Dicle University Journal of Engineering

Volume 12 | Issue 2 Article 4

2021

The Analysis of Similarity Measurements in Content-Based Scientific Paper Recommender System

Emine Deniz
Eskisehir Osmangazi University, denzemine@gmail.com

V. Karani Öz
Eskisehir Osmangazi University, veyskoz@gmail.com

Sinem BOZKURT KESER
Eskisehir Osmangazi University, sbozkurt@ogu.edu.tr

Savaş Okyay
Eskisehir Osmangazi University, osavas@ogu.edu.tr

Yusuf Kartal
Eskisehir Osmangazi University, ykartal@ogu.edu.tr

Recommended Citation

Deniz, Emine; Öz, V. Karani; BOZKURT KESER, Sinem; Okyay, Savaş; and Kartal, Yusuf (2021) "The Analysis of Similarity Measurements in Content-Based Scientific Paper Recommender System," *Dicle University Journal of Engineering*: Vol. 12 : Iss. 2 , Article 4.
DOI: 10.24012/dumf.838084
Available at: https://duje.dicle.edu.tr/journal/vol12/iss2/4

Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi / Research Article

İçerik Tabanlı Bilimsel Yayın Öneri Sisteminde Benzerlik Ölçümlerinin İncelenmesi

The Analysis of Similarity Measurements in Content-Based Scientific Paper Recommender System

Emine Deniz[1], V. Karani Öz[2], Sinem Bozkurt Keser[3*], Savaş Okyay[4], Yusuf Kartal[5]

[1] Eskişehir Osmangazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, denzemine@gmail.com
[2] Eskişehir Osmangazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, veyskoz@gmail.com
[3] Eskişehir Osmangazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, sbozkurt@ogu.edu.tr
[4] Eskişehir Osmangazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, osavas@ogu.edu.tr
[5] Eskişehir Osmangazi Üniversitesi, Bilgisayar Mühendisliği Bölümü, Eskişehir, ykartal@ogu.edu.tr

MAKALE BİLGİLERİ

Makale geçmişi:

Geliş: 9 Aralık 2020
Düzeltme: 15 Şubat 2021
Kabul: 20 Şubat 2021

Anahtar kelimeler:

Akademik makale öneri sistemi, ters doküman frekansı, içerik tabanlı filtreleme.

ÖZ

Öneri sistemleri, kullanıcılara kişiselleştirilmiş öneri sunan bilgi filtreleme sistemleridir. Öneri tabanlı uygulamalar *e*-ticaret, film, makale, restoran ve seyahat gibi birçok alanda kullanılmaktadır. Özellikle metin tabanlı depolamaya sahip sistemler üzerinde geleneksel anahtar kelime tabanlı arama tekniğiyle karşılaştırıldığında, öneri sistemleri büyük veri için daha etkili ve özelleştirilmiş sistemler olarak ön plana çıkmaktadır. Bilimsel çalışma paylaşımının yapıldığı platformlarda içerik havuzunun genişlemesiyle birlikte metinsel veri kullanımında önemli artış görülmektedir. Bu durum, araştırmacıların kendi alanlarıyla ilgili güvenilir ve doğru yayınlara erişimini zorlaştırmaktadır. Araştırmacılar, çalışmalarına katkı sağlayacak en doğru yayınları bulmakta çok fazla zaman harcayabilmektedir. Bilimsel çalışma öneri sistemi, karşılaşılan bu sorunlara çözüm üreterek araştırmacılara ilgi alanlarına uygun yayınları hızlı bir şekilde bulmalarına yardımcı olmaktadır. Sınırlı deneyime sahip kişiler için bilimsel çalışma öneri sistemleri araştırmacıların ufuklarını ve araştırma ilgi alanlarını genişletmeleri doğrultusunda yayınlar sunmaktadır. İçerik tabanlı filtreleme yöntemi, bilimsel çalışma öneri sistemi tasarımında en yaygın kullanılan yöntem olup kullanıcıdan bağımsız modellenir. Bu çalışmada, içerik tabanlı yeni bir bilimsel çalışma öneri sistemi farklı benzerlik yöntemleri üzerinden karşılaştırmalı olarak tavsiye edilmektedir. Yöntemler ve öneri ağırlıkları değişiklik gösterse de aynı veri seti içerisinde aynı yayınların önerildiği görülmektedir. Ölçüm değeri olarak birbirine yakın yöntemler arasında seçim yapmak gerektiğinde ise hesaplama süresini dikkate almak gerektiği sonucuna varılmaktadır.

Doi: 10.24012/dumf.838084

ARTICLE INFO

Article history:

Received: 9 December 2020
Revised: 15 February 2021
Accepted: 20 February 2021

Keywords:

Scientific journal recommender system, inverse document frequency, content-based filtering.

