Türk Kütüphaneciliği, 36, 2 (2022), 169-203
Doi: 10.24146/tk.1062751

İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi: Pennant Erişimle Desteklenen Yeni Bir Yöntem Önerisi*

Incremental Refinement of Relevance Rankings: Introducing a New Method Supported with Pennant Retrieval

Müge Akbulut** ve **Yaşar Tonta*****

Öz

Amaç: *İlgi sıralaması algoritmaları erişilen belgeleri arama sorgularıyla belgeler arasındaki konusal benzerlik (ilgi) derecelerine göre sıralar. Bu çalışmanın amacı; bir olasılıksal konu modelleme algoritması ile atıf verilerinin birleşiminden oluşan yeni bir ilgi sıralaması yöntemi geliştirmektir.*

Veri Kaynakları ve Yöntem: *Geliştirdiğimiz yöntemi yaklaşık 435 bin fizik makalesinden oluşan iSearch derlemi üzerinde uyguladık. Önce 65 sorgu için derlemdeki tüm makalelerin başlıkları ve özetleri üzerinde konu modelleme algoritmasını çalıştırarak ilgi sıralamalarını elde ettik. Daha sonra atıf bilgilerini mevcut ilgi sıralamalarını tümleştirmek ve daha da geliştirmek için kullandık. Böylece hem aranan konunun farklı yönlerini kapsayan hem de konuyla marjinal ilgili olan makalelerden oluşan daha iyi ilgi sıralamaları elde ettik. Önerdiğimiz yöntemin erişim performansını değerlendirdik.*

Bulgular: *Bulgular konu modelleme algoritması ile elde edilen ilgi sıralamalarında makalelerin başlıklarında ve özetlerinde geçen bazı terimlerin bazen göz ardı edilebildiğini göstermektedir. Ama bu sıralamalar atıf verileriyle desteklendiğinde, kullanılan terimlerin bağlamları hakkında ek bilgiler elde edilmekte ve sonuçta ilgi düzeyleri daha yüksek ve çeşitli makaleler içeren daha zenginleştirilmiş ilgi sıralamaları oluşturulmaktadır. Dahası, erişim çıktıları araştırmacıların önceliklerine göre kolayca yeniden sıralanabilmektedir.*

Sonuç: *Önerdiğimiz yöntemde pennant erişim tekniklerini kullanarak mevcut ilgi sıralaması algoritmalarının artırımlı olarak iyileştirilmesi üzerinde odaklandık. Bu yöntemin hesaplama yükü, sağlamlık, tekrarlanabilirlik ve ölçeklenebilirlik açılarından dinamik derlemler üzerinde sınandıktan sonra zamanla TR-Dizin, Web of Science ve Scopus gibi hem yerel hem de uluslararası bilgi sistemlerinde de kullanılabileceği kanısındayız.*

---

* Bu araştırma Müge Akbulut'un Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümünde yaptığı doktora tezine dayanmaktadır (http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/26338).

** Ankara Yıldırım Beyazıt Üniversitesi, Bilgi ve Belge Yönetimi Bölümü, Ankara, Türkiye. E-posta: mugeakbulut@gmail.com
Ankara Yıldırım Beyazıt University, Department of Information Management, Ankara, Turkey. E-mail: mugeakbulut@gmail.com
*** Hacettepe Üniversitesi, Bilgi ve Belge Yönetimi Bölümü, Ankara, Türkiye. E-posta: yasartonta@gmail.com
Hacettepe University, Department of Information Management, Ankara, Turkey. E-mail: yasartonta@gmail.com

Gönderim Tarihi –*Received:* 25.01.2022
Kabul Tarihi – *Accepted:* 10.04.2022
Yayımlanma Tarihi–*Published:*30.06.2022

Özgünlük: *Bu araştırmada yeni bir ilgi sıralaması yöntemi önerilmektedir. Bildiğimiz kadarıyla bu çalışma, LDA konu modelleme algoritması ile elde edilen ilgi sıralamalarının atıf verileriyle artırımlı olarak geliştirilebileceğini gösteren ilk çalışmadır.*
***Anahtar Sözcükler*:** *İlgi sıralamaları; olasılıksal konu modellemesi; Gizli Dirichlet Ayırımı (LDA) algoritması; pennant erişim; Maksimum Marjinal İlgi (MMR).*

Abstract

Purpose: *Relevance ranking algorithms rank retrieved documents based on the degrees of topical similarity (relevance) between search queries and documents. This paper aims to introduce a new relevance ranking method combining a probabilistic topic modeling algorithm with citation data.*

Data and Method: *We applied this method to the iSearch corpus of c. 435,000 physics papers. We first ran the topic modeling algorithm on titles and summaries of all papers for 65 search queries and obtained the relevance ranking lists. We then used citation data with the existing relevance rankings, thereby incrementally refining the results. The outcome produced better relevance rankings with papers covering various aspects of the topic searched as well as the more marginal ones. Finally, we evaluated the retrieval performance of the proposed method.*

Findings: *Findings suggest that the topic modeling algorithm might sometimes overlook the terms used in different contexts in the papers. However, the fusion of citation data to relevance ranking lists provides additional contextual information, thereby enriching the results further with various papers of higher relevance. Moreover, results can easily be re-ranked.*

Implications: *We argue that once it is tested on dynamic corpora for computational load, robustness, replicability, and scalability, the proposed method can, in time, be used in both local and international information systems such as TR-Dizin, Web of Science, and Scopus.*

