütunları A.A^{T} 'nin özvektörüdür ve A matrisinin sol tekil vektörlerini (left singular vectors) gösterir. V matrisinin ilk **r** sütunları ise A^{T}.A 'nın özvektörleridir ve A matrisinin sağ tekil vektörlerini (right singular vectors) gösterir.

SVD ile etkili bir şekilde boyut indirgemek için sıfır olmayan **r** değerleri alınır ve U,S,V sırasıyla *m*×*r* , *r*×*r* , *r*×*n* boyutlarında oluşur.

SVD tekniği kullanıcılar ve nesneler arasındaki gizli kalmış ilişkileri gösterdiği için A matrisini en iyi ifade eder. k<<r tekil değerleri elde edilir, geri kalan değerler ise atılır ve S matrisinin girişleri sıralanır. Böylelikle orjinal A matrisinde gözükmeyen, gizli kalmış ancak önemli olan kullanıcılar arasındaki benzerlikler, etkileşimler bulunmuş olunur. Elde edilen köşegen matris S_k ; U ve V matrisleri ise U_k ve V_k olarak adlandırılır. U_k U matrisinin r-k sütunlarını atarak, V_k ise V matrisinin r-k sütunlarını atarak üretilir. A_k matrisi denklem 4.2 deki şekilde gösterilir:

$A_k = U_k \times S_k \times V_k^T$ (4.2)

Böylelikle kullanıcılar ve nesneler **k** boyut uzayına indirgenerek gösterilmiş olunur. **k** orjinal A matrisine en yakın linear yakınsamadır.

4.2 Tavsiye Sistemlerinde SVD Kullanımı

SVD hem hafıza tabanlı hem model tabanlı tavsiye sistemlerinde kullanılmaktadır. Latent Semantic Indexing-LSI tekniğidir ve veri seyrekliği, ölçekleme, eşanlamlılık, gizlilik gibi İF yaklaşımlarında rastlanan problemleri çözmede oldukça etkili bir yöntemdir. SVD nesne tabanlı İF tekniğinde aşağıdaki gibi kullanılmaktadır.

4.2.1 Nesne Tabanlı İF Tekniklerinde SVD Kullanımı

A *m*×*n* boyutlarında orjinal kullanıcı-nesne matrisi olmak üzere SVD yöntemi ile tekil özdeğerlerine ayrıştırılmış biçimi aşağıdaki gibidir:

$A_k = U_k \times S_k \times V_k^T$ (4.3)

Nesneler arasındaki gizli ilişkileri, benzerlikleri yakalayıp kullanıcılara daha doğru önerilerde bulunmak için A matrisi **k** boyutuna indirgenir. $U\sqrt{S}$ *m*×*k* matrisinin **k** boyutundaki kullanıcıları, $\sqrt{S}V$ ise *n*×*k* boyutundaki nesneleri gösterir. Kullanıcıya daha güvenilir nesne tahmininde bulunmak için orjinal A matrisi U_k, S_k ve V_k olmak üzere k boyut uzayında gösterilir. Nesne tabanlı İF tekniklerinde SVD kullanımı eşitlik 4.4'de gösterildiği gibi yapılmaktadır:

$p_{ui} = A_u + U_k \sqrt{S_k}\,(u) \times \sqrt{S_k}\, V_k^T\,(i)$ (4.4)

p_{ui} **u** kullanıcısının **i** nesnesine verdiği puan hesaplanırken A_u **u** kullanıcısının tüm nesnelere verdiği ortalama puanı; U,S ve V sırasıyla orjinal A matrisine SVD uygulandıktan sonra elde edilen özdeğer, köşegen matrisi ve özvektörleri gösterir.

$A_{m\times \mathrm{n}} = U_{m\times \mathrm{r}} \sqrt{S_{r\times \mathrm{r}}} \sqrt{S_{r\times \mathrm{r}}} V_{r\times \mathrm{n}}$ olmak üzere nesne tabanlı İF tekniklerinde SVD aşağıdaki biçimde gösterilir.

$A_{m\times \mathrm{n}} = U_{m\times \mathrm{k}} \sqrt{S_{k\times \mathrm{k}}}\,(u) \sqrt{S_{k\times \mathrm{k}}}\, V_{n\times \mathrm{k}}^T(i)$ (4.5)

$A_{m\times \mathrm{n}} = U_{m\times \mathrm{k}} \sqrt{S_{k\times \mathrm{k}}}\,(u) \sqrt{S_{k\times \mathrm{k}}}\, V_{k\times \mathrm{n}}\,(i)$ (4.6)

Tavsiye sistemlerinde nesne tabanlı ile SVD'nin kullanımı eşitlik 4.5, 4.6, 4.7 ve 4.8'de gösterilmiştir. U özdeğerlerinde aktif kullanıcının (u) satırı $\sqrt{S_{k\times \mathrm{k}}}$ tüm sütunları ile çarpılırken ; $\sqrt{S_{k\times \mathrm{k}}}$ tüm satırları $V_{k\times \mathrm{n}}$ ise puanı bulunmak istenen nesnenin (i) sutunu çarpılır.

$$\begin{bmatrix} \\ A_{m\times \mathrm{n}} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ U_{m\times \mathrm{k}} \\ \\ \end{bmatrix} \begin{bmatrix} \\ \sqrt{S}_{kxk}\,(u) \\ \\ \end{bmatrix} \begin{bmatrix} \\ \sqrt{S}_{kxk} \\ \\ \end{bmatrix} \begin{bmatrix} \\ V^{T}_{n\times \mathrm{k}}(i) \\ \\ \end{bmatrix} \tag{4.7}$$

$$\begin{bmatrix} \\ A_{m\times \mathrm{n}} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ U_{m\times \mathrm{k}} \\ \\ \end{bmatrix} \begin{bmatrix} \\ \sqrt{S}_{kxk}\,(u) \\ \\ \end{bmatrix} \begin{bmatrix} \\ \sqrt{S}_{kxk} \\ \\ \end{bmatrix} \begin{bmatrix} \\ V_{kxn}\ (i) \\ \\ \end{bmatrix} \tag{4.8}$$

BÖLÜM 5

ZAMANA BAĞLI HAFIZA TABANLI İŞBİRLİKÇİ FİLTRELEME TEKNİĞİ

Bu çalışmada geleneksel İF tekniklerinin zamanı dikkate almayan eksik yönleri giderilmeye çalışılmıştır. Geleneksel İF yöntemlerinde olduğu gibi sadece müşterinin ürüne verdiği puana göre değil de puan verdiği tarih de dikkate alınarak müşteriye özgü tavsiyeler verilmiştir. Bu sistem e-ticaret, telekom, reklamcılık, medya, müzik siteleri gibi çok geniş alanlarda kullanılarak daha fazla kişiye hitap edecek alanlar oluşturacaktır. Örneğin; bir kitap/cd satan Amazon gibi siteler çeşitli kategorilerde binlerce ürüne sahiptir. Çok geniş alanlarda müşterilerin beğeni ve beklentilerini belirlemek oldukça zordur.

Bizim tarafımızdan oluşturulacak sistemle benzerlik ve tahmin hesabı yapılırken zaman değişkeni de sistemde kullanılmıştır. Böylelikle kullanıcı, nesne ve hibrit tabanlı İF yaklaşımlarına kullanıcı puanlarının yaş bilgisi de ekleyerek her kullanıcıya gerçek ihtiyaç ve ilgi alanına yönelik güncel önerilerde bulunabilen yenilikçi, performans açısından hızlı ve güvenilir bir sistem oluşturulmuştur. Ayrıca müşteriyi memnun edecek kullanışlı bir sistem tasarlanarak firmanın rakipleri karşısında müşteri sayısını arttırma potansiyeli sağlayan bir sistem oluşturulmuştur.

5.1 Puanların Yaşını Kullanarak Yapılan Zamana Bağlı İF Tekniği

Geleneksel kullanıcı tabanlı İF tekniğinde kullanıcılar arasındaki ve geleneksel nesne tabanlı İF tekniğinde nesneler arasındaki benzerlik hesaplanırken sadece kullanıcıların ürünler için vermiş olduğu puanlar kullanılmaktadır[1,16]. Kullanıcıların ürünlere vermiş olduğu puanlar Çizelge 5.1'de kullanıcı-ürün matrisi formunda gösterilmektedir. Bu

sistemlerde hesaplanan kullanıcı benzerliğine göre aktif kullanıcının puanı tahmin edilmeye çalışılmaktadır.

Bu çalışmada zamana bağlı İF çalışması ile daha önceden uygulanmamış bir yöntem geliştirilmiştir. Geleneksel İF tekniğinden farklı olarak kullanıcının vermiş olduğu puanın yaşı değerlendirme sistemine eklenmiştir. Diğer bir deyişle, puanların yaşına göre kullanıcı benzerliği ve tahmin edilen puan değişecektir.

Zamana bağlı İF çalışmasında Age_{nm} **n** kullanıcısının **m** ürününe vermiş olduğu puanın yaşını ifade eder. Aktif kullanıcının 3 nolu ürüne verdiği puan (r_{a3}) test örneği olsun. Önerilen bu sistemde bir kullanıcının veya ürünün benzerliği hesaplanırken her puanın yaşına da ihtiyaç duyulur. Örneğin; r_{13}'ün yaşı hesaplanırken test verisinin ve r_{13}'ün puan tarihleri arasındaki farka bakılacaktır. Test verisinin hesaplanacağı tarih 27.01.2005 ve r_{13}'ün tarihi 27.01.2003 olduğundan arasındaki fark 731 gün diğer bir deyişle yaklaşık 2 yıl olur. Bu süre r_{13} 'ün mevcut tarihe göre (27.01.2005 test tarihi) yaşını gösterir. Çizelge 5.1'de de görülen puanların yaşları mevcut tarihe göre değişen dinamik bir bilgidir.

Çizelge 5.1'de görüldüğü gibi verilen puanların yaşları test tarihine göre gün bazında hesaplanmaktadır. Geliştirilen sistemde bu puanların yaşları yıla çevrilerek kullanılmıştır. Amaç kullanıcı tabanlı İF tekniğinde verilen puanların yaşlarını kullanarak puan tahmininde hatayı azaltmaktır.

Önerdiğimiz yöntemde kullanıcıların vermiş olduğu puanlar; puanların yaşları büyükse (yaşlıysa) zayıflatılarak ya da yaşları küçükse (gençse) kuvvetlendirilerek kullanılmıştır. Bu zayıflatma ya da kuvvetlendirme işleminde çeşitli dönüşüm fonksiyonları denenerek en uygun dönüşüm fonksiyonu bulunmaya çalışılmıştır. Şekil 5.1 de puanların yaşını ağırlıklandırmak için kullanılan dönüşüm fonksiyonları gösterilmiştir. Bu dönüşüm fonksiyonu eşitlik 5.1 deki şekilde uygulanmıştır.

$$Age_{\mathrm{w}} = mAge_r + n \tag{5.1}$$

Bu fonksiyonda Age_r mevcut puanın yaşını, Age_{w} ise mevcut puanın ağırlıklandırılmış yaşını göstermektedir. Önerilen yöntem netflix müşterilerinin çeşitli filmlere son iki yılda

verdikleri puanlar baz alınarak yapıldığından Age_r [0,…,2] aralığında değer alırken Age_{w} Şekil 5.1 de gösterilen dönüşüm fonksiyonu ile hesaplanan değerleri alır.

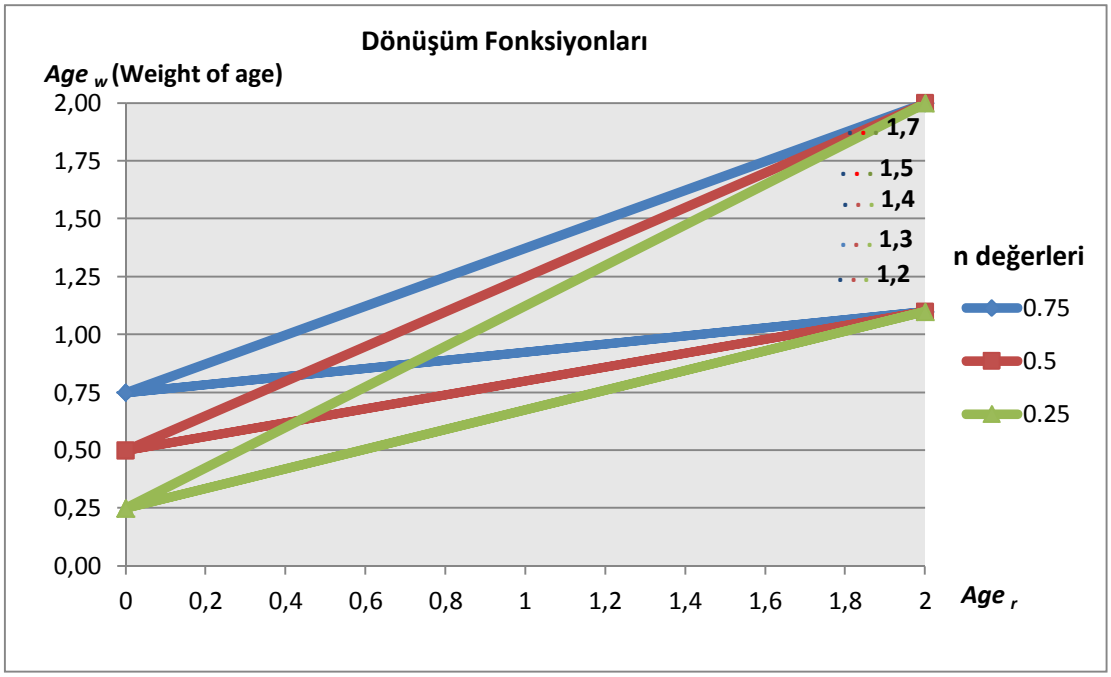
Çizelge 5.1 Kullanıcı-ürün matrisindeki puanların test örneğine göre yaşlarını gün olarak gösteren tablo

Kullanıcılar/ Ürünler	$\boldsymbol{i_1}$	$\boldsymbol{i_2}$	$\boldsymbol{i_3}$		$\boldsymbol{i_m}$
$\mathbf{u_1}$	$r_{1,1}$ *ve* $Age_{1,1}$ *(352 gün) 10.02.2004*		$r_{1,3}$ *ve* $Age_{1,3}$ *(731 gün) 27.01.2003*		
$\mathbf{u_2}$	$r_{2,1}$ *ve* $Age_{2,1}$ *(434 gün) 20.11.2003*				$r_{2,m}$ *ve* $Age_{2,m}$ *(26 gün) 01.01.2005*
a	$r_{a,1}$ *ve* $Age_{a,1}$ *(468 gün) 17.10.2003*		$\boldsymbol{r_{a,3}}$ **=?** $\boldsymbol{Age_{a,3}}$ ***Test Örneği (27.01.2005)***		
$\mathbf{u_n}$	$r_{n,1}$ *ve* $Age_{n,1}$ *(135 gün) 14.09.2004*				$r_{n,m}$ *ve* $Age_{n,m}$ *(235 gün) 06.06.2004*

Dönüşüm fonksiyonu ile hesaplanan Age_{w} mevcut puanın ağırlıklandırma sonucu elde edilen zamana dayalı yeni değerini hesaplamada eşitlik 5.2 de gösterildiği gibi kullanılır.

r_{aged}=r_{org}/Age_{w} (5.2)

Bu denklemde r_{org} mevcut puanı, r_{aged} ise zamana bağlı olarak ağırlıklandırılmış puanı gösterir. Bu dönüşüm işleminden sonra Bölüm 2 de anlatılan geleneksel yöntemlerde r yerine r_{aged} kullanılarak öneri sistemi zamana bağlı hale getirilmiştir.