ABSTRACT

Recommender systems are information filtering systems that offer personalized suggestions to users. Recommender-based applications are used in numerous areas, such as e-commerce, streaming services, textual media, restaurants, and tourism. Compared to traditional keyword-based search techniques, especially on systems with text-based storage, recommender systems stand out as more effective and customized systems for big data. With the expansion of the content pool on platforms where scientific study sharing is made, there is a significant increase in textual data use. This situation makes it difficult for researchers to access reliable and accurate papers in their domains. Researchers may spend a lot of time finding the most accurate publications to contribute to their studies. The scientific study recommender system helps researchers quickly find papers that are suitable for their interests. Scientific study recommender system offers publications to researchers, who have limited experience, to broaden their horizons and research interests. The content-based filtering method is the most widely used procedure in scientific study recommender system designs and is modeled independently from the user. In this study, a new content-based scientific study recommender system is comparatively suggested over different similarity methods. Although the methods and recommender weights vary, it is seen that the same publications are recommended within the same data set. When it is necessary to choose between methods that are close to each other as measurement values, it is concluded that the calculation time should be considered.

* Sorumlu yazar / Correspondence
✉ sbozkurt@ogu.edu.tr

Please cite this article in press as E. Deniz, V. Karani Öz, S. Bozkurt Keser, S. Okyay, Y. Kartal, "İçerik Tabanlı Bilimsel Yayın Öneri Sisteminde Benzerlik Ölçümlerinin İncelenmesi", DUJE, vol. 12, Iss.2, pp. 221-228, March 2021.

Giriş

Öneri sistemleri kullanıcılara ilgilerini çekebilecek öğeler veya ürünler için anlamlı öneriler üreten programlar olarak tanımlanabilir. Web sayfalarını önermekten müziğe, kitaplara, filmlere, akademik yayınlara ve diğer tüketici ürünlerine kadar çeşitli alanlarda uygulamalar geliştirilmektedir [1]. Amazon'dan satın alınmak istenen alternatif öğeler veya Netflix'te dizi ya da filmler için gerçekleştirilen öneriler, endüstriye güç veren öneri sistemlerinin işleyişinin gerçek dünya örneklerinden birkaçıdır [2]. Öneri sistemlerinin mimarisi ve gerçek dünya problemlerine ilişkin uygulamaları aktif bir araştırma alanıdır. Öneri sistemleri, kullanıcılar ve öğeler arasında iyi eşleşmiş çiftleri tanımlamak üzere kullanılabilecek yakınlık kavramları geliştirmek için veri kaynakları analiz yöntemlerini kullanır. Öneri sistemleri, istatistik, makine öğrenimi, veri madenciliği ve bilgi erişiminin çeşitli alt disiplinleri ile kesişen, kişiselleştirilmiş deneyim tabanlı, aktif bir araştırma alanı olarak her geçen gün güçlenmektedir [3].

Son yıllarda, bilgi teknolojilerinde yaşanan gelişmeler ile birlikte öneri sistemleri, akademik alanda da büyük ivme kazanmıştır. Bilimsel çalışma öneri sistemleri ile ilgili literatürde yapılan çalışmalar incelendiğinde, çoğunlukla içerik tabanlı filtreleme yönteminin uygulandığı tespit edilmiştir. İçeriğe dayalı filtreleme yöntemini, işbirlikçi filtreleme ve karma yöntemler izlemektedir [4].

İşbirlikçi filtreleme terimi 1992 yılında Goldberg ve arkadaşları tarafından "insanlar filtreleme sürecine dahil olduğunda bilgi filtrelemenin daha etkili olabileceği" fikri ile öne sürülmüştür. Bu yöntemde, benzer tercihlere sahip kullanıcılara benzer öğeler önerilmektedir. İçerik tabanlı filtreleme ile karşılaştırıldığında, işbirlikçi filtreleme yönteminde hataya açık öğe işlemeye gerek yoktur. Derecelendirmeleri kullanıcılar yaptığı için gerçek kalite değerlendirmelerini dikkate alır. İşbirlikçi filtreleme yönteminin tesadüfi öneriler sunması beklenir çünkü öneriler öğe benzerliğine değil, kullanıcı benzerliğine dayanmaktadır. Kullanıcının katılımını gerektirir, ancak genellikle katılma motivasyonu düşüktür. Bu soruna "soğuk başlangıç" (yeni kullanıcılar, yeni öğeler, yeni topluluklar veya disiplinler) sorunu denir. Yeni bir kullanıcı birkaç öğeyi derecelendirirse veya hiç öğe derecelendirmezse, sistem öneriler sunamaz. Bir öğe sistemde yeniyse ve henüz en az bir kullanıcı tarafından derecelendirilmediyse önerilemez. Soğuk başlangıç sorununun üstesinden gelebilmek için, kullanıcılar ve öğeler arasındaki etkileşimlerden örtük derecelendirmeler çıkarılabilir [5].