Originality: *The proposed method is, as far as we know, the first one that shows that relevance rankings produced with a topic modeling algorithm can be incrementally refined using citation data.*

Keywords: *Relevance rankings; probabilistic topic modeling; the Latent Dirichlet Allocation (LDA) algorithm; pennant retrieval; Maximal Marginal Relevance (MMR).*

Giriş

Bilgi erişim kullanıcının bilgi ihtiyacını tanımladığı sorgu terimleri ile belgelerde geçen terimlerin eşleştirilmesine dayanır. Fakat bu süreçte belge ve sorgu temsilinde aynı kavramın farklı biçimlerde temsil edilebilme olasılığından kaynaklanan bazı belirsizlikler söz konusudur (Ganguly ve Jones, 2018). Temsil için belirlenen terimler öznel olduğu için kişiye, zamana ve duruma göre değişebilir (Swanson, 1986a). Bilgi erişimin mantıksal organizasyonundan kaynaklanan bu erişim problemlerine karşın erişim çıktısındaki belgeler kullanıcının mevcut bilgisi ve tercihleri ile uyumluysa ve işleme çabasına (processing effort) değmişse "ilgili" (relevant) olarak değerlendirilmektedir (Saracevic, 2021; Wilson ve Sperber, 2002). Belgenin konusu, ilgisi, bilgi ihtiyacının ne olduğu gibi hususlar öznel olduğu için sistemdeki tüm ilgili ve sadece ilgili belgelere erişim sağlayacak ideal bilgi sistemi tasarlamak genellikle mümkün değildir (Mizzaro, 1997; Wilson, 1978). Bu noktada ilgi sıralamaları (relevance rankings)

kullanıcı tatmini açısından önemli rol oynamaktadır (Lei ve diğerleri, 2001). Çünkü kullanıcılar genellikle birkaç ilgili belgeye fazla çaba harcamadan eriştiklerinde tatmin olmaktadırlar (Tonta, 1995, s. 302).

İyi bir bilgi erişim sisteminin kullanıcının sorgusuna göre koleksiyondaki hangi belgelerin daha ilgili olduğunu öngörmesi ve bu belgeleri olasılık sıralama ilkesine (probability ranking principle) göre sıralaması beklenir (Robertson, 1977). Fakat bazen sıralamada birbirine çok benzeyen kaynaklar yerine (ya da onlara ek olarak) sorgulanan konunun çeşitli yönlerini ele alan kaynaklara ihtiyaç duyulur. Bu bakımdan özellikle literatür taraması gibi konunun tüm yönlerinin araştırıldığı sorgular için erişilen kaynakların çeşitliliği de önemlidir (Kucuktunc ve Ferhatosmanoglu, 2011, s. 481).

İlgi sıralamalarının oluşturulmasında kullanılan yöntemlerden birisi de konu modellemesidir (topic modeling). Konu modelleme algoritmaları herhangi bir terim için benzer ya da eş anlamlı terimlerin de geçtiği belgeleri listeler. Örneğin, olasılıksal konu modelleme yaklaşımlarından birisi olan LDA (Latent Dirichlet Allocation – Gizli Dirichlet Ayırımı) algoritması bilgi erişim sistemlerinde sorgu-belge, konu-belge, konu-sorgu ve belge-belge benzerliklerinin hesaplanmasına ve dolayısıyla ilgi sıralamaları oluşturulmasına olanak sağlamaktadır (Blei ve diğerleri, 2003; Li ve McCallum, 2006).

İlgi sıralaması oluşturma problemi çoğunlukla belge ve sorgu arasında eşleşme (token matching) problemine indirgenmektedir. Nitekim LDA algoritmasında da bu problem belge-belge ve sorgu-belge benzerliği olarak tanımlanmaktadır. Oysaki ilgi sıralamaları açısından sorgularda ve belgelerde geçen terimler arasındaki anlamsal (semantic) benzerlikler de önemlidir. Başarılı bir ilgi sıralaması oluşturmak için sorguyla belgeler arasında eşleşme olması gerektiği gibi sorgu teriminin önemi, belgelerin tematik bağlamları ve bu bağlamlar kullanılarak ilgi olasılıklarının tahmin edilmesi de gerekmektedir (Guo ve diğerleri, 2016; Wu ve diğerleri, 2007).

Atıf dizinlerinde ise belgelerin bağlamı ve ilgisi hakkında bibliyometrik bilgiler mevcuttur (Carevic ve Schaer, 2014). Örneğin, atıf yapan ve atıf yapılan yayın arasında bir anlamsal ilişki olabileceği fikrinden hareketle geliştirilen ortak atıf analizi ile konusal ilgi örüntüleri ortaya çıkarılabilmektedir (Han, 2020; Knoth ve diğerleri, 2017; Küçüktunç ve diğerleri, 2015; White, 2010). Bu sayede, atıflardan iz sürerek sisteme sunulan bir bilimsel yayına (çekirdek makale) benzer diğer yayınlar saptanabilmektedir. Bu tarz bibliyometri destekli (bibliometric-enhanced, bibliometrics-aided) uygulamalarda bilgi erişim performansı ciddi düzeyde artmaktadır (Mayr ve Mutschke, 2013). Benzeri bir biçimde, farklı alanlar arasındaki dolaylı ama önemli bağlantıların ortaya çıkarılması için bibliyometrik veriler kullanılarak literatür tabanlı keşif (literature-based discovery) (Swanson, 1986b) ve örüntü tabanlı ilişki çıkarma (pattern-based relationship extraction) çalışmaları gerçekleştirilmektedir (Yang ve diğerleri, 2017).