Şekil 5.1 Puanların yaşını ağırlıklandırmak için kullanılan dönüşüm fonksiyonları

Lineer dönüşüm işlemi eşitlik 5.1'de tanıtılan fonksiyon ile yapılmaktadır. Çizelge 5.2 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için m değerleri gösterilmiştir. Age_r nin en küçük değeri olan 0 için Age_{w} 0,25 değerini alırken (n=0,25), Age_r=2 için sırasıyla dönüşüm sonucu 2 , 1,7, 1,5, 1,4, 1.3, 1,2 ve 1,1 değerlerini alacak şekilde 6 farklı dönüşüm fonksiyonu denenmiştir. Bu dönüşüm fonksiyonunda Age_r=0 ise Age_{w}=0,25 olduğunda r_{aged} = r_{org} / Age_{w} işlemi sonucunda orjinal puan 4 katına çıkmaktadır. Diğer tarafta Age_{w}=1,1 değerini aldığında ise orjinal puan %10 zayıflatılmaktadır.

Şekil 5.1'de mevcut puanların yaşını ağırlıklandırmak için kullanılan dönüşüm fonksiyonları kullanılmıştır. Grafiğe göre mevcut puanları kuvvetlendirmek için n= 0,25, 0,5 ve 0,75 değerlerini alırken; zayıflatmak için ise Age_{w} 2'den 1,1'e kadar farklı değerler almaktadır. Dönüşüm fonksiyonlarındaki eğimin artması ya da azalmasına bağlı olarak puan tahmini yapılmaktadır.

Çizelge 5.2 n=0,25 için bulunan dönüşüm katsayıları

Age_r	Age_w	m
[0…2] ⇨	[0,25.....2,0]	0,875
	[0,25.....1,7]	0,725
	[0,25.....1,5]	0,625
	[0,25.....1,4]	0,575
	[0,25.....1,3]	0,525
	[0,25.....1,2]	0,475
	[0,25.....1,1]	0,425

Çizelge 5.3 n=0,5 için bulunan dönüşüm katsayıları

Age_r	Age_w	m
[0…2] ⇨	[0,5.....2,0]	0,750
	[0,5.....1,7]	0,600
	[0,5.....1,5]	0,500
	[0,5.....1,4]	0,450
	[0,5.....1,3]	0,400
	[0,5.....1,2]	0,350
	[0,5.....1,1]	0,300

Çizelge 5.3 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için m değerleri gösterilmiştir. Age_r nin en küçük değeri olan 0 için Age_w 0,5 değerini alırken, Age_r=2 için sırasıyla dönüşüm sonucu 2, 1,7, 1,5, 1,4, 1,3, 1,2 ve 1,1 değerlerini alacak şekilde dönüşüm fonksiyonları oluşturulmuştur.

Çizelge 5.4 n=0,75 için bulunan dönüşüm katsayıları

Age_r	Age_w	m
[0…2] ⇨	[0,75.....2,0]	0,625
	[0,75.....1,7]	0,475
	[0,75.....1,5]	0,375
	[0,75.....1,4]	0,325
	[0,75.....1,3]	0,275
	[0,75.....1,2]	0,225
	[0,75.....1,1]	0,175

Çizelge 5.4 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için m değerleri gösterilmiştir. Age_r nin en küçük değeri olan 0 için Age_w 0,75 değerini alırken, Age_r=2 için sırasıyla dönüşüm sonucu 2 , 1,7, 1,5, 1,4, 1,3, 1,2 ve 1,1 değerlerini alacak şekilde dönüşüm fonksiyonları denenmiştir. Böylelikle kullanıcılar için en doğru, en güvenilir tahminlerin yapıldığı dönüşüm fonksiyonu belirlenmeye çalışılmıştır.

BÖLÜM 6

VERİ KÜMELERİ

Bu çalışmada netflix ve movielens olmak üzere iki tane veri kümesi kullanılmıştır. Movielens verisinde tarihsel bilgi mevcut değildir.

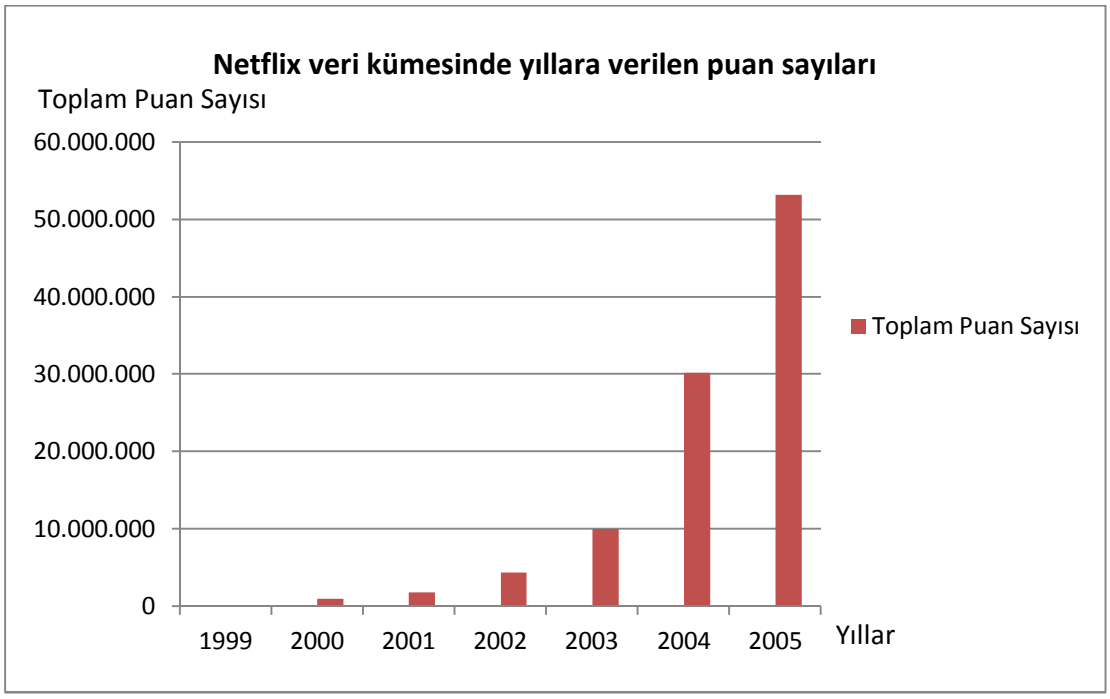
6.1 Netflix Veri Kümesinin Analizi

Önerilen sistemin başarısını karşılaştırmalı olarak ölçmek amacı ile netflix müşterilerinin son iki yılda çeşitli filmlere vermiş oldukları puanlar baz alınarak çalışılmıştır. Netflix veri kümesi yaklaşık yarım milyon müşterinin 1999'dan 2005'e kadar geçen sürelerde 17770 film için verdikleri puanlardan oluşur [43,50]. 100 milyon üzerinde puan verilmiş kayıt vardır. Amerika'da film ve video kiralama üzerine kurulmuş bir şirketten alınan gerçek veri kümesidir ve tavsiye sistemlerinde sıklıkla kullanılır. Bu çalışmada uygulamanın sonuçları eşitlik 3.3'de verilen RMSE ile hesaplanmıştır.

Çizelge 6.1 Yıllara göre verilen toplam puan sayısı

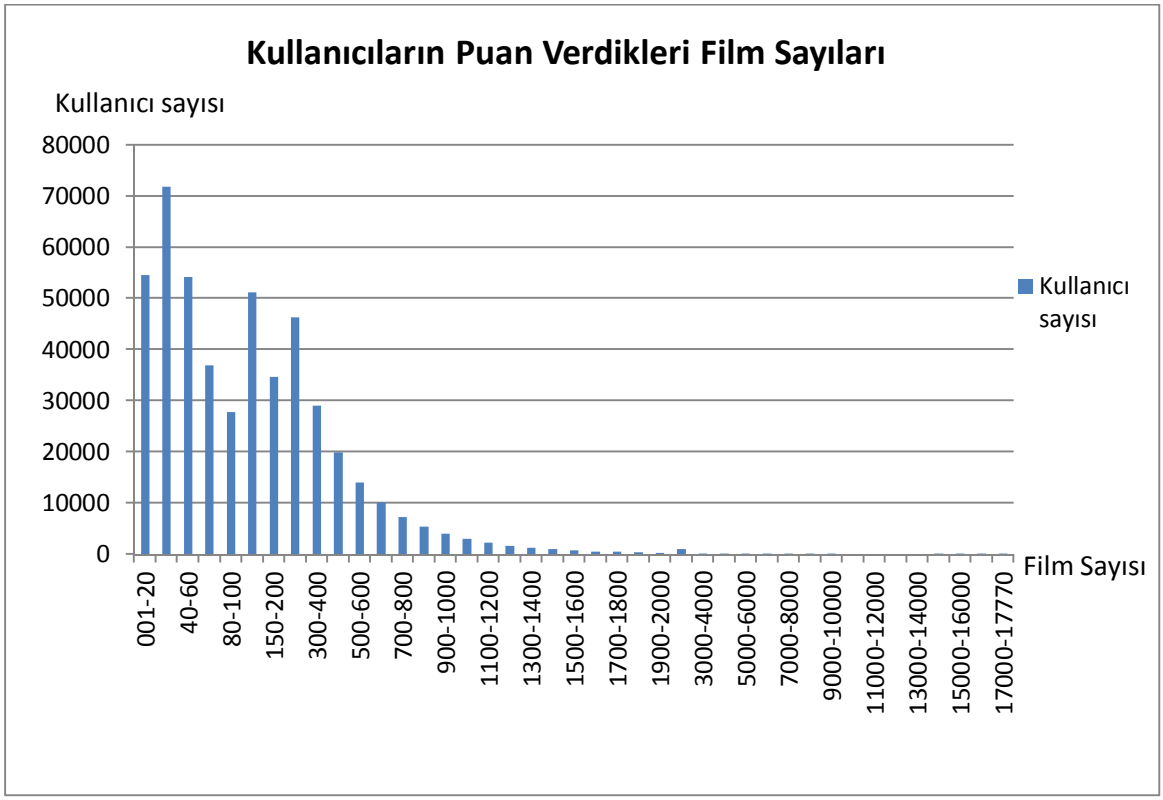
Yıllar	Toplam puan sayısı
1999	2178
2000	924.443
2001	1.769.031
2002	4.342.871
2003	9.985.337
2004	30.206.574
2005	53.250.073

Netflix veri kümesinde yıllara göre verilen puanlar Şekil 6.1 ve Çizelge 6.1'de gösterilmiştir. En fazla puan 2004 ve 2005 yılında verilmiştir.



Şekil 6.1 Netflix veri kümesinde müşterilerin yıllara göre verdikleri puan sayıları

Netflix veri kümesinde kullanıcıların puan verdikleri film sayıları Şekil 6.2'de gösterilmiştir. Netflix verisi üzerinde yapılan analize göre ortalama 39322 kullanıcı 0 ile 400 arasındaki filme puan vermiştir.



Şekil 6.2 Netflix veri kümesinde kullanıcıların puan verdikleri film sayıları

6.2 MovieLens Verisi

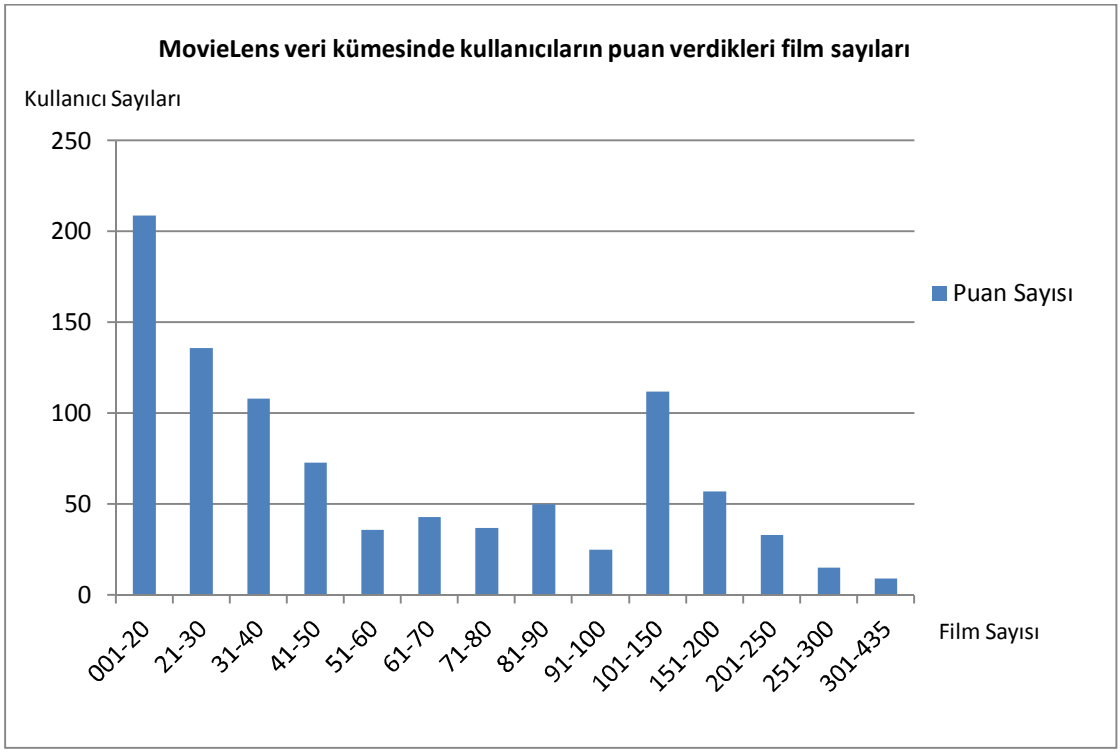
MovieLens veri kümesi Minnesota Üniversitesinde GroupLens Araştırma Projesi tarafından toplanan verilerden oluşur [51,52]. Bu veri kümesinde 943 kullanıcı 1682 filme 1-5 arasında puan vermiştir. 100.000 civarında puan vardır. Her kullanıcı en az 20 filme puan vermiştir. Ayrıca MovieLens veri kümesinde kullanıcılara ait yaş, cinsiyet, meslek gibi basit demografik bilgilerin yanı sıra; filmler türlerine ait bilgiler de mevcuttur. MovieLens veri kümesi çok seyrek (sparse) bir veri kümesidir. Veri seyrekliği oranı 100000/943*1682=0,937 'dir.