İçerik tabanlı filtreleme yöntemi, kullanıcı profilini analiz ederek, başka bir ifadeyle kullanıcının geçmiş seçimlerini dikkate alarak, bu seçimlere benzer seçimlerin önerilmesi esasına dayanır [5]. Kullanıcıların ilgi alanlarının, etkileşimde bulundukları öğelerden çıkarıldığı kullanıcı modelleme sürecidir. Öğeler genellikle metinseldir. Etkileşim genellikle bir içeriği indirme, etiketleme veya ilgilenilen içeriğe yazma gibi eylemler yoluyla oluşturulur. Öğeler, öznitelikler içeren model ile temsil edilir. Öznitelikler genellikle kelime tabanlıdır. Bazı öneri sistemleri ayrıca yazma stili, düzen bilgisi veya XML etiketleri gibi metinsel olmayan öznitelikleri de kullanabilmektedir. Kullanıcı modeli, kullanıcı öğelerinin özniteliklerinden oluşur. Öneriler oluşturmak için kullanıcı modeli ve öneri adayları benzerlik yöntemleri kullanılarak karşılaştırılır [6]. İçerik tabanlı yöntemlerde diğer kullanıcılar hakkında bilgiye ihtiyaç duyulmaz. Ayrıca, veri seyrekliği (*data sparsity*) gibi sorunlardan kaçınılmış ve gizlilik sağlanmış olur. Bu yöntem, öğe ve kullanıcı tanımının bütünlüğüne (*completeness*) karşı çok hassastır [5].

Öneri sonuçlarının doğruluğunu artırmak için, bazı bilimsel çalışmalarda iki veya daha fazla öneri tekniği birleştirilmektedir [7]. İçerik tabanlı ve işbirlikçi filtreleme yöntemlerinin avantajları ile birlikte dezavantajları vardır; ikisinin beraber kullanılması birbirini tamamlayabilir. Karma yöntemleri kullanan öneri sistemleri, tekil algoritma içeren öneri sistemlerine nazaran genellikle daha doğru ve net sonuç verebilir. Bu doğrultuda farklı

yaklaşımların avantajlarını birleştirmek ve dezavantajlarını ortadan kaldırmak için ağırlıklı, kademeli veya karışık kombinasyon teknikleri kullanılarak karma yöntemler önerilmiştir. Bu yöntemin avantajı, farklı öneri yöntemlerinin ve birçok kaynaktan gelen bilgilerin bir arada kullanılabilmesidir [1].

Öneri sistemlerinde kullanılan yöntemlerden işbirlikçi filtreleme yöntemi, daha çok geçmiş etkileşimleri analiz ederken, içerik tabanlı filtreleme yöntemleri profil özniteliklerine dayalı olup; karma teknikler ise bu tasarımları birleştirmeye çalışır. Bu çalışmada, içerik tabanlı yeni bir bilimsel çalışma öneri sistemi farklı benzerlik yöntemleri üzerinden karşılaştırmalı olarak tavsiye edilmektedir.

İzleyen başlıklarda öncelikle bilimsel çalışma öneri sistemi ve bilimsel çalışma öneri sistemlerinde içerik tabanlı yöntemlerin uygulandığı çalışmalar anlatılmakta sonrasında ise çalışmada kullanılan veri setleri analiz edilmektedir. Takip eden bölümlerde ise sırasıyla önerilen yöntem ve testler verilmektedir. Test sonuçlarının analizinin ardından çalışma, sonuçlar ve gelecek çalışmalar ile sonlandırılmaktadır.

Metodoloji

Akademik topluluktaki kişiler yapılan yeni çalışmaların farkında olma, belirli bir konuda kapsamlı araştırma yapma, derleme makaleler yazma veya belirli bir alanda yeni bir araştırma yapma gibi nedenlerden dolayı literatür araştırması yapmaktadırlar. Akademik çalışmaların artmasıyla birlikte araştırmacılar ilgilendikleri alandaki çalışmayı bulabilmek için çok fazla kaynak taramak zorunda kalmaktadır. Bu kaynaklar arasındaki benzer sayıdaki çalışmaların fazlalığı araştırmacının seçim yapmasını zorlaştırmaktadır. Profile özgü yayınların seçiminde, araştırmacılar çok zaman harcayabilmektedir. Araştırmacının, belli bir konuda ilgili tüm çalışmalara ulaşabilmesi, doğru yayınları okuyarak doğru bilgiyi edinmesi, araştırma yaparken zamandan tasarruf etmesi, en önemlisi tek bir platform üzerinden farklı kaynaklardaki çalışmalara kolaylıkla erişebilmesi gibi ihtiyaçlar doğrultusunda bilimsel çalışma öneren bir sisteme ihtiyaç duyulmaktadır [8]. Bilimsel çalışma öneri sistemleri, araştırmacıların ilgi alanlarına ve araştırma odaklarına uygun bilgiyi karmaşık hesaplamalara sahip algoritmalar ile gereksiz veriyi filtreleyerek sunmaktadır [9]. Bu sistemlerin amacı, kullanıcıya gerekli olan veriyi farklı kaynakları kullanarak hızlı bir şekilde ulaştırmaktır. Günümüzde bilimsel çalışma öneri sistemleri vazgeçilmez bir araçtır.