Kelime tabanlı konu modelleme yaklaşımı (örneğin, LDA) ve atıf tabanlı yaklaşım ayrı ayrı değerlendirildiğinde ikisinin de bazı eksik yanlarının olduğu bilinmektedir. Örneğin, kelime tabanlı yaklaşımlar, farklı alanlardaki özdeş kavramların değişik kullanımlarının neden olduğu karışıklıktan etkilenmektedir (bazen "yapay öğrenme" ile "makine öğrenmesi" eş anlamlı olarak kullanılmaktadır). Öte yandan, iki farklı kavram farklı alanlarda aynı adla anılabilir (Zarrinkalam ve Kahani, 2012). Bu durum ilgili yayınların göz ardı edilmesine ya da

listede ilgisiz yayınların yer almasına yol açabilir (Küçüktunç ve diğerleri, 2015, s. 2). Atıf tabanlı yaklaşımlar tematik bağlam yakalamada başarılı olsa da, bir çalışmanın atıf alması için belli bir zaman geçmesi gerekmektedir (Ke ve diğerleri, 2015). Atıf tabanlı yaklaşımlarda genellikle bir çekirdek makaleye ihtiyaç vardır. LDA algoritmasının atıflarla desteklendiği uygulamalarda LDA'nın performansı artmakta, önemli ve etkili çalışmalara erişim sağlanmaktadır (Guo ve diğerleri, 2013; Huang ve diğerleri, 2016; Huang ve diğerleri, 2018; Li ve diğerleri, 2017; Nguyen ve Do, 2018; Wang ve diğerleri, 2013; Xia ve diğerleri, 2012; Zhou ve diğerleri, 2017; Zou ve diğerleri, 2021).

Konularla atıflar arasındaki ilişki genel olarak kabul edilenden daha belirsizdir (Ballester ve Penner, 2022). Ama farklı erişim algoritmalarının farklı ilgili belgelere eriştikleri de bilinmektedir (Croft, 2002). Konu modellemesi ve atıflara dayanan algoritmaların ilgi sıralamalarının iyileştirilmesi için birlikte kullanıldığı bir çalışma bildiğimiz kadarıyla literatürde henüz bulunmamaktadır.

Bu çalışmanın temel amacı, söz konusu iki yaklaşımı bir arada kullanarak ilgi ve çeşitlilik (interdisiplinerlik) oranı yüksek, sorgudaki terimlerin ya da yöntemin farklı uygulamalarının gözlenebildiği, marjinal[1] ve sorguyla ilgili kaynakları içeren ilgi sıralamaları oluşturmaktır. Çalışmanın temel araştırma sorusu ise "LDA konu modelleme algoritması uygulanarak elde edilen ilgi sıralamaları ilgi kuramı, bilgi erişim ve bibliyometriye dayanarak geliştirilen ve atıf verilerini kullanan pennant erişim yöntemiyle desteklenerek ilgi ve çeşitlilik oranları artırılmış ilgi sıralamaları geliştirilebilir mi?" şeklinde formüle edilmiştir (Akbulut ve diğerleri, 2020; White, 2007a, 2007b, 2009). Bu amaçla 2009 yılına kadar arXiv'e eklenen tüm fizik konulu makaleleri içeren iSearch derlemi üzerinde bir uygulama yapılmıştır. Bulgular, önerilen yöntemle çeşitlilik ve ilgi oranları daha yüksek ve kullanıcıların ihtiyaçlarına göre kişiselleştirilebilen ilgi sıralamaları elde edilebileceğini göstermektedir.

Literatür Değerlendirmesi

Yayın sayılarının hızla artması araştırmacıların ilgili kaynaklara erişmelerini giderek zorlaştırmaktadır (Bornmann ve diğerleri, 2021). 1950'lerin başından beri ilgi ve dolayısıyla ilgi sıralamaları bilgi erişim sistemlerinin tasarımı, optimizasyonu ve değerlendirilmesinde odak noktası olmuştur (Saracevic, 2021; Verma ve diğerleri, 2016). İlgi sıralamaları özelinde konusal ilginin (topical relevance) temeli kullanıcı sorgusu ve dizin terimleri arasındaki tam çakışma ya da benzerlik oranına dayanmaktadır (Carevic ve Schaer, 2014; White, 2007b). Diğer bir deyişle, kullanıcılar için en az çaba gerektiren ve en kolay çıkarımlar terim eşleşmesine dayananlardır. Oysaki ilgi sıralamalarında yenilik, popülerlik, çeşitlilik gibi farklı özellikler de önemlidir.

Sorgu terimleri ve erişim çıktılarında tam eşleşme durumu olmadığında belgeleri işlemek daha zordur. Çünkü çakışma oranları yüksek olmasa bile sorgu ve belgeler arasında konusal ilgi olabilir (Akbulut, 2016, s. 9). Bunun için belgelerin tam metinleri üzerinde doğal dil işleme yöntemleri kullanılarak ilgi belirlenmektedir (örneğin, Chen ve Décary, 2018; Cambria ve White, 2014). Olasılıksal konu modelleme yöntemi ile belgelerin konularını ve bu konuların hangi oranda hangi terimleri potansiyel olarak içerebileceğini ortaya çıkarmak ve

[1] "Marjinal" kelimesi "aykırı", "sınırda", "uçta", "sıra dışı" anlamına gelmektedir. Bu kelime bu araştırmada konuyla ya da sorguyla ilgili olan ama, örneğin, anahtar kelime eşleşmesi yoluyla kolayca erişilemeyen kaynaklar anlamında kullanılmaktadır.