MovieLens veri kümesinde tarihsel bilgi bulunmamaktadır. Bu veri kümesinin kullanılma nedeni SVD boyut indirgeme yönteminin nesne tabanlı İF yöntemiyle nasıl kullanıldığını araştırmaktır. Çünkü movielens verisi netflix verisine göre daha az kullanıcı ve daha az nesne içermektedir. Nesne tabanlı İF tekniğinin uygulanabilirliğini netflix verisinde daha kolay analiz etmek için movielens verisi kullanılmıştır. Amaç; hem zamansal İF tekniğini hem de SVD yöntemini netflix verisi üzerinde uygulamaktır.

Çizelge 6.2 Kullanıcıların puan verdikleri film sayıları

Kullanıcı Sayıları	**Puan Sayısı**
209	1-20
136	21-30
108	31-40
73	41-50
36	51-60
43	61-70
37	71-80
50	81-90
25	91-100
112	101-150
57	151-200
33	201-250
15	251-300
9	301-435

Çizelge 6.2 'de kullanıcıların puan verdikleri film sayıları gösterilmiştir. MovieLens verisi üzerinde yapılan analize göre genelde kullanıcıların birçoğu 20 filme puan vermiştir. Kullanıcıların büyük bir çoğunluğu az sayıda filme puan vermiştir.



Şekil 6.3 MovieLens veri kümesinde film sayılarına göre puan veren kullanıcıların sayısı

Şekil 6.3'te movielens veri kümesinde film sayılarına göre puan veren kullanıcıların sayısı gösterilmiştir. Yapılan analize göre movielens verisinde ortalama 135 kullanıcı 50 filme oy vermiştir.

BÖLÜM 7

DENEYSEL ÇALIŞMA

Klasik kullanıcı tabanlı, nesne tabanlı ve hibrit tekniği ile zamana bağlı önerilen sistemin tahmin başarısı ölçülmüş ve sonuçlar karşılaştırılmıştır.

7.1 Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği

Önerilen sistem ile klasik kullanıcı tabanlı İF tekniğini karşılaştırmak için yapılan analizde az sayıda filme puan veren kullanıcılar ve çok sayıda filme puan veren kullanıcılar için 2 ayrı deneme kümesi için oluşturulmuştur. Az sayıda puan veren kullanıcılar 100 film ile 800 film arasındaki filme puan veren kullanıcılar grubunu oluştururken; çok sayıda puan veren kullanıcılar ise 7000 ile 10000 filme puan veren kullanıcılar grubunu oluşturur.

7.1.1 Az Sayıda Filme Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği

Az sayıda filme puan veren kullanıcılar ile yapılan analizde puan tahmininde her iki yöntemde de hata yüksek çıkmıştır. Bunun nedeni seçilen aktif kullanıcıların az sayıda filme puan vermiş olması ve bunun sonucunda aktif kullanıcıya benzer kullanıcıların tespitinin zorlaşmasıdır. Az sayıda filme puan veren kullanıcılar ile çok filme puan veren aktif kullanıcılar ile yapılan analizde elde edilen sonuçlar aşağıdaki tablolarda ve grafiklerde gösterilmiştir.

Çizelge 7.1 Az puan veren kullanıcılar için klasik kullanıcı tabanlı İF'nin tahmin hatası

Klasik Kullanıcı Tabanlı İF	RMSE	Yüzdelik Hata (%)
Az puan veren kullanıcılar	2,38	47,75

Çizelge 7.1 ‘de az sayıda puan veren kullanıcılar grubu için klasik kullanıcı tabanlı İF tekniğinin hata oranları verilmiştir. Klasik kullanıcı tabanlı İF tekniği ile yaklaşık %47 hata elde edilmiştir.

Çizelge 7.2 Az sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	1,94	38,91
[0,75...1,7]	**1,81**	**36,32**
[0,75...1,5]	1,98	39,74
[0,75...1,4]	1,99	39,93
[0,75...1,3]	2,10	42,02
[0,75...1,2]	2,17	43,46
[0,75...1,1]	2,29	45,81

Çizelge 7.2 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatası gösterilmiştir. Dönüşüm fonksiyonlarında eğim azaldıkça hata oranında artış gözlemlenmektedir. En minimum hata oranı $\mathrm{Age_w}$ = 1,7 olduğunda yaklaşık %36 bulunmuştur.

Çizelge 7.3 Az sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

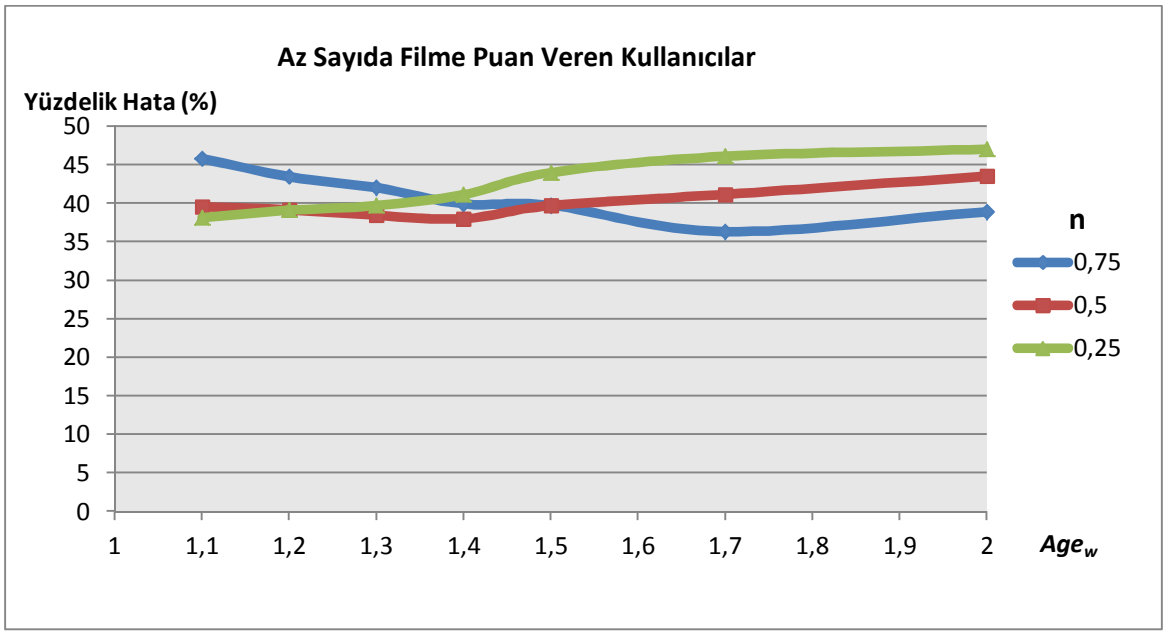
n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	2,17	43,51
[0,5...1,7]	2,05	41,17
[0,5...1,5]	1,98	39,72
[0,5...1,4]	**1,89**	**37,99**
[0,5...1,3]	1,92	38,45
[0,5...1,2]	1,95	39,16
[0,5...1,1]	1,97	39,52

Çizelge 7.3 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,4 olduğunda yaklaşık %38 bulunmuştur.

Çizelge 7.4 Az sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	2,35	47,03
[0,25...1,7]	2,30	46,10
[0,25...1,5]	2,19	43,99
[0,25...1,4]	2,05	41,12
[0,25...1,3]	1,98	39,71
[0,25...1,2]	1,95	39,10
[0,25...1,1]	**1,90**	**38,12**

Çizelge 7.4 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata oranı $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %38 bulunmuştur.



Şekil 7.1 Zamana bağlı kullanıcı tabanlı İF tekniğinin sonuçları(az oylama yapan kullanıcılar için)

Klasik kullanıcı tabanlı İF tekniğinin az filme puan veren kullanıcılar grubu için hata oranı %47,75 dir. Şekil 7.1'de zamana bağlı kullanıcı tabanlı İF çalışmasının az filme puan veren kullanıcılar grubu için sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim düştükçe hata oranında sürekli bir azalış olduğunu göstermektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim düştükçe hata oranında artış olduğu gözlemlenmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride Age_w dönüşüm fonksiyonu 1,7 olduğunda yaklaşık %37 ile hata oranının en düşük olduğu değer elde edilir.

7.1.2 Çok Sayıda Filme Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Kullanıcı Tabanlı İF Tekniği

Zamana bağlı kullanıcı tabanlı İF tekniğinin çok sayıda filme puan veren diğer grubu için yapılan çalışmalar aşağıdaki çizelgelerde ve grafikte gösterilmiştir.

Çizelge 7.5 Çok puan veren kullanıcılar için klasik kullanıcı tabanlı İF'nin tahmin hatası

Klasik Kullanıcı Tabanlı İF	**RMSE**	**Yüzdelik Hata (%)**
Çok puan veren kullanıcılar	1,67	33,51

Çizelge 7.5'de klasik kullanıcı tabanlı İF tekniğinin çok sayıda puan veren kullanıcılar için tahmin hatası verilmiştir. Klasik kullanıcı tabanlı İF tekniği ile yaklaşık %33 hata elde edilmiştir.

Çizelge 7.6 Çok sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

n=0,75		
Age_w **için Değişim Aralığı**	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	**1,08**	**21,67**
[0,75...1,7]	1,19	23,91
[0,75...1,5]	1,34	26,88
[0,75...1,4]	1,48	29,77
[0,75...1,3]	1,52	30,43
[0,75...1,2]	1,64	32,85
[0,75...1,1]	1,69	33,98

Çizelge 7.6 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin başarısı gösterilmiştir. Dönüşüm fonksiyonunda eğilim azaldıkça hata oranında artış gözlemlenmektedir. Eğim azaldıkça Age_r =2 olduğunda mevcut puanın yaşı diğer fonksiyonlara göre daha fazla zayıflatıldığından; Age_r =0 olduğunda mevcut puanın yaşı daha fazla kuvvetlendirildiğinden hata oranı yüksek çıkmaktadır. En minimum hata oranı $\mathrm{Age_w}$ = 2 olduğunda yaklaşık %21 bulunmuştur.

Çizelge 7.7 Çok sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

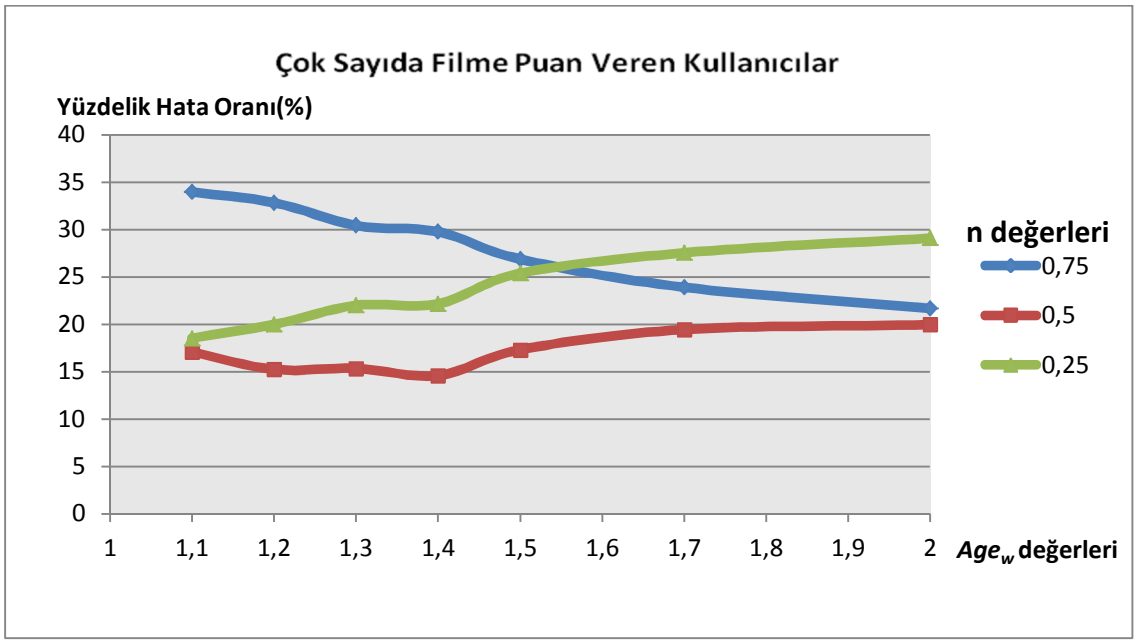
n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	0,99	19,97
[0,5...1,7]	0,97	19,46
[0,5...1,5]	0,86	17,61
[0,5...1,4]	**0,73**	**14,60**
[0,5...1,3]	0,75	15,36
[0,5...1,2]	0,76	15,24
[0,5...1,1]	0,85	17,09

Çizelge 7.7 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmektedir. En minimum hata $\mathrm{Age_w}$ = 1,4 olduğunda yaklaşık %14 bulunmuştur.

Çizelge 7.8 Çok sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	1,45	29,10
[0,25...1,7]	1,37	27,54
[0,25...1,5]	1,26	25,39
[0,25...1,4]	1,10	22,15
[0,25...1,3]	1,10	22,01
[0,25...1,2]	1,04	20,99
[0,25...1,1]	**1,00**	**20,15**

Çizelge 7.8 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmektedir. En minimum hata oranı $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %20 bulunmuştur.



Şekil 7.2 Zamana bağlı kullanıcı tabanlı İF tekniğinin sonuçları(çok oylama yapan kullanıcılar için)

Klasik kullanıcı tabanlı İF tekniğinin çok filme puan veren kullanıcılar grubu için hata oranı %33,51 dir. Şekil 7.2 'de zamana bağlı kullanıcı tabanlı İF çalışmasının çok sayıda filme puan veren kullanıcılar grubu için yapılan sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu gözlemlenmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış olduğu gösterilmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w dönüşüm fonksiyonu 1,4 olduğunda yaklaşık %14 ile hata oranının en düşük olduğu değer elde edilir.

7.2 Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği

Önerilen sistem ile nesne kullanıcı tabanlı İF tekniğini karşılaştırmak için yapılan analizde az sayıda puan verilmiş filmler ve çok sayıda puan verilmiş filmler için 2 ayrı deneme kümesi için oluşturulmuştur. Az sayıda oylanmış filmler grubunu 100 film ile 750 film arasında oylanmış filmler oluştururken; çok sayıda oylanmış filmler grubunu ise 55000 ile 13000 arasında oylanmış filmler oluşturur.