Bilimsel çalışma öneri sistemlerinde, içerik tabanlı yöntemlerin uygulandığı çalışmalar literatürde sıklıkla karşılaşılmaktadır [10]. Bu tür çalışmalarda makale başlığı, özetçe, anahtar kelimeler, yazar bilgileri, vb. olmak üzere genellikle metinsel özellikler girdi olarak kullanılmaktadır. Bu özelliklerin kimi zaman birinin kimi zaman da birkaçının bir araya getirilmesi ile elde edilen varyasyonları öznitelik çıkarma adımı olarak ön işleme sürecini oluşturmaktadır [11]. Kullanıcıların geçmiş tercihleri dikkate alınır ve kullanıcı profili olarak adlandırılan ilgi alanı modelini çıkarmak için kişisel kütüphane oluşturulur. Daha sonra, kullanıcı profillerinden ve mevcut içerikten çıkarılan anahtar kelimelerin benzerliği hesaplanır. Benzerlik oranları sıralandıktan sonra, yüksek benzerliğe sahip makaleler kullanıcılara önerilir.

Veri setleri ve analizi

Bu çalışmada, bilimsel çalışma öneri sistemleri alanında içerik tabanlı yöntemler için geliştirilmiş ve araştırmacıların erişimine açık iki veri seti kullanılmaktadır.

- ARXIV veri setinde bilgisayarla görü, örüntü tanıma, makine öğrenmesi, yapay zeka, istatistiksel öğrenme, hesaplama ve dil, sinirsel ve evrimsel hesaplama, bilgiye erişim, optimizasyon ve kontrol, robotik ve yüksek enerji fiziği alanlarında oluşturulmuş 1992 ve 2018 yılları arasında yayınlanan makaleler ile ilgili veri bulunmaktadır [12]. Bu veri setine ait özellikler açıklamaları ile birlikte Tablo 1'de verilmektedir.
- NIPS veri seti, 1987 ve 2016 yılları arasında düzenlenen "Sinirsel Bilgi İşleme Sistemleri Konferansı" (*Conference on Neural Information Processing Systems*) kapsamında derin öğrenme ve bilgisayarla görüden, bilişsel bilimler ve pekiştirmeli öğrenmeye kadar çeşitli alanlarda yayınlanmış bilimsel

çalışmaları içermektedir [13]. Bu veri setinde, makalelerinin başlığı, yazarları, özetleri ve tüm metin bilgileri bulunmaktadır. İlgili özellikler açıklamaları ile birlikte Tablo 2'de verilmektedir.

Tablo 1. ARXIV veri setinin özellikleri

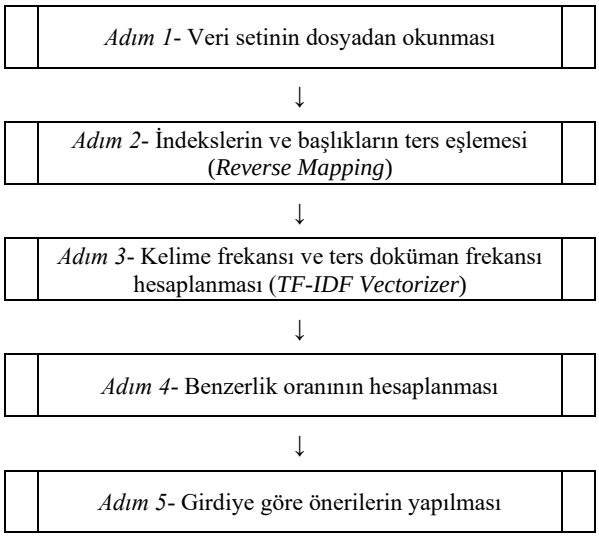
Öznitelik	Tanım	Tip
author	Yazar bilgisini içerir.	Metin
day	Basıldığı günü içerir.	Numerik
month	Basıldığı ayı içerir.	Numerik
year	Basıldığı yılı içerir	Numerik
id	Kimliğini içerir.	Metin
title	Başlığını içerir	Metin
link	Çevrimiçi bağlantısı sağlar.	Metin
summary	Özetini içerir.	Metin
tag	Kapsamı kısaltılmış etiketleri içerir.	Metin

Tablo 2. NIPS veri setinin özellikleri

Öznitelik	Tanım	Tip
id	Kimliğini içerir.	Numerik
year	Basıldığı yılı içerir.	Numerik
title	Başlığını içerir.	Metin
event_type	Konferans, poster, vb. bilgisini içerir.	Metin
pdf_name	İndirilebilir dokümanının ismidir.	Metin
abstract	Özetini içerir.	Metin
paper_text	Çalışma metnini içerir.	Metin

Önerilen yöntem

Bu makalede önerilen ve beş adımdan oluşan yöntemin akış şeması Şekil 1'de gösterilmektedir.