gizli tematik bilgileri belirlemek mümkündür (Boyd-Graber ve Blei, 2010). Konu modellemede belgeler, her konunun kelimelerin dağılımına göre karakterize edildiği gizli (latent) konular üzerine rastgele karışımlar olarak temsil edilir (Blei ve diğerleri, 2003, s. 996). Bu bağlamda konu modelleme amacıyla kullanılan en popüler algoritmalardan biri LDA'dır. Bu algoritmada, ilgiyi belirleyebilmek için derlemde yer alan belgeler hem belli bir belgede geçen terimlerin hem de farklı belgelerde geçen terimlerin birlikte geçiş sıklıkları açısından incelenir. Böylece her belgenin bir veya birden fazla konuya ait olabileceği sonucunu veren model oluşturulur ve her belge için saptanan konuların olasılık dağılımı bulunur (Blei ve diğerleri, 2003; Chang ve diğerleri, 2009).

Ancak LDA algoritmasının bazı dezavantajları bulunmaktadır. LDA ile tutarlı konular oluşturmak ve güvenilir istatistikler sağlamak için büyük miktarda veriye ihtiyaç duyulmaktadır (Chen ve Liu, 2014, s. 1116; Leydesdorff ve Nerghes, 2017; Xie ve diğerleri, 2019; Nguyen ve Do, 2018). Diğer yandan büyük derlemlerde konu sayısı artmakta ve tutarlılık sorunları oluşmaktadır (Hecking ve Leydesdorff, 2018). Bunun dışında terim düzeyinde hesaplama söz konusu olduğu için büyük derlem, çoklu dil, tam metin gibi durumlarda matris boyutu ve dolayısıyla hesaplama süresi ciddi oranda artmaktadır. Ayrıca LDA algoritması kelime torbası (bag of words) yaklaşımına dayalı olduğu için kelimelerin sadece belge içerisindeki konumları dikkate alınmaktadır (Chang ve diğerleri, 2009; Ekinci ve İlhan Omurca, 2020). Dolayısıyla modelde terimlerle ilgili anlamsal bilgi ya da bağlam bilgisi yer almamaktadır.

Öte yandan kelime tabanlı yöntemler ile ortaya çıkarılamayan belgeler arasındaki anlamsal ilişkiler kaynakça benzerliği ya da atıflar yoluyla açığa çıkarılabilir. Araştırmacılar atıf yaparak hem söz konusu çalışmalarının entellektüel ve bilişsel katkısını kayıt altına almış hem de yazarlarına kredi vermiş olurlar (Tonta ve Akbulut, 2021, s. 391). Bu süreçte farklı alanlardaki araştırmalar ile kurulan bağlantılar makalelerin kavramsal ve anlamsal içerikleri ve bağlam ile ilgili ipuçları barındırmaktadır. Atıf bilgilerinin algoritmalara dâhil edildiği durumlarda bilgi erişim performansı önemli ölçüde (%25) artmaktadır (Pao, 1993, s. 104). Atıf bilgileri hem erişim sonuçlarını sıralamak için hem de öneri sistemlerinde (recommendation systems) kullanılmaktadır (Beel ve Gipp, 2009; Beel ve diğerleri, 2016).

Arama yapılan makaleye benzer makaleler araştırmacılara sunulurken çoğunlukla makalelerin kaynakça benzerliğinden faydalanılmaktadır (Kessler, 1963; Carevic ve Mayr, 2014; Vergoulis ve diğerleri, 2019). Atıf dizinleri bağlamında, temel düzeyde de olsa atıf bilgileri ilgi sıralaması oluşturmak amacıyla kullanılmaktadır (Belter, 2017). Örneğin, Web of Science'ın (WoS) ilgili kayıtlar (related records) özelliği makalelerin kaynakçaları arasındaki örtüşmeye (bibliographic coupling) dayanmaktadır. İlgili kayıtlar sıralanırken, kaynakçası en çok örtüşen çalışmadan en az örtüşene doğru listelenmektedir. Bunun yanı sıra ortak atıflar da ilgi sıralamalarında kullanılmaktadır (Beel ve Gipp, 2009; Zarrinkalam ve Kahani, 2012). İki farklı makale arasındaki konusal ve anlamsal benzerliğin bir diğer göstergesi de her iki makalenin kaynakçalarında aynı kaynaklara ya da yazarlara ortak atıf yapılmasıdır (White ve McCain, 1998). Bazen kaynakça benzerliğiyle ortak atıflar birlikte kullanılmaktadır (Bichteler ve Eaton, 1980).

İster kelime tabanlı yaklaşımlar isterse kaynakça ve ortak atıf verileri kullanılsın, öznel bir kavram olan "ilgi"nin ölçülmesi zordur. Sperber ve Wilson'ın (1995) ilgi teorisine (relevance theory) göre bir girdinin ilgisini belirleyen şey o girdinin "bilişsel etki"si (cognitive effect) ile o girdiyi işlemek için gereken işleme kolaylığının (ease of processing) birbirine oranıdır.