7.2.1 Az Sayıda Puan Verilmiş Filmler için Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği

Az sayıda puan verilmiş filmler ile yapılan analizde puan tahmininde her iki yöntemde de hata yüksek çıkmıştır. Bunun nedeni seçilen nesnelere az sayıdaki kullanıcı tarafından puan verilmesi ve bunun sonucunda puanı bulunmak istenen nesneye benzer nesnelerin tespitinin zorlaşmasıdır. Az sayıda puan verilmiş filmler ile çok sayıda puan verilmiş filmler ile yapılan analizde elde edilen sonuçlar aşağıdaki tablolarda ve grafiklerde gösterilmiştir.

Çizelge 7.9 Az puan verilen filmler için klasik nesne tabanlı İF tekniğinin tahmin hatası

Klasik Nesne Tabanlı İF	**RMSE**	**Yüzdelik Hata (%)**
Az puan verilmiş filmler	2,76	55,38

Çizelge 7.9'da klasik nesne tabanlı İF tekniğinin az puan verilen filmler grubu için sonuçları verilmiştir. Klasik nesne tabanlı İF tekniğinin az puan verilen filmler grubu için hata oranı yaklaşık %55 bulunmuştur.

Çizelge 7.10 Az sayıda puan verilen filmler için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	**2,35**	**47,17**
[0,75...1,7]	2,39	47,93
[0,75...1,5]	2,43	48,64
[0,75...1,4]	2,50	50,03
[0,75...1,3]	2,59	51,99
[0,75...1,2]	2,64	52,95
[0,75...1,1]	2,79	55,92

Çizelge 7.10 'da n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatası gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış gözlemlenmektedir. n=0,75 olduğunda $\mathrm{Age_w}$ =2 olduğunda en minimum hata yaklaşık %47 bulunmuştur.

Çizelge 7.11 Az sayıda puan verilen filmler için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

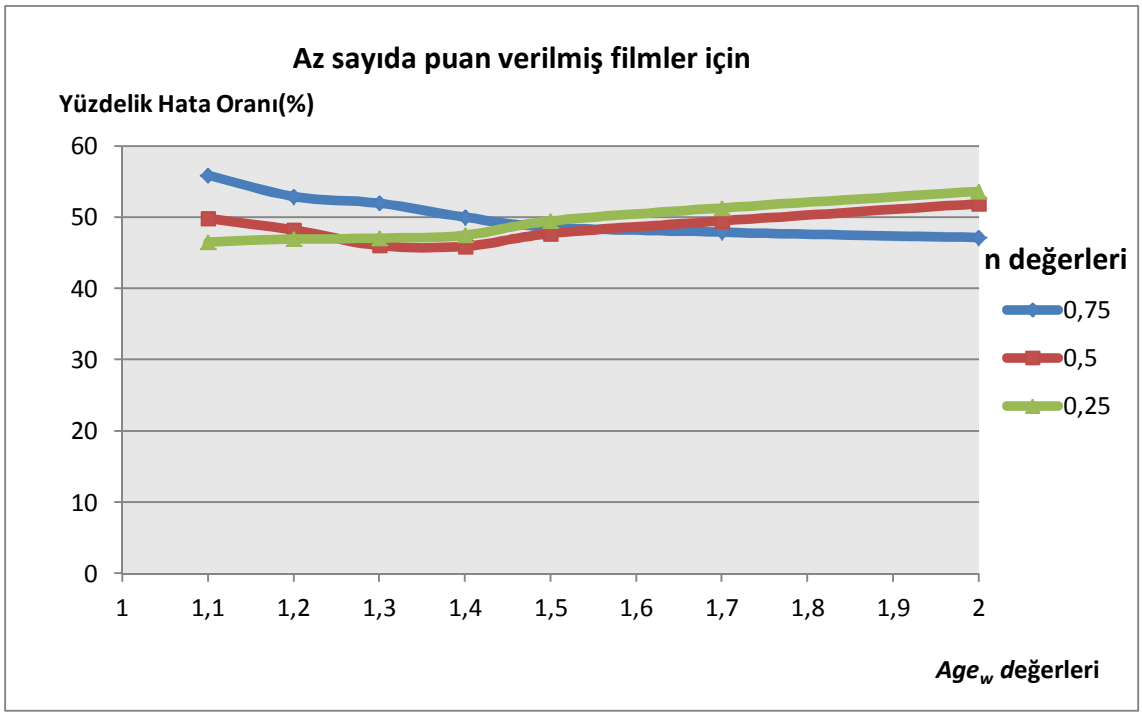
n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	2,59	51,93
[0,5...1,7]	2,47	49,51
[0,5...1,5]	2,38	47,73
[0,5...1,4]	**2,29**	**45,92**
[0,5...1,3]	2,30	46,05
[0,5...1,2]	2,41	48,24
[0,5...1,1]	2,49	49,86

Çizelge 7.12 Az sayıda puan verilen filmler için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	2,68	53,67
[0,25...1,7]	2,56	51,34
[0,25...1,5]	2,47	49,50
[0,25...1,4]	2,37	47,50
[0,25...1,3]	2,35	47,08
[0,25...1,2]	2,34	46,95
[0,25...1,1]	**2,32**	**46,52**

Çizelge 7.11 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmektedir. En minimum hata $\mathrm{Age_w}$ = 1,4 olduğunda yaklaşık %45 bulunmuştur.

Çizelge 7.12 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %46 bulunmuştur.



Şekil 7.3 Zamana bağlı nesne tabanlı İF tekniğinin sonuçları(az oylanmış filmler için)

Klasik nesne tabanlı İF tekniğinin az sayıda puan verilmiş filmler grubu için hata oranı %55,38 dir. Şekil 7.3' de zamana bağlı nesne tabanlı İF çalışmasının az sayıda puan verilmiş filmler grubu için yapılan sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu görülmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış olduğu gözlemlenmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w dönüşüm fonksiyonu 1,4 olduğunda yaklaşık %46 ile hata oranında en düşük değer elde edilmektedir.

7.2.2 Çok Sayıda Puan Verilen Filmler için Klasik ve Zamana Bağlı Nesne Tabanlı İF Tekniği

Klasik ve zamana bağlı nesne tabanlı İF tekniğinin çok sayıda puan verilen diğer grubu için yapılan çalışmalar aşağıdaki çizelgelerde ve grafikte gösterilmiştir.

Çizelge 7.13 Çok puan verilen filmler için klasik nesne tabanlı İF tekniğinin RMSE tahmin hatası

Klasik Nesne Tabanlı İF	**RMSE**	**Yüzdelik Hata (%)**
Çok puan verilmiş filmler	2,14	42,83

Çizelge 7.13'de klasik nesne tabanlı İF tekniğinin çok sayıda puan verilen filmler grubu için elde edilen RMSE tahmin hatası gösterilmiştir. Klasik nesne tabanlı İF tekniğinin çok puan verilen filmler grubu için hata oranı yaklaşık %42 bulunmuştur.

Çizelge 7.14 Çok sayıda puan verilen filmler için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	**1,15**	**30,72**
[0,75...1,7]	1,67	33,50
[0,75...1,5]	1,75	35,13
[0,75...1,4]	1,90	38,09
[0,75...1,3]	2,00	40,02
[0,75...1,2]	2,05	41,05
[0,75...1,1]	2,10	42,04

Çizelge 7.14 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatası gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış gözlemlenmektedir. En minimum hata $\mathrm{Age_w} = 2$ olduğunda yaklaşık %30 bulunmuştur.

Çizelge 7.15 Çok sayıda puan verilen filmler için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

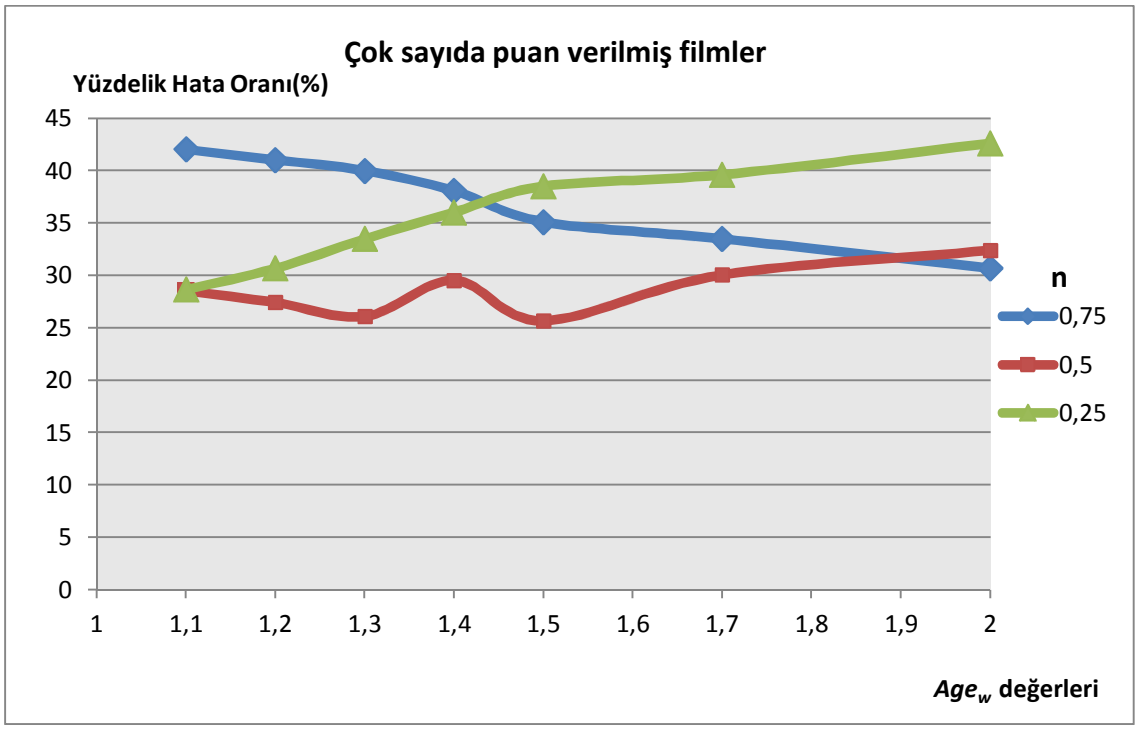
n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	1,62	32,41
[0,5...1,7]	1,50	30,06
[0,5...1,5]	**1,28**	**25,68**
[0,5...1,4]	1,47	29,55
[0,5...1,3]	1,30	26,09
[0,5...1,2]	1,37	27,44
[0,5...1,1]	1,42	28,53

Çizelge 7.16 Çok sayıda puan verilen filmler için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	2,13	42,61
[0,25...1,7]	1,98	39,64
[0,25...1,5]	1,92	38,52
[0,25...1,4]	1,80	36,04
[0,25...1,3]	1,67	33,49
[0,25...1,2]	1,53	30,67
[0,25...1,1]	**1,43**	**28,65**

Çizelge 7.15 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmektedir. En minimum hata $\mathrm{Age_w}$ = 1,5 olduğunda yaklaşık %25 bulunmuştur.

Çizelge 7.16 'da n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin başarısı gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %28 bulunmuştur.



Şekil 7.4 Zamana bağlı nesne tabanlı İF tekniğinin sonuçları(çok oylanmış filmler için)

Klasik nesne tabanlı İF tekniğinin çok sayıda puan verilmiş filmler grubu için hata oranı %42,83 dir. Şekil 7.4'de zamana bağlı nesne tabanlı İF çalışmasının çok sayıda puan verilmiş filmler grubu için yapılan sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu görülmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış olduğu gözlemlenmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w dönüşüm fonksiyonu 1,5 olduğunda yaklaşık %25 ile hata oranının en düşük olduğu değer elde edilmektedir.

7.3 Klasik ve Zamana Bağlı Hibrit İF Tekniği

Önerilen sistem ile hibrit İF tekniğini karşılaştırmak için yapılan analizde az sayıda puan verilmiş filmler ve az sayıda filme puan veren kullanıcılar ile çok sayıda puan verilmiş ve çok sayıda filme puan veren kullanıcılar için 2 ayrı deneme kümesi için oluşturulmuştur. İlk grup için az sayıda puan veren kullanıcılar ve az sayıda oylanan filmlerin keşişim

kümesini, ikinci grup için ise çok sayıda puan veren kullanıcılar ve çok sayıda oylanmış filmler kümesinin kesişimi alınmıştır.

7.3.1 Az Puan Verilmiş Filmler ve Az Sayıda Puan Veren Kullanıcılar için Zamana Bağlı Hibrit Tekniği

Az sayıda puan verilmiş ve az sayıda filme puan veren kullanıcılar ile yapılan analizde puan tahmininde her iki yöntemde de hata yüksek çıkmıştır. Bunun nedeni seçilen nesnelere ve kullanıcılara az sayıdaki kullanıcı tarafından puan verilmesi ve bunun sonucunda puanı bulunmak istenen nesneye benzer nesnelerin tespitinin zorlaşmasıdır. Klasik hibrit İF tekniği, az sayıda puan verilmiş ve az sayıda kullanıcının puan verdiği filmler ile çok sayıda puan verilmiş ve çok sayıda kullanıcının puan verdiği filmler ile yapılan analizde elde edilen sonuçlar aşağıdaki tablolarda ve grafiklerde gösterilmiştir.

Çizelge 7.17 Az puan verilmiş filmler ve az sayıda puan veren kullanıcılar grubu için klasik hibrit İF tekniğinin RMSE tahmin hatası

Klasik Hibrit İF Tekniği	RMSE	Yüzdelik Hata (%)
Az sayıda puan verilen filmler ve puan veren kullanıcılar	2,17	43,52

Çizelge 7.17'de az puan verilen filmler ve az sayıda puan veren kullanıcılar grubu için klasik hibrit İF tekniğinin RMSE tahmin hatası verilmiştir. Klasik hibrit tabanlı İF tekniğinin az puan veren kullanıcılar ve az puan verilen filmler grubu için hata oranı yaklaşık %43 bulunmuştur.