Şekil 1. Önerilen yönteme ait akış diyagramı

Adım 1- Veri Setinin Dosyadan Okunması: İlk adımda veri seti dosyadan okunmaktadır. ARXIV ve NIPS verisinden sırasıyla 6000 ve 2000 örneklem bu çalışmaya dahil edilmiştir. Analiz edilen veride `NULL` değer bulunmamaktadır. Veri setlerinde sırasıyla 37.810 ve 265.724 adet kelime veya kelime grubu yer almaktadır. Bu kelimelere ait karşılaşılma sıklığı istatistikleri iki veri seti için de birleştirilmiş şekilde Tablo 3'te, istatistiksel gösterim ise Şekil 2'de verilmektedir.

Tablo 3. En sık gözlenen kelime istatistikleri

	(a) ARXIV			(b) NIPS	
	kelime	**sıklık**		**kelime**	**sıklık**
1	*model*	8067	**1**	*model*	35067
2	*network*	5712	**2**	*network*	28493
3	*method*	5287	**3**	*function*	23102
4	*data*	4899	**4**	*algorithm*	22924
5	*learn*	4762	**5**	*data*	20479
6	*algorithm*	4568	**6**	*set*	19494
7	*propose*	4459	**7**	*show*	16958
8	*learning*	3995	**8**	*result*	16319
9	*show*	3828	**9**	*neural*	16207
10	*problem*	3765	**10**	*learning*	16177

(a) ARXIV (b) NIPS

Şekil 2. En sık gözlenen kelime görseli

Adım 2- İndekslerin ve Başlıkların Ters Eşlemesi (*Reverse Mapping*): Bu aşamada, bir başlık bilgisini girdi olarak alan ve en çok benzeyen `10` çalışma listesi çıkarılmaktadır. Bu doğrultuda, makale başlıkları ve veri indeksleri üzerinde ters eşleme işlemi gerçekleştirilir.

Adım 3- Kelime Frekansı ve Ters Doküman Frekansı Hesaplanması (*TF-IDF Vectorizer*): Terim Frekansı – Ters Doküman Frekansı yönteminde (TF-IDF, *Term Frequency – Inverse Document Frequency*) bir doküman içerisinde geçen terimlerin çıkarılması ve bu terimlerin geçtiği miktara göre çeşitli hesapların yapılmasına dayanmaktadır. TF-IDF, bilgi erişiminde özellik çıkarma amacıyla kullanılır ve Doğal Dil İşlemenin (*Natural Language Processing*) bir alt alanıdır. TF-IDF;

- Arama motorları sıralama ve derecelendirme sırasında kullanabilir.
- Bağlaç, noktalama işaretleri, vb. (*stopwords*) içeren terim olmayan ifadeleri filtrelemede kullanılabilir.
- Metin özetlemesinde ve sınıflandırmasında kullanılabilir.

TF-IDF ağırlığı iki terimden oluşur. İlk terim bir belgedeki bir sözcüğün görünme sayısının o belgedeki toplam sözcük sayısına bölünmesiyle elde edilen sayı olan normalleştirilmiş TF, ikinci terim ise kitaplıktaki belge sayısının logaritmasının belirli terimin göründüğü belge sayısına bölünmesiyle hesaplanan IDF terimidir.

TF, bir terimin bir belgede ne sıklıkta geçtiğini ölçer. Her dokümanın uzunluğu farklı olduğundan, bir terimin uzun belgelerde daha kısa olanlara göre çok daha fazla görünmesi mümkündür. Bu nedenle frekans terimi, normalleştirme yöntemi olarak genellikle belge uzunluğuna, diğer bir deyişle belgedeki toplam terim sayısına bölünür (1).

$$Tf(t) = \frac{Bir\ dokümanda\ t\ teriminin\ görülme\ sıklığı}{Bir\ dokümandaki\ terimlerin\ toplam\ sayısı} \quad (1)$$

IDF, bir terimin ne kadar önemli olduğunu ölçer. TF hesaplanırken, tüm terimler eşit derecede önemli kabul edilir. Bununla birlikte, "eşittir", "/" ve "bu" gibi belirli terimlerin birçok kez görünebileceği, ancak çok az önem taşıdığı bilinmektedir. Denklem (2) ile terimlerin önem oranı hesaplanır.

$$Idf(t) = log_{10}(\frac{Dokümanların\ toplam\ sayısı}{t\ terimini\ içeren\ dokümanların\ sayısı}) \quad (2)$$

Denklem (2)'den anlaşıldığı üzere bir terim ne kadar az dokümanda tekrar ediyor ise IDF değeri o kadar büyük olur.