$$ilgi = bilişsel\ etki/erişim\ kolaylığı\ (işleme\ çabası) \quad (1)$$

İlgi, bir belgenin ilgi düzeyini ve bağlamını saptamak için gereken çaba (effort) ile doğrudan ilişkilidir. İlgiyi bilişsel etki ve erişim kolaylığı ile ilişkilendiren ilk kavramsal ve ampirik çalışmalar Howard D. White tarafından gerçekleştirilmiştir (White, 2007a, 2007b, 2009, 2010, 2015, 2016, 2018). White'ın "pennant erişim" olarak adlandırdığı bu yaklaşımın temeli ilgi teorisine, Salton'un vektör uzayı bilgi erişim modeline (Salton ve diğerleri, 1975) ve bibliyometriye dayanmaktadır. White vektör uzayı modelindeki *tf***idf* (terim sıklığı * ters belge sıklığı) formülünü yeniden tanımlamış, bilişsel etkiyi (konusal ilgi) ve erişim kolaylığını (bilgiyi elde etme kolaylığı) hesaplamak için sırasıyla belgelerin ortak atıf (*tf*) ve toplam atıf (*idf*) sayılarından yararlanmıştır. Böylece *tf***idf* formülünün farklı bir biçimde yorumlandığı pennant erişim yöntemi ile erişim kolaylığı (çaba) bilgisi de kullanılarak ilgi sıralamaları elde edilebilmektedir (Akbulut ve diğerleri, 2020).[2]

Pennant erişim yöntemi arama sorgusu olarak kullanılan çekirdek makalenin (seed article) önceden yayımlanan çalışmalar ile ilişkilerini ortaya çıkarmakta ve bu çalışmanın hangi modellerin ya da yapıların oluşmasında etkili olduğunu gözlemeye olanak sağlamaktadır. Ayrıca, araştırmacıların bir konu hakkında ilgili literatürü, belli bir kavramın veya yöntemin ortaya çıkışı ve gelişimiyle birlikte daha kolay takip edebilmelerine yardımcı olmaktadır. Örneğin, bilgi erişim literatüründe atıf klasiği haline gelmiş olan ve kaynakçasında sadece iki kaynak listelenen Maron ve Kuhns'un (1960) olasılıksal bilgi erişim ile ilgili çalışması için pennant erişim yöntemi ile ilgi sıralaması oluşturulduğunda, bu yöntemin kaynakça benzerliğine dayalı ilgili kayıtlar özelliğinden çok daha iyi bir performans gösterdiği saptanmıştır (Akbulut ve diğerleri, 2020). WoS'un ilgili kayıtlar özelliği ile oluşturulan ilgi sıralamasındaki makalelerin çoğu çekirdek makale ile ilgili değilken, pennant erişim yöntemi ile erişilen makalelerin tümünün ilgili olduğu ortaya çıkmıştır. Fakat her ne kadar pennant erişim yöntemi ile oluşturulan ilgi sıralamaları için gerekli veriler (toplam atıf ve ortak atıf sayıları) atıf dizinlerinde yer alsa da, atıf dizinlerinde pennant erişim yöntemi pratikte henüz hayata geçirilmemiştir.

İlk zamanlarda bilgi erişim performans değerlendirme çalışmalarının çoğu ilgi düzeyini belirlemeye odaklanmıştır. Fakat ilgi tek başına yeterli bir performans göstergesi değildir (Bradley ve Smyth, 2001; Herlocker ve diğerleri, 2004; McNee ve diğerleri, 2006). Sadece ilgiye odaklanıldığında konunun farklı bağlamlarını yakalamaya yarayan yenilik, çeşitlilik gibi özellikler genellikle göz ardı edilmektedir (Adomavicius ve Kwon, 2011). Örneğin, yüzeysel olarak farklı görünen ancak temelinde benzer özellikler gösteren birbirine yakın alanlardaki interdisiplinerlik derecesi (degree of interdisciplinarity) yüksek çeşitli kaynakları listeleyen ilgi sıralamaları kullanıcılar için daha faydalı olabilmektedir (Abramo ve diğerleri, 2018; Akbulut, 2016; Rafols ve diğerleri, 2012). İlgi, azalan ilgi düzeyine göre sıralanmış bir liste oluşturmayı amaçlarken, erişim çıktısının çeşitlendirilmesi geniş bir konu yelpazesini kapsayan sıralanmış bir yayın listesi oluşturmaya odaklanmaktadır (Li ve diğerleri, 2020).

Yeniden sıralama (re-ranking) algoritmaları ilginin yanında çeşitlilik, popülerlik gibi özellikleri de algoritmalara dâhil ederek erişim çıktısının sorguyla ilgisini daha da artırmakta, kullanıcı ihtiyaçları açısından daha dengeli ilgi sıralamaları oluşturulmasını sağlamaktadır (Liu ve diğerleri, 2022). Bu amaçla birden fazla algoritmadan elde edilen sıralamalar veri

[2] Bilgi erişimde erişim kolaylığını dikkate alarak gerçekleştirilen diğer çalışmalar için bkz. Yılmaz ve diğerleri (2014) ve Verma ve diğerleri (2016).

tümleştirme (data fusion) yoluyla birleştirilmekte ve sürece dâhil olan tüm algoritmalardan daha iyi performans gösteren ilgi sıralamaları elde edilmektedir (Baeza-Yates ve Ribeiro-Neto, 1999; Meng ve diğerleri, 2002). Birleştirme aşamasında artırımlı (incremental, boosting) hesaplamalar kullanılması ise yüksek hesaplama maliyetini önlemektedir (Jin ve diğerleri 2008; Ma ve diğerleri, 2022).