Çizelge 7.18 Az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	**1,80**	**36,02**
[0,75...1,7]	1,81	36,19
[0,75...1,5]	1,94	38,94
[0,75...1,4]	1,99	39,96
[0,75...1,3]	2,12	42,52
[0,75...1,2]	2,15	43,12
[0,75...1,1]	2,18	43,74

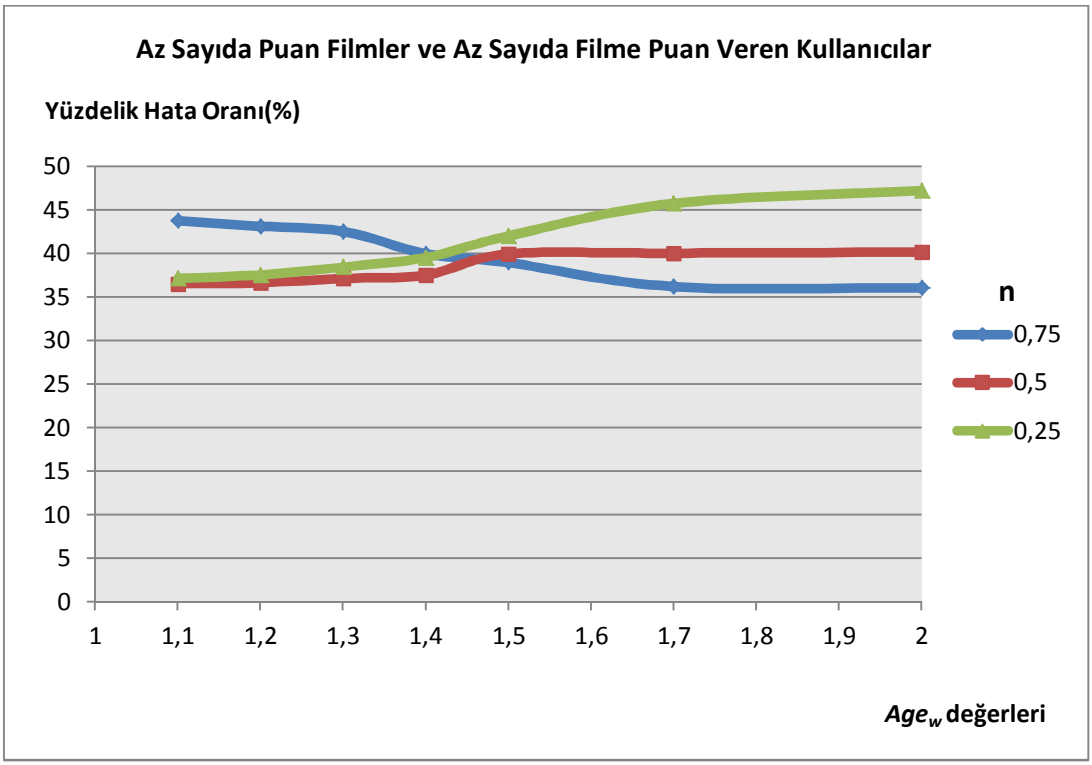
Çizelge 7.18 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatası gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış gözlemlenmektedir. En minimum hata $\mathrm{Age_w} = 2$ olduğunda yaklaşık %36 bulunmuştur.

Çizelge 7.19 Az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	2,01	40,15
[0,5...1,7]	2,00	40,03
[0,5...1,5]	1,99	39,94
[0,5...1,4]	1,87	37,52
[0,5...1,3]	1,85	37,13
[0,5...1,2]	1,83	36,64
[0,5...1,1]	**1,82**	**36,51**

Çizelge 7.20 Az sayıda puan verilen filmler ve az sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	2,36	47,23
[0,25...1,7]	2,28	45,74
[0,25...1,5]	2,10	42,01
[0,25...1,4]	1,97	39,52
[0,25...1,3]	1,92	38,42
[0,25...1,2]	1,87	37,53
[0,25...1,1]	**1,85**	**37,13**



Şekil 7.5 Zamana bağlı hibrit yönteminin sonuçları(az puan veren kullanıcılar ve az oylanmış filmler için)

Çizelge 7.19 'da n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %36 bulunmuştur.

Çizelge 7.20 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %37 bulunmuştur.

Klasik hibrit İF tekniğinin az sayıda puan verilmiş filmler ve az sayıda puan veren kullanıcılar grubu için hata oranı %43,52 dir. Şekil 7.5'de zamana bağlı hibrit çalışmasının az sayıda puan verilmiş filmler ve az sayıda puan veren kullanıcılar grubu için yapılan sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu görülmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış gözlemlenmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w dönüşüm fonksiyonu 1,4 olduğunda yaklaşık %37 ile hata oranının en düşük olduğu değer elde edilir.

7.3.2 Çok Sayıda Puan Verilen Filmler ve Çok Sayıda Puan Veren Kullanıcılar için Klasik ve Zamana Bağlı Hibrit Tekniği

Klasik ve zamana bağlı hibrit tekniğinin çok sayıda puan verilen ve çok sayıda puan veren diğer grubu için yapılan çalışmalar aşağıdaki çizelgelerde ve grafikte gösterilmiştir.

Çizelge 7.21 Çok puan verilen filmler ve çok sayıda puan veren kullanıcılar için klasik hibrit İF tekniğinin RMSE tahmin hatası

Klasik Hibrit İF Tekniği	RMSE	Yüzdelik Hata (%)
Çok sayıda puan verilen filmler ve puan veren kullanıcılar	1,59	31,86

Çizelge 7.21'de çok sayıda puan verilen filmler ve çok sayıda puan veren kullanıcılar grubu için klasik hibrit İF tekniğinin RMSE tahmin hatası verilmiştir. Klasik hibrit tabanlı İF tekniğinin çok puan veren kullanıcılar ve çok puan verilen filmler grubu için hata oranı yaklaşık %31 bulunmuştur.

Çizelge 7.22 Çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar için n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	0,89	17,83
[0,75...1,7]	**0,88**	**17,60**
[0,75...1,5]	1,27	25,57
[0,75...1,4]	1,37	27,52
[0,75...1,3]	1,46	29,30
[0,75...1,2]	1,58	31,72
[0,75...1,1]	1,77	35,52

Çizelge 7.23 Çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar için n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	1,02	20,55
[0,5...1,7]	1,08	21,74
[0,5...1,5]	0,86	17,32
[0,5...1,4]	0,78	15,67
[0,5...1,3]	**0,72**	**14,57**
[0,5...1,2]	0,75	15,08
[0,5...1,1]	0,86	17,32

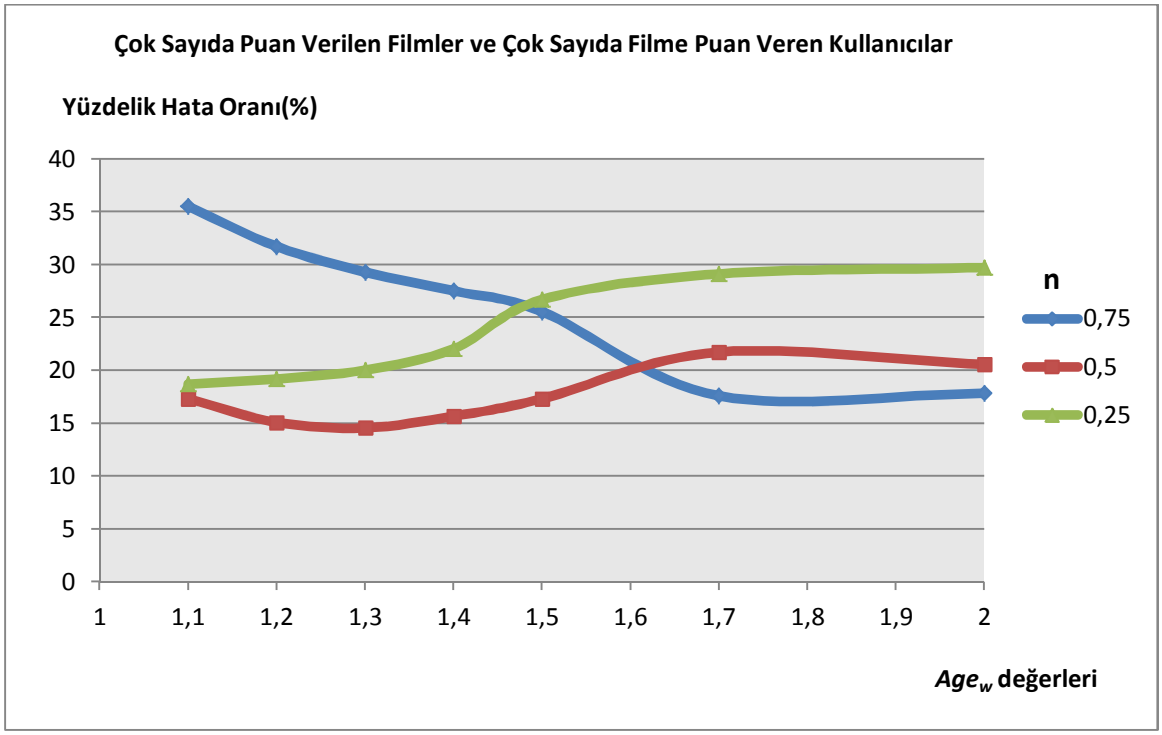
Çizelge 7.22 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatsı gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış gözlemlenmektedir. En minimum hata $\mathrm{Age_w}$= 1,7 olduğunda yaklaşık %17 bulunmuştur.

Çizelge 7.23 'de n=0,5 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,3 olduğunda yaklaşık %14 bulunmuştur.

Çizelge 7.24 Çok sayıda puan verilen filmler ve çok sayıda filme puan veren kullanıcılar için n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen RMSE tahmin hatası

n=0,25		
$\boldsymbol{Age_w}$ **için Değişim Aralığı**	**Zamana Bağlı Kullanıcı Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	1,48	29,72
[0,25...1,7]	1,45	29,14
[0,25...1,5]	1,33	26,70
[0,25...1,4]	1,10	22,05
[0,25...1,3]	1,00	20,04
[0,25...1,2]	0,95	19,19
[0,25...1,1]	**0,93**	**18,72**

Çizelge 7.24 'de n=0,25 sabit tutularak farklı dönüşüm aralıkları için tahmin hatası gösterilmiştir. En minimum hata $\mathrm{Age_w}$ = 1,1 olduğunda yaklaşık %18 bulunmuştur.



Şekil 7.6 Zamana bağlı hibrit yönteminin sonuçları(çok puan veren kullanıcılar ve çok oylanmış filmler için)

Klasik hibrit İF tekniğinin çok sayıda puan verilmiş filmler ve çok sayıda puan veren kullanıcılar grubu için hata oranı %31,86' dir. Şekil 7.6' da zamana bağlı hibrit çalışmasının çok sayıda puan verilmiş filmler ve çok sayıda puan veren kullanıcılar grubu için yapılan sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu görülmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış olduğu gözlemlenmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w değeri 1,3 olduğunda yaklaşık %14 ile hata oranının en düşük olduğu değer elde edilir.

7.4 MovieLens Verisine SVD Uygulanması

MovieLens verisine nesne ve kullanıcı tabanlı İF ile SVD uygulanması sonucu aşağıdaki grafikler elde edilmiştir.

Çizelge 7.25 Rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımı (düşük boyutlu k değerleri için)

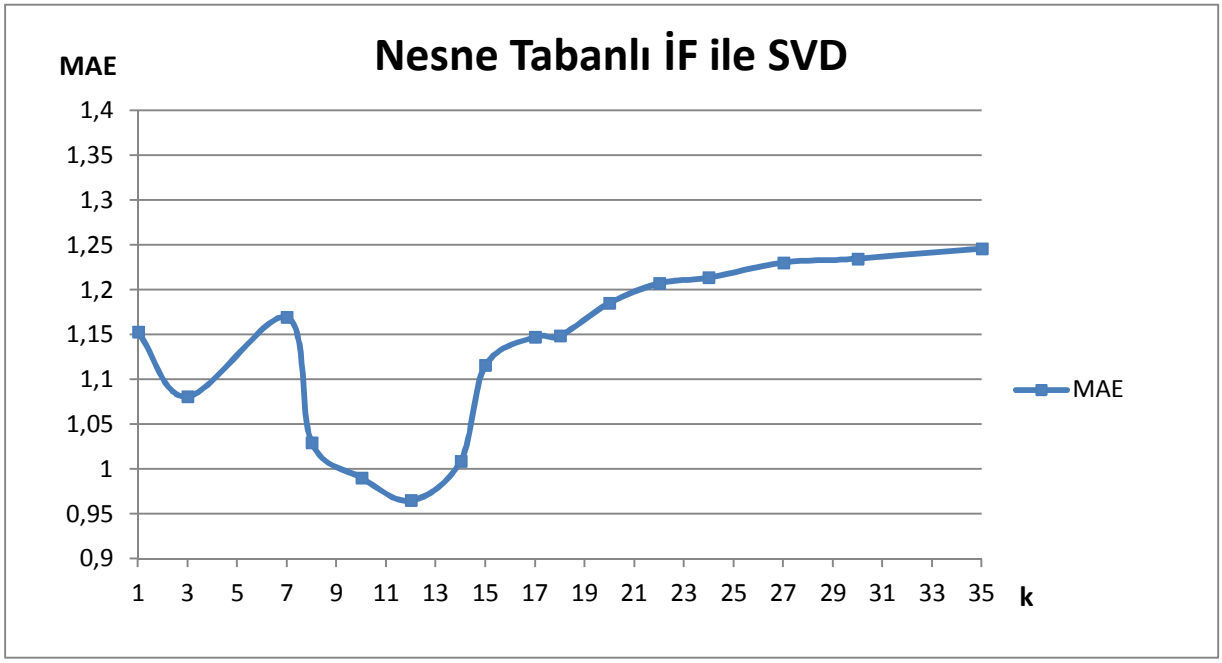
k değerleri	MAE sonuçları
1	1,15
3	1,08
7	1,16
8	1,02
10	0,99
12	0,96
14	1,00
15	1,11

Çizelge 7.25'de rastgele seçilmiş 100 test örneğinin nesne tabanlı İF tekniği ile SVD kullanımının düşük boyutlu k değerleri için sonuçları verilmiştir. Rastgele seçilmiş örnekler için klasik nesne tabanlı İF tekniğinin MAE sonucu 1,23 'dir. En minimum hata k=12 olduğunda MAE 0,96 ile bulunmuştur.

Çizelge 7.26 Rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımı (yüksek boyutlu k değerleri için)

k değerleri	MAE sonuçları
17	1,14
18	1,14
20	1,18
22	1,21
24	1,21
27	1,23
30	1,23
35	1,24

Çizelge 7.26' da rastgele seçilmiş 100 test örneğinin nesne tabanlı İF tekniği ile SVD kullanımının yüksek boyutlu k değerleri için sonuçları verilmiştir. k değeri arttıkça hata oranı artmaktadır.



Şekil 7.7 Rastgele seçilmiş 100 test örneği için nesne tabanlı İF tekniği ile SVD kullanımı

Şekil 7.7'de rastgele seçilmiş 100 test örneğinin nesne tabanlı İF tekniği ile SVD kullanımının **k** değerleri için sonuçları verilmiştir. Nesne tabanlı İF ile SVD kullanımı düşük boyutlu k değerlerinde daha iyi sonuç vermektedir.