Adım 4- Benzerlik Oranının Hesaplanması: TF-IDF oranları hesaplanmış içerikler arasındaki benzerlik oranları hesaplanır. Öneri sistemlerinde benzerlik hesabında lineer kernel, sigmoid kernel, öklid mesafesi ve pearson korelasyonu sıklıkla kullanılan yöntemlerdir. Bu yöntemlerin hangisinin daha iyi olduğu ile ilgili ortak bir fikir bulunmamakla birlikte farklı senaryolarda farklı yöntemlerin uygulanması genellikle iyi bir fikirdir.

Lineer Kernel fonksiyonu, x ve y sütun vektörleri ise, bunların doğrusal çekirdeği Denklem (3) ile hesaplanır:

$$k(x, y) = x^T y \quad (3)$$

Sigmoid Kernel fonksiyonu, iki vektör arasındaki sigmoid çekirdeğini hesaplar. Sigmoid çekirdek aynı zamanda hiperbolik tanjant veya çok katmanlı algılayıcı olarak da bilinir (çünkü sinir ağı alanında genellikle nöron aktivasyon işlevi olarak kullanılır). x, y giriş vektörlerini, γ eğim ve c_0 kesişme parametrelerini içeren *Sigmoid Kernel* fonksiyonu Denklem (4)'te tanımlanmıştır.

$$k(x, y) = \tanh(\gamma x^T y + c_0) \quad (4)$$

Öklid Mesafesi (*Euclidean Distance*), iki noktayı birleştiren düz bir çizginin uzunluğunu hesaplayarak iki nokta arasındaki mesafeyi bulmaktadır. Benzerlik oranı asla negatif olmaz

ve değer sıfıra yaklaştıkça daha benzer olduğu anlamına gelir. Öklid mesafesi Denklem (5) ile hesaplanır.

$$\mathrm{k}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (5)$$

Pearson Korelasyonu (*Pearson's Correlation*), iki değişkenin doğrusal olarak ilişkili olduğu dereceyi gösteren `-1` ile `+1` arasında bir sayıdır. Pearson korelasyonu aynı zamanda "ürün moment korelasyon katsayısı" veya basitçe "korelasyon" olarak da bilinir. `-1` ile `+1` arasında değişir ve `0`, korelasyon olmadığını gösterir. `-1` veya `+1` korelasyonları, tam bir doğrusal ilişki anlamına gelir. Pozitif korelasyonlar, *x* arttıkça *y*'nin de arttığını gösterir. Negatif korelasyonlar, *x* arttıkça *y*'nin azaldığını gösterir.

Adım 5- Girdiye Göre Önerilerin Yapılması: Bu adımda, girilen başlık bilgisine karşılık gelen rakamsal indeks hesaplanarak o indekse sahip makalenin diğer çalışmalar ile olan benzerlik değerleri hesaplanır. Benzerlik yöntemleri kullanılarak hesaplanmış bu değerler sıralanarak ilk on makalenin önerilmesi sağlanır. Sıralama işlemi öklid mesafesi için küçükten büyüğe yapılırken diğer yöntemler için büyükten küçüğe doğrudur.

Testler

Tavsiye edilen içerik tabanlı akademik makale öneri sisteminde benzerlik ölçümlerinin değerlendirilebilmesi için ARXIV ve NIPS veri setleri kullanılmaktadır. Her iki veri setinde, gereksiz kelimeler (*stopwords*) ve uzunluğu üç karakterden az kelimeler çıkarılarak veri setleri ön işlemden geçirilmektedir. Özniteliklerin elde edilmesinde farklı seviyelerde (*bigram, trigram, fourgram*) bilgi çıkarılmasına imkân veren karakter düzey *N-gram* modeli kullanılmıştır. Bu modelde, n karakter uzunluğundaki kelimeler öznitelikleri oluşturmaktadır [14]. Böylece, özniteliklerin dilden bağımsız olması, kısaltma kullanımı ve yazım yanlışı gibi durumlara karşı dayanıklı olması sağlanmaktadır. Daha sonra, tavsiye edilen sistemin farklı veri setleri üzerinde farklı benzerlik ölçümlerinin etkisi analiz edilmektedir. Bu amaçla, *lineer kernel*, *sigmoid kernel*, *öklid mesafesi* ve *pearson korelasyonu* benzerlik ölçümleri kullanılarak aynı başlık bilgisine göre ilk on makalenin önerilmesi işlemi tavsiye edilen sisteme göre gerçekleştirilmektedir. Her iki veri setinde de bulunan aynı başlık bilgisi girdi olarak kullanılarak deneysel çalışmalar gerçekleştirilmiştir. Tablo 3 ve Tablo 4'te "*Semi-supervised Learning with Ladder Networks*" başlık bilgisi işlenerek sırasıyla ARXIV ve NIPS veri setleri için öneriler listelenmektedir.