Bu çalışmada, makalelerin başlık ve özetleri üzerinde LDA algoritması uygulanarak oluşturulan erişim çıktıları pennant erişim ile desteklenerek ilgi ve çeşitlilik oranları nispeten daha yüksek ilgi sıralamaları elde edilmiştir. İlgi sıralamalarında bağlam ve etki bilgisi atıflardan, ilgi bilgisi de hem atıf hem de kelime sıklıklarından elde edilmiştir. Önerilen yöntem ile LDA algoritması belgelerin başlık ve özetlerine uygulandığı için hem konu modellemesi daha hızlı yapılmakta hem de bağlam bilgisi ve belli bir alandaki temel çalışmaları öne çıkaran ve kullanıcıların ihtiyaçlarına göre kişiselleştirilebilen sıralama listeleri oluşturulmaktadır.

Veri Kaynakları ve Yöntem

Çalışmanın bu kısmında araştırmada kullanılan veri kaynakları, yöntem ve teknikler ayrıntılı olarak açıklanmaktadır.

Veri Kaynakları

Bu araştırmada Lykke ve diğerleri (2010) tarafından oluşturulan iSearch derlemi kullanılmıştır. Söz konusu derlem 2009 yılına kadar arXiv'e eklenen 434.813 fizik makalesinden ve bu makalelere ait 3,7 milyondan fazla dâhili referanstan oluşmaktadır. Derlemde 65 sorgu[3] ve her bir sorgu için uzmanlar tarafından derecelendirilmiş ortalama 200 ilgi değerlendirmesi bulunmaktadır. Neredeyse tüm fizik makaleleri bilimsel dergilerde yayımlanmadan önce arXiv'e yüklendiğinden, arXiv'in böyle bir araştırma için iyi bir koleksiyon olduğu düşünülmektedir. Fakat iSearch derleminde makalelerin yayın yılı bilgileri ile özetleri bulunmadığından bu makalelere dair temel konu başlıkları da dâhil tüm üst veriler arXiv API (https://arxiv.org/help/api) aracılığıyla arXiv.org'dan indirilmiştir.[4] Daha sonra çeşitli makrolar yazılarak makalelerde atıf yapılan kaynakların her birinin temel konu kategorileri belirlenmiştir.[5] iSearch derlemindeki yayınların %98'inin temel konusu fiziktir.[6] Yayınların %2'sinin temel konuları ise bilgisayar bilimi, matematik, kantitatif biyoloji, kantitatif finans ve istatistiktir. Ama bu yayınların ikincil konuları fizik olarak tanımlandığı için onlar da arXiv derleminde yer almaktadır.

Yayınların fiziğin alt konularına dağılımı homojen değildir. Örneğin, astrofizik alt konusu diğerlerine göre daha fazla makale içermektedir (tüm makalelerin %22'si). Astrofizik alanındaki gelişmeler nedeniyle alt sınıflama yapılması gereksinimi doğmuş ve 2008 yılında altı alt kategori tanımlanmıştır. Bu tarihe kadar olan makaleler ise doğrudan astrofizik alt konusu altında sınıflandırılmıştır. Miras (legacy) sınıflama sistemlerinde karşılaşılan bu sorun arXiv ve dolayısıyla iSearch derlemi için de geçerlidir. Öte yandan bazı konu sınıfları ise birden fazla arşiv altında listelenmektedir. Örneğin makine öğrenmesi (machine learning) konu sınıfı hem istatistik (stat.ML) hem de bilgisayar bilimleri (cs.LG) arşivi altında listelenmektedir.

Çalışma kapsamında kullanılan pennant erişim algoritmasının performansı toplam atıf ve ortak atıf sayıları ile doğrudan ilgilidir. Bu yüzden iSearch derleminin ortak atıf oranları

[3] iSearch derleminde 20. sorgu yer almamaktadır (sorgu 19, sorgu 21 şeklinde devam etmektedir). Dolayısıyla her ne kadar en sonuncu sorgu 66 olarak geçse de toplam 65 sorgu bulunmaktadır.
[4] Kodlar için bkz. https://mugeakbulut.com/phd/codes/iSearch_verilerini_Arxivden_indirme.py
[5] Makaleler arXiv'de birden fazla konu altında listelenebilmektedir. Bu çalışma kapsamında ise her makale için sadece temel (primary category) konu kategorisi esas alınmıştır.
[6] Etkileşimli grafik için bkz. https://mugeakbulut.com/phd/gorsellestirme/bubble.html

hesaplanmıştır. iSearch derlemi için atıf ağı yoğunluğu 0,0021'dir. Diğer bir deyişle potansiyel olarak kullanılabilecek bağlantıların sadece 0,0021'i kullanıldığı için iSearch derlemi atıf ağı seyrek (sparse) bir ağdır. iSearch derlemindeki makalelerin atıf sayılarının ortalaması 15'tir (ortanca=5). Toplam 65 sorgu için pennant erişim algoritması uygulanarak erişilen makaleler için ortalama ortak atıf sayısı ise ikidir (ortanca=1).[7]