Çizelge 7.27 Çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımı (düşük boyutlu k değerleri için)

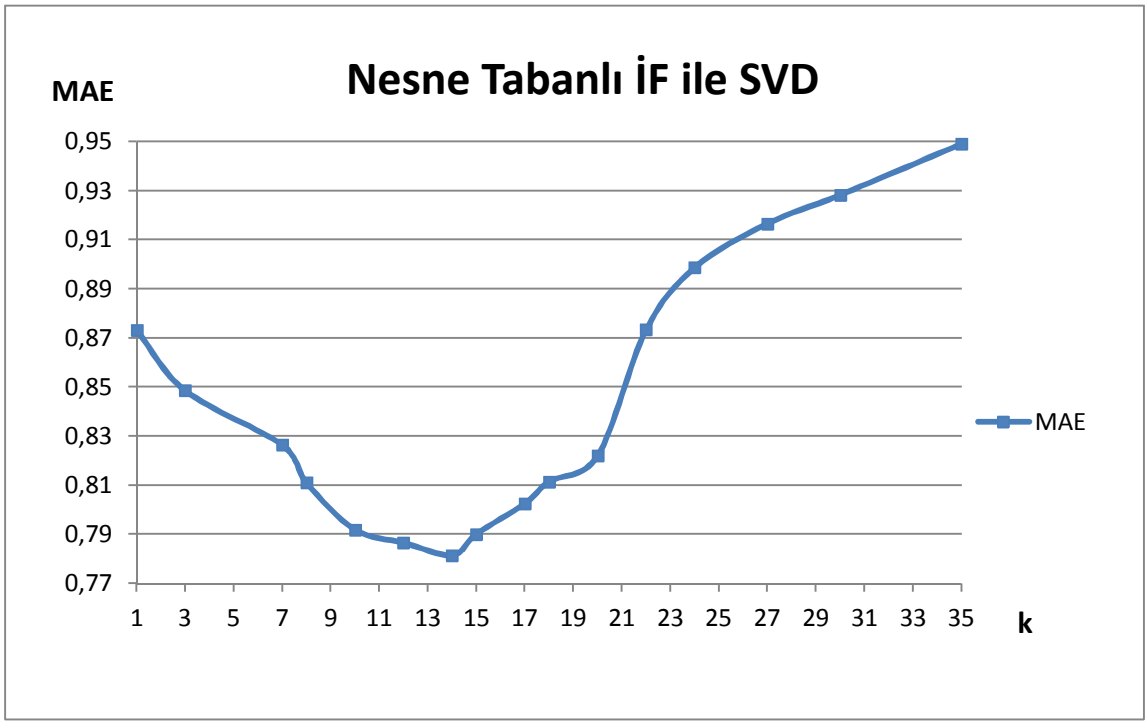
k değerleri	MAE sonuçları
1	0,87
3	0,84
7	0,82
8	0,81
10	0,79
12	0,78
14	0,78
15	0,79

Çizelge 7.27'de çok oylanmış filmlerin nesne tabanlı İF tekniği ile SVD kullanımının düşük boyutlu k değerleri için sonuçları verilmiştir. En minimum hata k=12 olduğunda MAE 0,78 ile bulunmuştur.

Çizelge 7.28 Çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımı (yüksek boyutlu k değerleri için)

k değerleri	MAE sonuçları
17	0,80
18	0,81
20	0,82
22	0,87
24	0,89
27	0,91
30	0,92
35	0,94

Çizelge 7.28' de çok oylanmış filmlerin nesne tabanlı İF tekniği ile SVD kullanımının yüksek boyutlu k değerleri için sonuçları verilmiştir. k değeri arttıkça hata oranı sürekli artmaktadır.



Şekil 7.8 Çok oylanmış filmler için nesne tabanlı İF tekniği ile SVD kullanımı

Şekil 7.8' de çok oylanmış filmlerin nesne tabanlı İF tekniği ile SVD kullanımının k değerleri için sonuçları verilmiştir. Nesne tabanlı İF ile SVD tekniği çok oylanmış filmler için yapılan analizinde düşük boyutlu k değerlerinde daha iyi sonuç vermektedir.

Çizelge 7.29 Tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımı (düşük boyutlu k değerleri için)

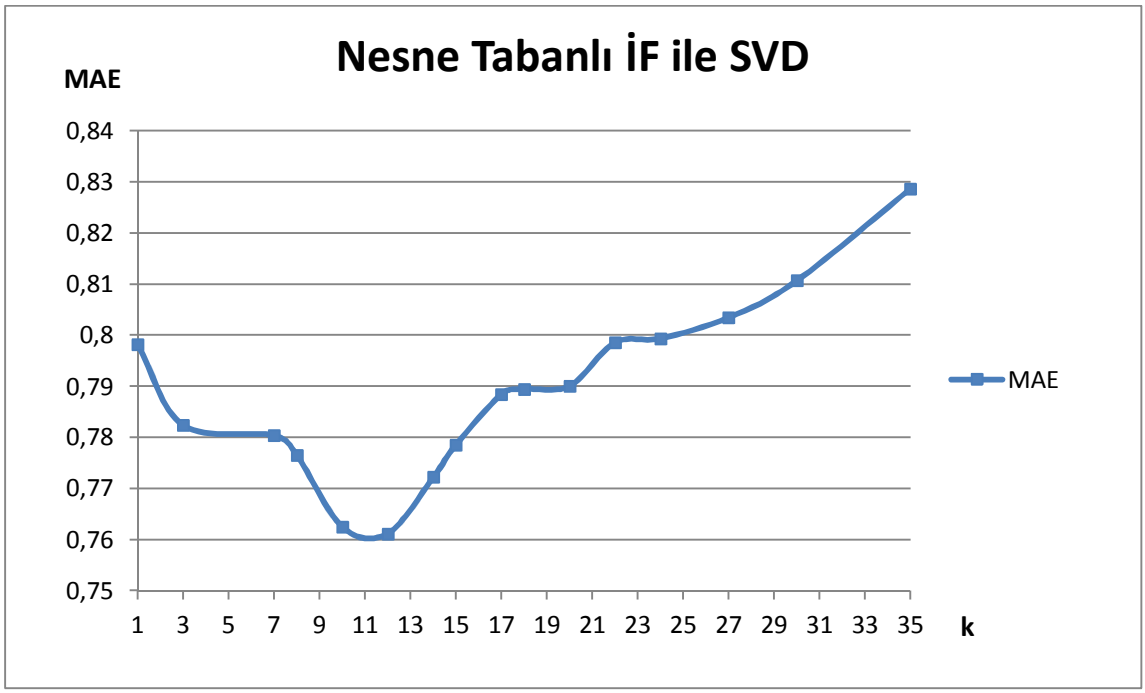
k değerleri	MAE sonuçları
1	0,79
3	0,78
7	0,78
8	0,77
10	0,76
12	0,76
14	0,77
15	0,77

Çizelge 7.29' da tüm test verisinin nesne tabanlı İF tekniği ile SVD kullanımının düşük boyutlu k değerleri için sonuçları verilmiştir. Tüm test verisine klasik nesne tabanlı İF tekniğinin MAE sonucu 0,84'dir. En minimum hata k=12 olduğunda MAE 0,76 bulunmuştur.

Çizelge 7.30 Tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımı (yüksek boyutlu k değerleri için)

k değerleri	MAE sonuçları
17	0,78
18	0,78
20	0,79
22	0,79
24	0,79
27	0,80
30	0,81
35	0,82

Çizelge 7.30' da tüm test verisinin nesne tabanlı İF tekniği ile SVD kullanımının yüksek boyutlu k değerleri için sonuçları verilmiştir. k değeri arttıkça hata oranı artmaktadır.



Şekil 7.9 Tüm test verisi için nesne tabanlı İF tekniği ile SVD kullanımı

Şekil 7.9'da tüm test verisinin nesne tabanlı İF tekniği ile SVD kullanımının k değerleri için sonuçları verilmiştir. Grafiğe göre küçük k değerlerinde başarı daha yüksektir. k değeri arttıkça öneri niteliği ve doğruluğu azalır.

Çizelge 7.31 k=12 optimum değer için nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen MAE sonuçları

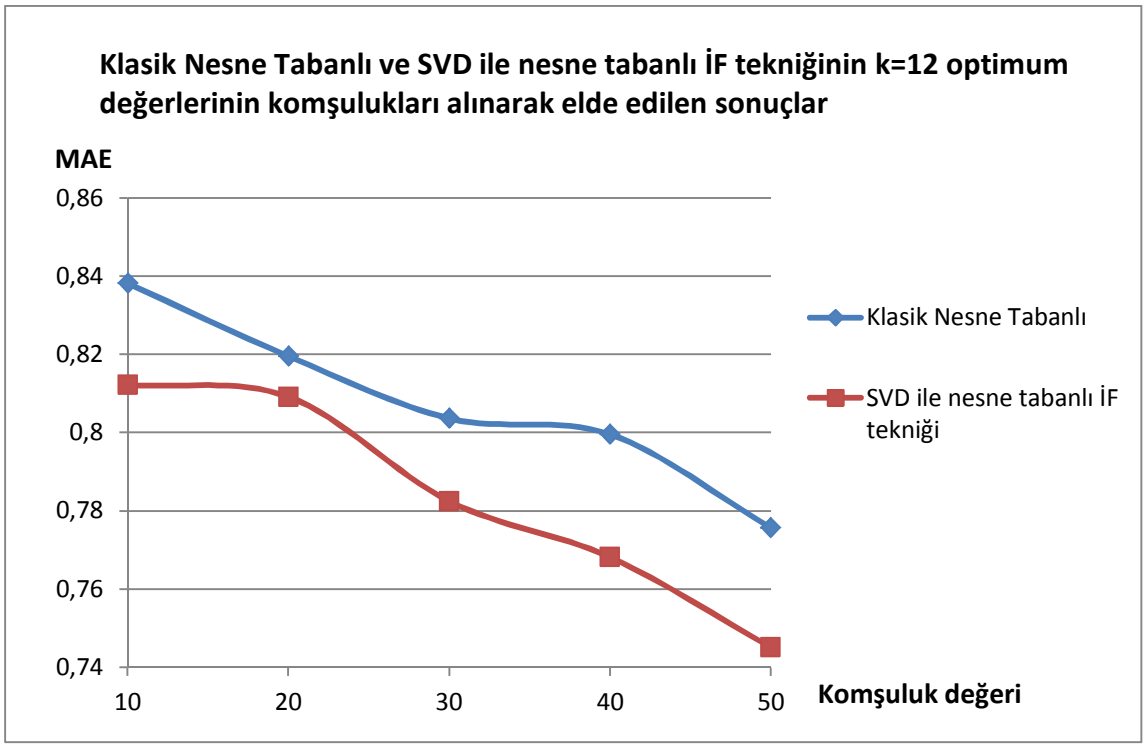
Komşuluk	**MAE**
10	0,81
20	0,80
30	0,78
40	0,76
50	**0,74**

Çizelge 7.31'de k=12 optimum değeri için nesne tabanlı İF tekniğinin komşuluklarını alarak elde edilen SVD sonuçları gösterilmiştir. Komşuluk arttıkça hata oranı azalmaktadır.

Çizelge 7.32 Klasik nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen MAE sonuçları

Komşuluk	**MAE**
10	0,83
20	0,81
30	0,80
40	0,79
50	**0,77**

Çizelge 7.32'de klasik nesne tabanlı İF tekniğinin çeşitli komşulukları alınarak elde edilmiş sonuçları gösterilmiştir.



Şekil 7.10 Klasik nesne tabanlı İF tekniği ve k=12 optimum değerinin komşulukları alınarak elde edilen MAE sonuçları

Şekil 7.10'da klasik nesne tabanlı İF ile SVD uygulanarak elde edilen optimum k değerinin komşulukları alınarak elde edilmiş sonuçları gösterilmiştir. Grafikte de görüldüğü gibi SVD uygulayarak daha doğru öneriler verilmektedir. SVD uygulamak hem veri seyrekliğini hem de veri ölçekleme probleminin üstesinden gelir.

Çizelge 7.33 k değerlerine bağlı olarak puan tahminde bulunurken geçen süre(düşük boyutlu için)

k değerleri	**Süre (saniye)**
1	612,00
3	614,89
7	623,12
8	626,72
10	631,60
12	637,08
14	643,83
15	647,73

Çizelge 7.34 k değerlerine bağlı olarak puan tahminde bulunurken geçen süre(yüksek boyutlu için)

k değerleri	**Süre (saniye)**
17	649,80
18	650,63
20	653,63
22	657,79
24	660,48
27	664,66
30	669,70
35	677,25

Çizelge 7.33'de ve Çizelge 7.34'de SVD ile nesne tabanlı İF tekniğinin tüm test verisi için puan tahmininde bulunurken **k** değerlerine bağlı olarak aldıkları süreler verilmektedir. K değeri arttıkça hesaplama süresi de artmaktadır.

Çizelge 7.35 Klasik nesne tabanlı İF tekniklerinde puan tahmininde bulunurken geçen süre

Komşuluk	**Süre (saniye)**
10	1487,65
20	1496,87
30	1584,65
40	1671,65
50	1796,87

Çizelge 7.35'de ise klasik nesne tabanlı İF tekniğinin tüm test verisi için puan tahmininde bulunurken aldıkları süreler verilmektedir. Elde edilen sonuçlara göre SVD uygulayarak nesne tabanlı İF tekniğinin veriye uygulanması zaman olarak da kısa sürede tamamlanmıştır. Klasik nesne tabanlı İF tekniğine göre 2 kat daha erken sürede hesaplanmıştır.

7.5 Netflix Veri Kümesi için Uygulama Sonuçları

Netflix veri kümesinin 2005 yılını test kümesi olarak kullanıp yapılan deneysel çalışmada klasik nesne tabanlı İF tekniğinin RMSE sonucu 1,97'dir ve yüzdelik hata oranı %39,52 hesaplanmıştır.

Çizelge 7.36 Klasik nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen RMSE sonuçları

Komşuluk	RMSE	Yüzdelik Hata(%)
100	1,94	38,95
200	1,85	37,13
300	1,81	36,28
400	1,79	35,92
500	**1,77**	**35,47**

Çizelge 7.36'da klasik nesne tabanlı İF tekniğinin komşulukları alınarak elde edilen RMSE sonuçlarında hata oranı giderek azalmaktadır. Komşuluk değeri arttıkça hata oranında sürekli bir azalış gözlemlenmektedir.

Çizelge 7.37 Nesne tabanlı İF tekniği ile SVD kullanımı

k değerleri	RMSE	Yüzdelik Hata(%)
5	1,71	34,2
10	1,67	33,4
20	1,60	32,0
30	1,51	30,2
35	**1,49**	**29,9**
50	1,68	33,7
100	1,78	35,6
150	1,94	38,8
200	2,14	42,8

Çizelge 7.37'de test verisinin nesne tabanlı İF tekniği ile SVD kullanımının çeşitli k değerleri için sonuçları verilmiştir. Şekil 7.10'da tüm test verisinin kullanıcı tabanlı İF tekniği ile SVD kullanımının k değerleri için sonuçları verilmiştir. Grafiğe göre küçük k değerlerinde başarı daha yüksektir. k değeri arttıkça öneri niteliği ve doğruluğu azalır. En minumum hata k=35 olduğunda yaklaşık %29 bulunmuştur.