Öklid mesafesi yönteminde uzaklıklar hesaplandığı için birbirine en yakın olan makaleler en kısa mesafeye sahiptir yaklaşımı temel alınmaktadır. Bu nedenle, *öklid mesafesi* yönteminde sıralama işlemi küçükten büyüğe doğru yapılırken *sigmoid*, *lineer* ve *pearson korelasyonu* yöntemlerine göre sıralama büyükten küçüğe doğru yapılmaktadır. Tablo 3 ve Tablo 4'te *sigmoid kernel* yöntemi için γ eğim ve c_0 parametreleri için sırasıyla 0.8 ve 0.5 değerleri alınmaktadır. Bu değerler, farklı parametre setleri için gerçekleştirilen deneylerde elde edilen en uygun sonuçlara göre belirlenmiştir. Tablo 3 ve Tablo 4 ile verilen sonuçlar incelendiğinde aynı başlık bilgisi verildiğinde her iki veri setinde de benzer makalenin önerildiği görülmektedir; aynı çalışmanın geçtiği durumlar kalın yazım şekli ile gösterilmektedir.

Tablo 4. ARXIV veri setinde "Semi-supervised Learning with Ladder Networks" başlıklı bilimsel çalışma üzerinden ilk on öneri

	Lineer Kernel	Sigmoid Kernel	Öklid Mesafesi	Pearson Korel.	Başlık
1	0.409	0.679	1.087	0.409	*Virtual Adversarial Ladder Networks For Semi-supervised Learning*
2	0.313	0.636	1.172	0.313	***Recurrent Ladder Networks***
3	0.312	0.635	1.173	0.312	*Adversarial Ladder Networks*
4	0.306	0.632	1.178	0.306	*Semi-Supervised Learning with Deep Generative Models*
5	0.300	0.629	1.183	0.300	*Video Ladder Networks*
6	0.281	0.620	1.199	0.281	*Semi-Supervised Phoneme Recognition with Recur...*
7	0.274	0.616	1.205	0.273	*Deep Bayesian Active Semi-Supervised Learning*
8	0.264	0.611	1.213	0.264	*Semi-Supervised Phoneme Recognition with Recurrent Ladder Networks*
9	0.254	0.607	1.221	0.254	*Semi-supervised Learning with Density Based Distances*
10	0.138	0.544	1.313	0.252	*Supervised Learning with Growing Cell Structures*

Tablo 5. NIPS veri setinde "Semi-supervised Learning with Ladder Networks" başlıklı bilimsel çalışma üzerinden ilk on öneri

	Lineer Kernel	Sigmoid Kernel	Öklid Mesafesi	Pearson Korel.	Başlık
1	0.330	0.643	1.158	0.329	***Recurrent Ladder Networks***
2	0.303	0.631	1.180	0.303	*Iterative Double Clustering for Unsupervised and Semi-Supervised Learning*
3	0.296	0.627	1.187	0.295	*Good Semi-supervised Learning That Requires a Bad GAN*
4	0.284	0.621	1.197	0.284	*Semi-supervised MarginBoost*
5	0.277	0.618	1.202	0.277	*Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference*
6	0.226	0.592	1.245	0.225	*Semi-Supervised Support Vector Machines*
7	0.167	0.560	1.291	0.166	*Learning Disentangled Representations with Semi-Supervised Deep Generative Models*
8	0.145	0.548	1.308	0.144	*Using Unlabeled Data for Supervised Learning*
9	0.139	0.545	1.313	0.138	*Supervised learning from incomplete data via an EM approach*
10	0.138	0.544	1.313	0.137	*Supervised Learning with Growing Cell Structures*

Diğer yandan, aynı veri seti içerisinde farklı benzerlik ölçüm yöntemleri ile aynı ilk on makalenin önerildiği görülmektedir. Benzerlik ölçümlerini, hesaplama süresi açısından karşılaştırabilmek için Tablo 6 oluşturulmuştur.

Tablo 6. Farklı benzerlik ölçüm yöntemleri için elde edilen hesaplama süreleri (s)

Yöntem	ARXIV	NIPS
Sigmoid Kernel	0.22	0.94
Lineer Kernel	0.21	0.60
Öklid Mesafesi	0.21	0.99
Pearson Korel.	0.76	4.55

Tablo 6 incelendiğinde, aynı veri seti için en yüksek hesaplama maliyetine sahip yöntemin *pearson korelasyonu* yöntemi olduğu görülmektedir. Ayrıca, Tablo 3 ve Tablo 4'te *lineer kernel* ile *pearson korelasyonu* yöntemi ile elde edilen değerlerin birbirine oldukça yakın olduğu görülmektedir. Bu durumda, bu iki yöntem arasında *lineer kernel* yönteminin seçiminin daha uygun olacağı sonucuna varılmaktadır. Diğer yandan; NIPS veri seti ile incelenen kelime sayısı daha fazla olduğu için hesaplama süresi açısından ARXIV veri seti ile kıyaslandığında daha yüksek değerlerin elde edildiği görülmektedir.