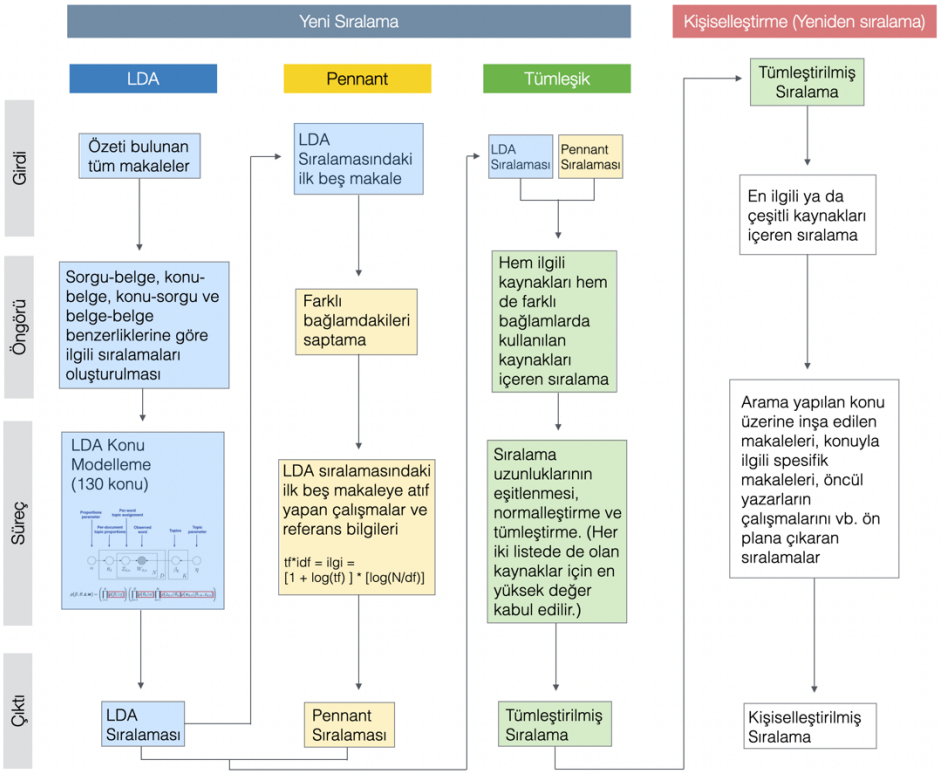
Yöntem

Bu araştırmada ilgi sıralamaları oluşturulurken hem kelime sıklıklarından hem de atıf ilişkilerinden yararlanılmıştır. İlgi oranı merkezilik değerleri ile, çeşitlilik oranı ise arXiv konu kategorileri baz alınarak ölçülmüştür. Makalelerin hem entellektüel içerikleri hem de belge bağlamları dikkate alınmıştır. İzlenen yol Şekil 1'deki gibidir. Süreçte tümleşik liste oluşturulması (LDA+Penannt erişim) ve sıralamaların kişiselleştirilmesi (yeniden sıralama) olmak üzere iki ana adım bulunmaktadır.

İlk aşamada makalelerin başlıkları ve özetleri üzerinde LDA olasılıksal konu modelleme algoritması[8] çalıştırılarak iSearch derleminde daha önceden tanımlanmış olan 65 sorgu için ilgi sıralamaları elde edilmiştir. Ardından ortak atıf ve toplam atıf verileri hesaplamaya dâhil edilerek pennant erişim yöntemi uygulanmıştır. Son olarak bu iki sıralama birleştirilerek tümleşik liste elde edilmiştir. Elde edilen sıralama sorguyla en ilgili makaleler ya da farklı alanlardan (ama sorguyla ilgili) makaleler üst sıralarda olacak şekilde yeniden sıralanabilmektedir.

Şekil 1

İlgi sıralaması oluşturulması sırasında uygulanan işlemler



[7] Sorguların 35'i için ortalama ortak atıf sayısı 1, 26'sı için 2, dördü için ise 3'tür.

[8] Kaynak kodlar için bkz. https://colab.research.google.com/drive/1dESqgDRL6WfyCSDgHxAa1kPDoHKBFK54