Çizelge 7.38 n=0,75 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası

n=0,75		
Age_w için Değişim Aralığı	**Zamana Bağlı Nesne Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,75...2,0]	**1,46**	**29,34**
[0,75...1,7]	1,50	30,17
[0,75...1,5]	1,57	31,42
[0,75...1,4]	1,62	32,52
[0,75...1,3]	1,71	34,36
[0,75...1,2]	1,82	36,57
[0,75...1,1]	1,93	38,75

Çizelge 7.38 'de n=0,75 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin başarısı gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında bir artış gözlemlenmektedir. En minimum hata $\mathrm{Age_w}$=2 olduğunda yaklaşık %29 bulunmuştur.

Çizelge 7.39 n=0,5 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası

n=0,5		
Age_w için Değişim Aralığı	**Zamana Bağlı Nesne Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,5...2,0]	1,42	28,57
[0,5...1,7]	1,41	28,20
[0,5...1,5]	1,40	28,13
[0,5...1,4]	1,37	27,54
[0,5...1,3]	1,32	26,54
[0,5...1,2]	1,30	26,12
[0,5...1,1]	**1,21**	**24,36**

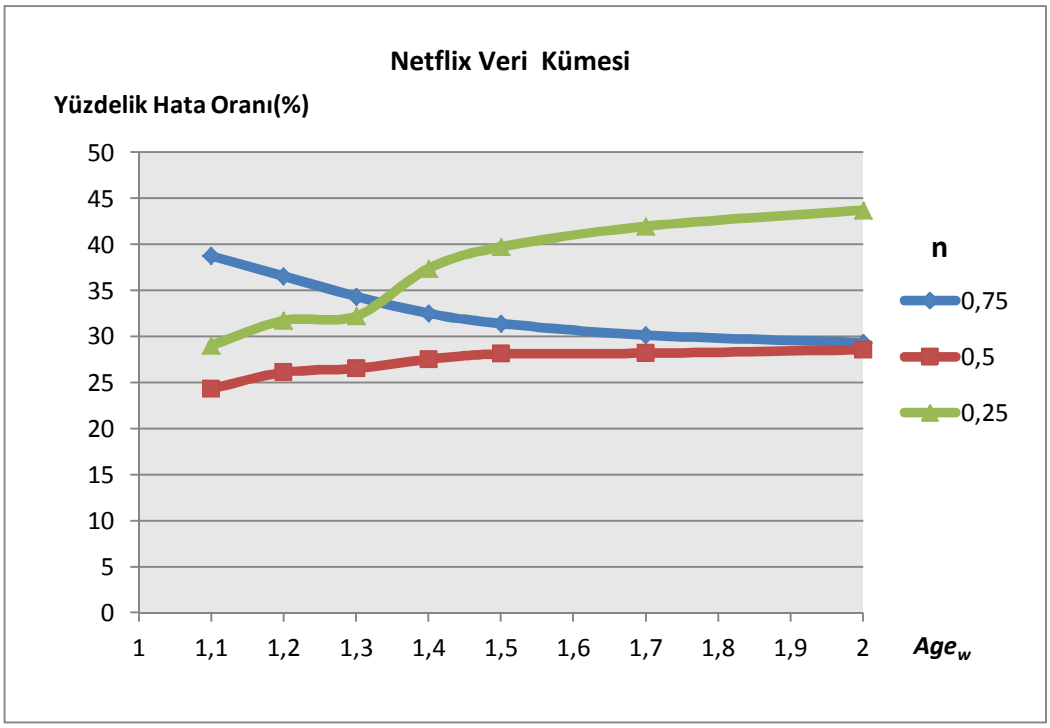
Çizelge 7.40 n=0,25 dönüşüm fonksiyonu kullanılarak elde edilen puan tahmin hatası

n=0,25		
Age_w için Değişim Aralığı	**Zamana Bağlı Nesne Tabanlı İF tekniği için RMSE**	**Yüzdelik Hata (%)**
[0,25...2,0]	2,18	43,75
[0,25...1,7]	2,09	41,98
[0,25...1,5]	1,98	39,75
[0,25...1,4]	1,87	37,41
[0,25...1,3]	1,61	32,24
[0,25...1,2]	1,58	31,75
[0,25...1,1]	**1,45**	**29,05**

Çizelge 7.39 'da n=0,5 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin hatası gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında

sürekli bir azalış gözlemlenmektedir. En minimum hata $\mathrm{Age_w}$=1,1 olduğunda yaklaşık %24 bulunmuştur.

Çizelge 7.40 'da n=0,25 sabit tutularak farklı dönüşüm aralıkları için elde edilen RMSE tahmin başarısı gösterilmiştir. Dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış gözlemlenmektedir. En minimum hata $\mathrm{Age_w}$=1,1 olduğunda yaklaşık %29 bulunmuştur.



Şekil 7.11 Zamana bağlı nesne tabanlı İF tekniğinin sonuçları(netflix veri kümesi için)

Şekil 7.11' de zamana bağlı nesne tabanlı İF çalışmasının az filme puan veren kullanıcılar grubu için sonuçları verilmiştir. Grafikte yeşil renkle gösterilen n=0,25 için verilen eğri Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında sürekli bir azalış olduğu görülmektedir. Mavi renkte gösterilen n=0,75 için verilen eğride ise Age_w dönüşüm fonksiyonunda eğim azaldıkça hata oranında artış olduğu gözlemlenmektedir. Kırmızı renkte gösterilen n=0,5 için verilen eğride Age_w dönüşüm fonksiyonu 1,1 olduğunda yaklaşık %29 ile hata oranının en düşük olduğu değer elde edilmektedir.

Çizelge 7.41 Age_w[0,5…1,1] için SVD sonuçları

k değerleri	RMSE	Yüzdelik Hata(%)
5	1,12	22,2
10	0,98	19,7
20	0,87	17,5
30	0,77	15,5
35	0,70	14,8
50	0,89	17,9
100	1,02	20,5
150	1,28	25,6
200	1,57	31,5

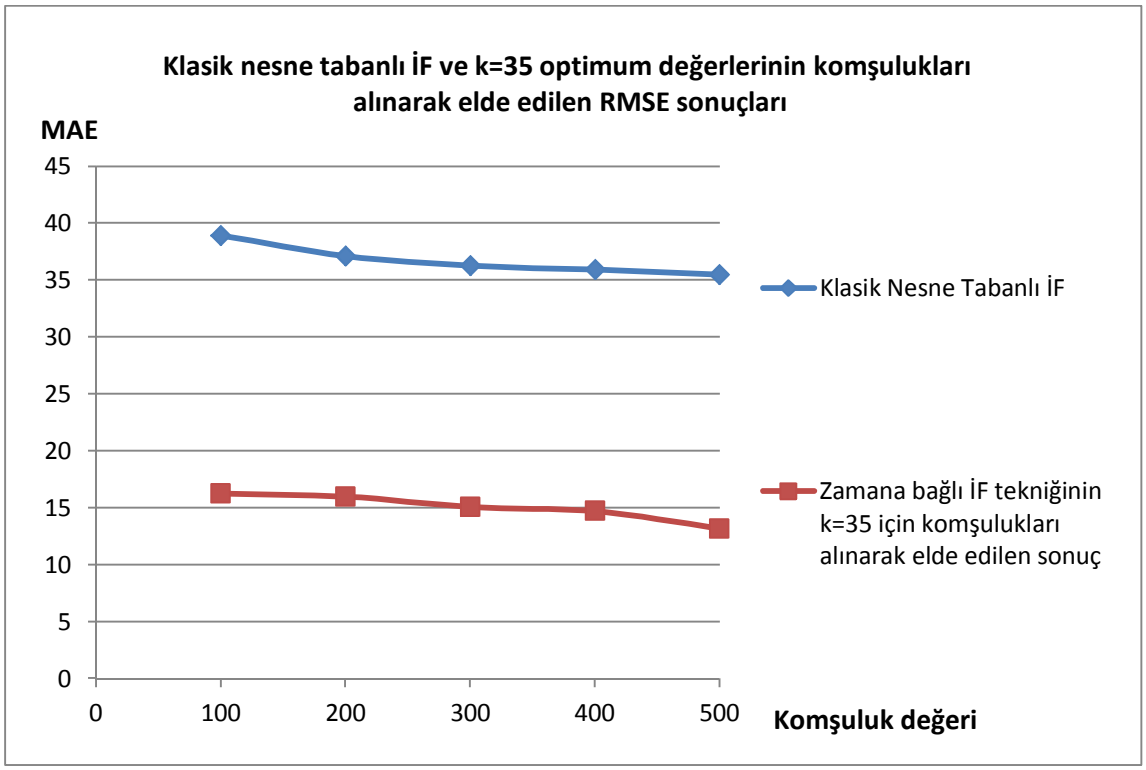
Çizelge 7.41 'de en doğru önerilerinin yapıldığı Age_w[0,5…1,1] dönüşüm aralığına SVD uygulayarak elde dilen sonuçlar gösterilmiştir. Hem SVD hem de zamansal dönüşüm katsayısı kullanarak elde edilen başarının arttırılması sağlanmıştır.

Çizelge 7.42 k=35 için komşulukları alınarak elde edilen puan tahmin hatası

Komşuluk	RMSE	Yüzdelik Hata(%)
100	0,81	16,25
200	0,79	15,98
300	0,75	15,07
400	0,73	14,71
500	**0,65**	**13,17**

Çizelge 7.42'de hem SVD hem de zamansal dönüşüm fonksiyonu kullanarak elde edilen en doğru k'nın 35 değeri için çeşitli komşulukları verilmiştir. Komşulukları alınarak elde

edilen RMSE sonuçlarında hata oranı giderek azalmaktadır. Komşuluk değeri arttıkça hata oranında azalış gözlemlenmektedir.



Şekil 7.12 Klasik nesne tabanlı İF ve k=35 optimum değerinin komşulukları alınarak elde edilen RMSE sonuçları

Şekil 7.12'de zamana bağlı nesne tabanlı İF tekniğinin puan başarısının en iyi çıktığı Age_w[0,5…1,1] aralığına SVD uygulanmasının k=35 için komşuluklarının alınmasıyla elde edilen sonuçların klasik nesne tabanlı İF tekniğine göre daha başarılı olduğu görülür.

BÖLÜM 8

SONUÇ VE ÖNERİLER

İnternet kullanımının hızla yayılması ve bunun sonucunda web içeriğindeki sürekli artış kullanıcıları daha çok ilgilendirecek kişiselleştirilmiş ürün ve servis sunma ihtiyacı doğurmuştur. Bunun sonucunda her şirket e-ticaretin gelişmesiyle birlikte müşterilerine web üzerinden hizmet sunarak karını arttırmayı hedeflemektedir. Ancak her müşterinin tercihi zamanla birlikte değişiklik göstermektedir. Benzer olarak ürünlerin popülerliği de sürekli değişmektedir. Bu durumda sistemde çok fazla kullanıcı ve çok fazla ürünün olması sebebiyle hem hızlı hem de kişinin ilgisini canlı tutarak kişiye özgü tavsiyelerde bulunmak oldukça zorlaşır. Bu çalışmayla müşterilerin zamanla değişen isteklerine, zevklerine hızlı ve doğru cevap verebilmek için yeni bir yöntem geliştirilmiştir. Geleneksel İF tekniklerinde olduğu gibi sadece kullanıcıların geçmişte verdikleri puanlara bakarak kullanıcıların tercihlerini, eğilimini kullanıcılar ya da ürünler arasındaki benzerlikleri bularak belirlemez. Kullanıcıların puanlarının yaşlarına bakarak kullanıcıların ürünlere vereceği puan tahmin edilmeye çalışılmıştır. Çeşitli dönüşüm fonksiyonları kullanılarak mevcut puanlar yaşlandırılarak ya da kuvvetlendirilerek kullanıcıların zevklerini, isteklerini belirleyebilecek en doğru aralık bulunmaya çalışılmıştır. Zamana bağlı kullanıcı kullanıcı tabanlı İF tekniğinde en başarılı dönüşüm fonksiyonu n=0,5 için Age_w 1,5 olarak bulunmuştur ve RMSE puan tahmin hatası %15 olarak hesaplanmıştır. Zamana bağlı nesne tabanlı İF tekniğinde en başarılı dönüşüm fonksiyonu n=0,5 için Age_w 1,4 olarak bulunmuştur ve RMSE puan tahmin hatası %25 olarak hesaplanmıştır. Zamana bağlı hibrit tabanlı İF tekniğinde ise en başarılı dönüşüm fonksiyonu n=0,5 için Age_w 1,3 olarak bulunmuştur ve RMSE puan tahmin hatası %15 olarak hesaplanmıştır.

Zamana bağlı İF tekniği geleneksel İF teknikleriyle karşılaştırıldığında yaklaşık %15'den fazla başarı sağlanmıştır. Geliştirilen sistem sayesinde zamana bağlı olarak mevcut puanların yaşına bakılarak tahmin başarısı arttırılarak kullanıcılara daha doğru öneriler sunulmuştur. Ayrıca SVD ile nesne tabanlı İF tekniği birleştirilerek hem puan tahmin başarısı arttırılmış hem de kullanıcılara daha hızlı öneriler verilmesi sağlanmıştır. SVD ile birleştirilmiş nesne tabanlı İF tekniği ve geleneksel nesne tabanlı İF tekniğinin çeşitli komşulukları alınarak hesaplanan puan tahmin hatası karşılaştırılmıştır. Movielens verisi üzerinde yapılan uygulamalara göre nesne tabanlı İF tekniği ile SVD yönteminin k=12 için 50 komşuluğu alınarak elde edilen MAE puan tahmin hatası 0,74 hesaplanırken, klasik nesne tabanlı İF tekniğinin MAE tahmin hatası 0,84 olarak hesaplanmıştır. Her iki yöntem süre olarak karşılaştırıldığında ise nesne tabanlı İF tekniği ile SVD yöntemi hem daha hızlı hem daha doğru önerilerde bulunmuştur. Ayrıca netflix verisi üzerinde nesne tabanlı İF, nesne tabanlı ile SVD, zamana bağlı nesne tabanlı İF, zamana bağlı İF ile SVD yöntemlerinin RMSE puan tahmin hatası hesaplanmıştır. SVD ile zamansal değişimler birleştirildiğinde puan tahmin hatası %14 olarak bulunmuştur.

Geliştirilen sistem ile birlikte kullanıcıların zamanla değişen isteklerine, beğenilerine daha doğru ve hızlı cevap verilmiştir. Böylelikle hem kullanıcı hem de firma bakımından daha güvenilir bir öneri sistemi geliştirilmiştir.