Sonuçlar

Öneri sistemleri alanında literatürde çok sayıda çalışma bulunmaktadır. Gerçek hayattaki farklı problemler için geliştirilmiş birçok öneri yöntemi olmasına karşın, hala kapsamlı bir şekilde araştırılmamış bazı alanlar vardır. Bilimsel çalışma öneri sistemleri hala araştırma aşamasında olan konular arasında gösterilebilir. Bu alanda önerilen yöntemler ağırlıklı olarak içeriğe dayalı yaklaşımı dikkate almaktadır. Bu çalışmada, yeni bir içerik tabanlı bilimsel çalışma öneri sistemi tavsiye edilmektedir. Tavsiye edilen sistem, farklı veri setleri için farklı benzerlik ölçümleri kullanılarak analiz edilmektedir. Analiz sonuçlarında aynı başlık bilgisi için aynı veri seti içerisinde aynı ilk on makalenin önerildiği ve bu makaleler için elde edilen benzerlik değerlerinin birbirlerine oldukça yakın olduğu görülmektedir. Birbirine oldukça yakın değerlerin elde edildiği benzerlik ölçümleri için ise hesaplama sürelerinin dikkate alınması yöntem seçiminde oldukça yardımcı olmaktadır.

Gelecekte yapılacak çalışmalar arasında, benzerlik yöntemlerinin ayırt ediciliğini arttırmak adına içerik tabanlı yöntemlerin, işbirlikçi filtreme gibi yöntemler ile bir araya getirilerek karma yöntemlerin oluşturulması verilebilir. Diğer yandan, başlık bilgisi yanında başlık, anahtar kelimeler ve/veya özet bilgilerinin kombinasyonları ile tavsiye edilen sistem analiz edilerek kullanıcı memnuniyetinin de dikkate alındığı öneriler gerçekleştirilebilir.

Teşekkür

Yazarlar finansal destek için TÜBİTAK'a (proje numarası 109M637) teşekkür ederler.

Kaynaklar

[1] Lu, J., D. Wu, M. Mao, W. Wang, and G. Zhang, *Recommender system application developments: a survey*. Decision Support Systems, 2015. **74**: p. 12-32.

[2] Melville, P. and V. Sindhwani, *Recommender systems.* Encyclopedia of machine learning, 2010. **1**: p. 829-838.

[3] Portugal, I., P. Alencar, and D. Cowan, *The use of machine learning algorithms in recommender systems: A systematic review*. Expert Systems with Applications, 2018. **97**: p. 205-227.

[4] Beel, J., S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger. *Research paper recommender system evaluation: A Quantitative Literature Survey*. 2013. ACM Press.

[5] Beel, J., B. Gipp, S. Langer, and C. Breitinger, *Research-paper recommender systems: a literature survey*. International Journal on Digital Libraries, 2016. **17**(4): p. 305-338.

[6] Wang, D.H., Y.C. Liang, D. Xu, X.Y. Feng, and R.C. Guan, *A content-based recommender system for computer science publications.* Knowledge-Based Systems, 2018. **157**: p. 1-9.

[7] Maleszka, B., *A Framework for Research Publication Recommendation System*. 2019, Springer International Publishing. p. 167-178.

[8] Dhanda, M. and V. Verma, *Recommender system for academic literature with incremental dataset.* Procedia Computer Science, 2016. **89**: p. 483-491.

[9] Sugiyama, K. and M.-Y. Kan. *Scholarly paper recommendation via user's recent research interests*. in *Proceedings of the 10th annual joint conference on Digital libraries*. 2010.

[10] Bai, X.M., M.Y. Wang, I. Lee, Z. Yang, X.J. Kong, and F. Xia, *Scientific Paper Recommendation: A Survey*. Ieee Access, 2019. **7**: p. 9324-9339.

[11] Lops, P., M. De Gemmis, and G. Semeraro, *Content-based recommender systems: State of the art and trends*, in *Recommender systems handbook*. 2011, Springer. p. 73-105.

[12] *ARXIV data from 24,000+ papers*. URL:https://www.kaggle.com/neelshah18/arxivdataset. (Erişim Zamanı; 19/11/2020);

[13] *NIPS Papers*. URL:https://www.kaggle.com/benhamner/nips-papers/notebooks. (Erişim Zamanı; 19/11/2020);

[14] Kanaris, K., I. Houvardas, and E. Stamatatos. *Words vs. Character n-grams for anti-spam filtering*. 2006.