Zamana Bağlı Kullanıcı ve Nesne Tabanlı İF ve User Based Collaborative Filtering with Temporal Dynamics Using Age of Scores çalışmalarında da zamana bağlı İF yöntemlerinin kullanıcılara verilen tavsiye niteliğindeki sonuçları gösterilmiştir [73,74]. Yapılan çalışmalarla kullanıcılara verilen tavsiyelerin başarısı gösterilmiştir.

KAYNAKLAR

[1] Koren Y., (2009). “Collaborative Fİltering with Temporal Dynamics,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 447-456.

[2] D.Goldberg, D.Nichols, B.M Oli ve D.Terry, (1992). "Using collaborative filtering to weave an information tapestry", Communications of ACM, 35(12): 61-70.

[3] P.Resnick, N.Iacovou, M.Suchak, P.Bengstrom ve J.Rield, (1994). "GroupLens:an open architecture for collaborative filtering of netnews", in Proceedings of the ACM Conference on Computer Supported Cooperative Work, 175-186.

[4] T.Hoffman, (2004). "Latent semantic models for collaborative filtering", ACM Transactions on Information Systems(TOIS), 22(1).

[5] K.Miyahara, M.J.Pazzani, (2000). "Collaborrative filtering with the simple Bayesian Classifier", PRICA'00 6th Pacific Rim International Conference on Artificial Intelligence, 1886: 679-689.

[6] L.H.Uygar, D.P.Foster, (1998). "Clustering Methods for collaborative filtering", AAAI Technicial Report.

[7] J.Rongfei, J.Maozhagn, L.Chao, (2010). "A new clustering method for collaborative filtering", Information Technology(ICNIT), 488-492.

[8] Rozie A., Peter H., (2014). “Swarm Collaborative Filtering through Fish School Search”, International Journal of Software Engineering and Its Applications, 8(3): 251-254.

[9] D.M.Pennock, H.Eric, S.Laurence ve C.L.Giles, (2000). "Collaborative filtering by Personality Diagnosis:A Hybrid Memory and Model-based approach", in Processings of the 16th Conference on Uncertainty in Artificial Intelligence(UAI'00), 473-480.

[10] X.Su ve T.M.Khsohgoftar, (2006). "Collaborative filtering for multi-class data using belief nets algorithms", in Proceedings of the Artificial Interrational Conference on Tools with Artificial Intelligence(ıctaı'06), 497-504.

[11] R.Greinemn, X.Su, B.Shen ve W.Zlow, (2005). "Structural extension to logistic regression:discriminative parameter learning of belief net classifiers," Machine Learning, 59(3): 297-322.

[12] T.Landaver, M.Littman ve Bell Communications Research, (1994). "Computerized cross-large document Retrieval using latent semantic indexing." USpatent 5301109.

[13] D.Billsus ve M.Pazzani, (1998). "Learning collaborative information filters", in Processings of the 15th International Conference on Machine Learning(ICML'98), 1998.

[14] Linden G., Smith B. ve York J., (2003). "Amazon.com Recommendations Item-to-Item Collaborative Filtering, " IEEE Distributed Systems Online, 76-80.

[15] Ziegler C.N, Lausen G. ve Thieme L.S, (2004). "Taxonomy-driven computation of product recommendation," 13th International Conference on Information and Knowledge Management (CIKM'04), 406-415.

[16] D.Y.Pavlov ve D.M.Pennock, (2002). "A maxiumum entropy approach to collaborative filtering in dynamic sparse, high-dimensional domains,"in Processing Systems, 1441-1448.

[17] R.Burke, (2002). "Hybrid recommender systems: Survey and experiments, user modeling and user-adapted Interaction", 12: 331-370.

[18] Goldberg K., Roeder T., Gupta D. and Perkins C., (2001). "Eigentaste:A Constant Time Collaborative Filtering Algorithm," Information Retrieval, 4(2): 133-151.

[19] Xiang L., Yuan Q., Zhao S. ve Chen L., (2014), "Temporal Recommendation on Graphs via Long-and Short-term Prefence Fusion", KDD'10 of the 16th ACM SIGKDD International Confer3ence on Knowledge discovery and data mining", 723.732.

[20] Luo C., Cai X ve Chowdhury N., (2014). "Self-training Temporal Dynamics Collaborative Filtering", PAKDD '14, 461-472.

[21] Shrot T., Rosenfeld A., Golbeck J. ve Kraus S., (2014). “CRISP-An Interruption management algorithm based on collaborative filtering”, CHI ’14, One of a Chind, 3035-3044.

[22] Choi K., Suh Y., (2013). “A new similarity function for selecting neighbors for each target item in collaborative filtering”, Knowledge based Systems, 37: 146-153.

[23] W.Cohen, W.Fan, (2000). "Web-collaborative filtering: Recommending music by crawling the Web", International Journal of computer and telecommunications, 33(1-6,): 685-698.

[24] Bobadilla J., Ortega F., Hernanda A. ve Alcala J., (2011). "Improving collaborative filtering recommender systems and perfomance using genetic algorithms", Knowledge-Based Systems, 1310-1316.

[25] Herlocker J.L., Konstan J.A, Borchers A. and Rield J., (1999). “An Algorithmic Framework for Performing Collaborative Filtering , ”ACM Conference on Research and Development Information Retrieval , 230-237.

[26] Breese J.S, Heckerman D., Kadie C., (1998). “Empricial Analysis of Predictive Algorithms for Colaborative Filtering,” UAI'98 Fourteeth Conference on Uncertainty in Artificial Intelligence, 43-52.

[27] Sarwar B., Karypis G., Konstan J. and Rield J. , (2001). "Item-based Collaborative Filtering Recommendation Algorithms,” in WWW'01: International Conference on World Wide Web, 285-295.

[28] Li Y., Lu L., Xuefeng L., (2005). “A hybrid collaborative method for multiple-interest and multiple-content recommendation in E-Commerce, ” International Journal of Expert System with Application, 28: 67-77.

[29] Zhang Z., Lin H., Liu K., Wu D., Zhang G. ve Lu J., (2013). “A hybrid fuzzy-based personalized recommender system for telecom products/services,” Elsevier Information Sciences, 117-129.

[30] Choi K., Yoo D., Kim G. ve Suh Y., (2012). "A hybrid online-product recommendation systems: Combining implicit rating-based collaborative filtering and sequential pattern analysis," Elseiver Elektronic Commerce Research and Applications, 309-317.

[31] M.Ankest, M.Breuning, (1999). H.Kriegel ve J.Sarder, "Optics:Ordering points to Identify the clustering structure",ACM SIGMOD International Conference on Management of data, 28(2): 49-60.

[32] M.Connar, J.Herlocker, (1999). "Clustering Items for Collaborative Filtering",ACM SIGIR Workshop on Recommender Systems, Berkeley.

[33] B.Mobasher, X.Jin ve Y.Zhou, (2003) "Semanticially Enhanced Collaborative Filtering on the Web", 1st European web mining forum(EWMF2003), 3209: 57-76.

[34] Meteren R., Somener M., (2000) "Using Content-based Filtering for Recommendation Systems", NetlinQ .

[35] Dejen M., (2008). "Recomender Systems and Conten-based filtering", Seminar in Electornic Business.

[36] Qui L., Berbasat I.,(2010). "A study of demografic embodiments of product recommendation agents in electronic commerce", Int.J.Human Computer Studies, 669-688.

[37] Yapriady B., Uitdenbogerd A.L, (2005). "Combining Demografic data with collaborative filtering for automatic music recommendation", Knowledge-based Intelligent Information and Engineering Systems, 3684: 201-207.

[38] Ansari A., Essengaier S., ve Kohli R., (2000). "Internet recommendation systems", Journal of Marketing Research, 37(3): 363-375.

[39] Balaboniovic M., Shoham Y., (1997). "Fab: Content-based collaborative recommendation", Comminications of the ACM, 40(3).

[40] Barragans-Martinez A., Costa-Montenegro E., Burguillo J.C ,(2010). "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition", Information Sciences, 180(22): 4290-4311.

[41] Gong J.G., (2010). "A Collaborative Filtring Recommendation Algorithm Based on User Clustering and Item Clustering," Academy Publisher, 5(7): 745-752.

[42] Ahn H.J, (2008). "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences, 178(1): 37-51.

[43] Nguyen A., Denos N., Berrut C., (2007). “Improving New user Recommendations with rule-based induction Cold User Data ,” ACM Conference on Recommender Systems, 121-128.

[44] Schein A.I, Popescul A., Ungar L.H ve Pennock D.M, (2002). “Methods and metrics for cold-start recommendations, ”25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(sıgır'02).

[45] B.M.Sarwar, G.Karypis, J.Konstan ve J.Rield, (2004). "Incremental SVD-Based algorithms and for highly Scaleable Recommender Systems", 5th International Conference on Computer and Information Techonology (ICCIT'02).

[46] Deerwester S., Dumais S.T, Furnas G.W, Landaver T.K ve Harshman R., (1990). "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41(6): 391-407.

[47] Hoffman T., (2003). "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis", ACM SIGIR Conference on Research and development Information Retrieval, 259-266.

[48] Herlocker J.L, Konstan J.A, Terveen L.G and Rield J.T, (2004). “Evaluating Collaborative Filtering Recommender Systems ,” CAM Transactions on Information Systems, 22: 5-53.

[49] M.G.Vozalis ve K.G.Margaritis, (2005). "Applying SVD on Item-based filtering", Intelligent Systems Design and Applications 5th International Conference(ISDA205), 464-469.

[50] Netflix veri kümesi, https://www.netflix.com/global, 5. Mayıs. 2013.

[51] Gong S., Ye H., Dai Y., (2009). "Combining Singular Value Decomposition and Item-based Recommender in Collaborative Filtering", WKDD'09 of the 2009 Second International Workshop on Knowledge Discovery and Data Mining, 769-772.

[52] MovieLens veri kümesi, http://grouplens.org/datasets/movielens/ , 5.Mayıs.2013.

[53] Su X., Khoshgoftaar T.M. ,(2009). “A Survey of Collaborative Filtering Techniques,” Advances in Artificial Intelligence, 3: 1-20.

[54] Huang Z., Chen H., Zeng D., (2004). "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collabrative Filtering", ACM Transactions on Information Systems, 22(1): 116-142.

[55] Ma H.,King I., Lyu M.R, (2007). "Effective Missing Data Prediction for Collabrative Filtering", ACM SIGIR Conference on Reseach and Development in Information Retrieval, 39-46.

[56] Karen H.L, Tso-Sutter, Leandro B.M ve Lars S.T, (2008). "Tag-aware Recommender Systems by Fusion of Collabrative Filtering Algorithms", SAC' 08:Processing of the 2008 ACM Symposium on Applied Computing, 1995-1999.

[57] Chang T., Hsiao W. ve Chag W., (2014). "An ordinal regression model with SVD hebbian learning for collaborative recommendation", Journal of Information Science and Engineering, 387-401.

[58] Gao M., Wu Z., Jiang F., (2011). "Userrank for item-based Collabrative Filtering Recommendation", *Inf.Process.Lett.111,* 9: 440-446.

[59] G.Adomavicius ve A.Tuzhilin, (2005)."Toward the next generation of Recommender Systems:A Survey of state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering*,* 17(6): 734-749.

[60] K.Yu, A.Schwaighofer, V.Tresp, X.Xu veH.P.Kriegel, (2004). "Probabilistic memory-based Collabrative Filtering", IEEE Transactions on Knowledge and Data Engineering*,* 16(1): 56-69.

[61] L.Mendo, (2009). "Estimation of a probability with guaranted normalized mean absolute error", *IEEE Communications Letters,* 3:11.

[62] Luo X., Xia Y. ve Zhe Q., (2013). "Applilying the learning rate adaptation to the matrix factorization based collaborative filtering", Knowledge-Based Systems, 37:154-164.

[63] D.Kim, B.Yum, (2005). "Collaborative Filtering based on iterative principal component analysis", Expert Systems with Applications*,* 823-830.

[64] Kim H., Ha I., Lee K. ,(2011). "Collaborative user modeling for enhanced content filtering in recommender systems", Decision Support Systems*,* 51(4): 772-781.

[65] Kim D., Yum B., (2005). "Collaborative filtering based on iterative principal component analysis", Expert Sytems with Applications*,* 28(4): 823-830.

[66] Lee j., Jun C. , (2005). "Classification-based collaborative filtering using market basket data", Expert Systems with Applications, 29(3): 700-704.

[67] Paterek A., (2007). "Improving regularized singular value decomposition for collaborative filtering", KDD Cup and Workshop.

[68] Zhang S., Wang W., Ford J.,(2005). "Using Singular Value Decomposition approximation for collaborative filtering", Seventh IEEE International Conference on E-Commerce Technology, 257-264.

[69] Zhang C., Liu L., Yan X., (2014). "Information via collaborative user clustering modeling", 396:195-203.

[70] Yang X., Liu P., (2014). "Collaborative Filtering Recommendation using Matrix Factorization: A MapReduce Implementation", International Journal of Grid Distributed Computing, 7(2): 1-10.

[71] Lin C., Xie R., Guan X. ve Li T., (2014)."Personalized news recommendation via implicit social experts, Information Sciences, 1-18.

[72] Si S., Sarma A., (2014)."Beyond Modeling private actions: Predicting Social Shares", International World Wide Web Conference Committee (IW3C2) ,377-378.

[73] Çiğdem B., Songül A., (2014). "Zamana Bağlı Kullanıcı ve Nesne Tabanlı İşbirlikçi Filtreleme", Sinyal İşleme Kurultayı, SİU'14, Trabzon.

[74] Çiğdem B., Songül A., (2014). "User Based Collaborative Filtering With Temporal Dynamics Using Age of Scores", 14th Industrial Conference on Data Mining, ICDM'14, St.Petersburg, Rusya.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı :Çiğdem BAKIR

Doğum Yeri :Eminönü

Yabancı Dili :İngilizce

E-posta :cigdem@ce.yildiz.edu.tr

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Lisans	Bilgisayar Müh.	Sakarya Üniversitesi	10.06.2010

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2013	Yıldız Teknik Üniversitesi	Araştırma Görevlisi
2011	Iğdır Üniversitesi	Araştırma Görevlisi
2011	Yalova Üniversitesi	Bilgisayar Yazılım
2009	Urlu Sağlık Hizmetleri	Stajer

2008	Çolaksoy Teknoloji	Donanım
2007	Mavi Eğitim Kurumları	Yazılım

YAYINLARI

Bildiri

1. Çiğdem B., Songül A., (2014). "Zamana Bağlı Kullanıcı ve Nesne Tabanlı İşbirlikçi Filtreleme", Sinyal İşleme Kurultayı, SİU'14, Trabzon.

2. Çiğdem B., Songül A., (2014). "User Based Collaborative Filtering With Temporal Dynamics Using Age of Scores", 14th Industrial Conference on Data Mining, ICDM'14, St.Petersburg, Rusya.