Konu Modellemesi

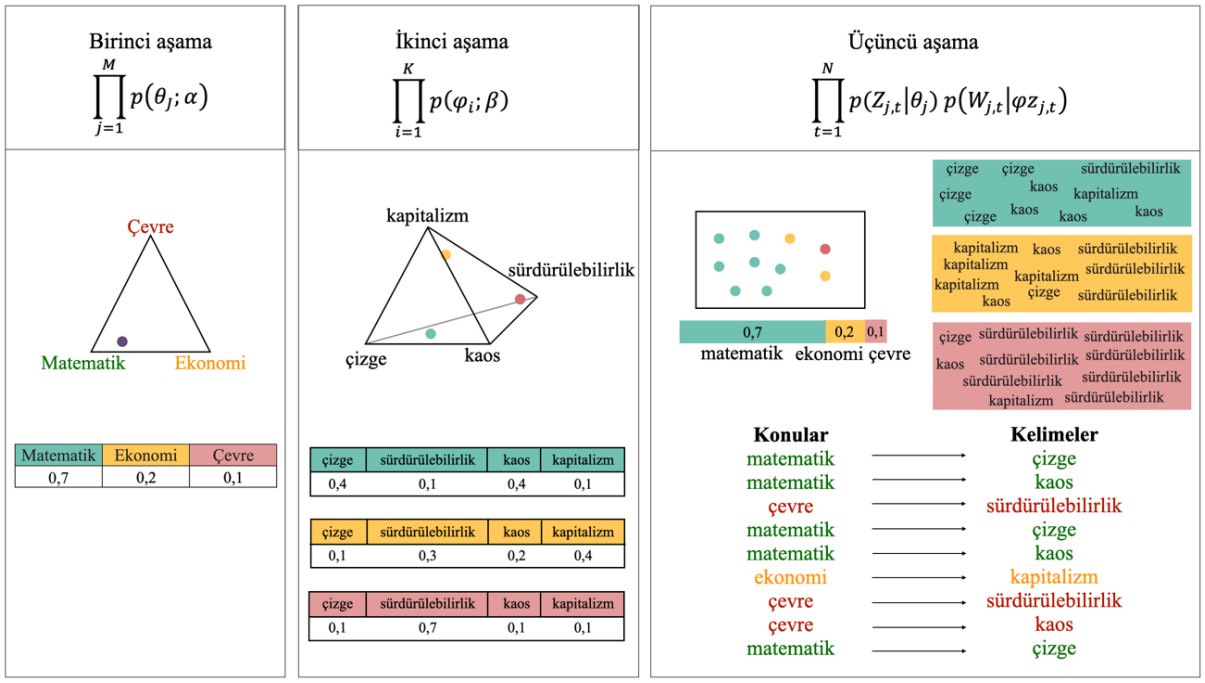
Bir belgenin, sorguyla ifade edilen bilgi ihtiyacını karşılama olasılığı 0 (ilgisiz) ile 1 (ilgili) arasında değişmektedir. Örneğin, bir belge belli bir konudaki bilgi ihtiyacını daha çok (diyelim ki %80 -veya 0,8 oranında), bir başka konudaki bilgi ihtiyacını ise daha az (%50 -veya 0,5 oranında) karşılıyor olabilir. Başka bir deyişle, söz konusu belgenin ilk konu için ilgi düzeyi 0,8, ikincisi için 0,5'tir (Akbulut, 2016, s. 11). Buradan hareketle LDA konu modelleme algoritması kullanılarak belge için saptanan konuların olasılık dağılımları elde edilir. LDA algoritması bir makalenin sınırlı sayıda konunun karışımından oluştuğu ve her kelimenin de makalenin konularından birisi ile ilişkilendirilebileceği varsayımına dayanır (Zhang ve diğerleri, 2015). LDA algoritması temelde üç aşamalı hiyerarşik Bayes modeline dayanır. Bayesçi çıkarım modeli önceki ilgili olayların çıktılarına ve bazı mantıksal varsayımlara dayanarak bir olayın olma olasılığını hesaplamak için kullanılır. Konu modellemede daha önceden analiz edilen bir dizi belgeye dayanarak belli bir konuyla ilgili kelimeler ve belli bir belgede işlenen konular algoritmadan elde edilen olasılık dağılımlarına göre öngörülür (tahmin edilir). Bayes yaklaşımında parametreler önsel (prior) bir dağılımdan gelen rastsal değişkenler olarak görülür (Alpaydın, 2017, s. 291). Bayes kuralı önsel olasılık ve olabilirliği birleştirip sonsal olasılık dağılımlarının (posterior probability distributions) hesaplanmasını sağlar. LDA modelindeki aşamalarda dağılımlar yeni bir öngörü dağılımı (bir sonraki aşama) için girdi olarak kullanılır.

$$p(W, Z, \theta, \varphi;\ \alpha, \beta) = \prod_{j=1}^{M} p(\theta_J; \alpha) \prod_{i=1}^{K} p(\varphi_i; \beta) \prod_{t=1}^{N} p(Z_{j,t}|\theta_j)\, p(W_{j,t}|\varphi z_{j,t}) \qquad (2)$$

Formül 2'de eşitliğin sol tarafı modelin olasılık değerini temsil etmektedir. Formülde kelimelerin konular, konuların da makaleler üzerinde olasılık dağılımları yer almaktadır (Blei, 2012, s. 80). *M* derlemdeki toplam makale sayısı, *K* toplam konu sayısı, *N* belli bir makaledeki kelime sayısı, *W* kelime, *Z* ise konu'dur. Kelimelerin konulardaki dağılımı φ, konuların makalede bulunma olasılığı ise θ ile temsil edilmektedir. Dirichlet parametreleri de α ve β'dır. Konuların makalelerdeki dağılımını α, kelimelerin konulardaki dağılımını ise β temsil eder (düşük α değeri makalelerin daha az sayıda konu içerdiğini belirtmektedir). Formülde üç ana adım bulunmaktadır (Şekil 2). Her adımda olasılık hesaplaması yapılır ve bu üç olasılığın çarpımı modelin olasılık değerini verir. Birinci aşamada her makale için, konuların (θ) makalelere dağılma olasılığı (p) hesaplanır. İkinci aşamada Dirichlet dağılımına göre kelimelerin (φ) konulara dağılım olasılığı belirlenir (β). Her makale için o makalede yer alan kelimelerin makalenin konuları ile ne kadar ilişkili olduğunun hesaplandığı üçüncü aşamada ise makalelere konuların atanması iki adımda gerçekleşmektedir. Önce makalede yer alan her kelime geçici olarak rastgele bir konuya atanır ve kelimelerin konulardaki dağılımı verildiğinde belli bir kelimenin o konuya ait olma olasılığı hesaplanır. Ardından da makaledeki kelimeler olasılık dağılımı olarak temsil edilir ve buna göre makalenin konuları belirlenir. Diğer bir deyişle konuların makalelerde bulunma olasılığı verildiğinde belli bir konunun o makaleye ait olma olasılığı belirlenir. Böylece her bir kelimenin belli konularla ilişkili olma olasılığı hesaplanır. Bu işlem tekrarlıdır (iterative). Herhangi bir konu için ulaşılan en yüksek değer bir kelimenin o konuyu temsil edebileceğini gösterir. Kelimelerin konu dağılımı yapıldıktan sonra makale-kelime matrisi oluşturulur. Bu sayede kelimelerin konulardaki ağırlıkları elde edilmiş olur ve makalenin konuları da bu ağırlıklar dikkate alınarak belirlenir.

Şekil 2

LDA algoritmasının aşamaları



Olasılıksal modellemede en kritik aşamalar ön işleme (pre-processing), veri temizleme ve konu sayısının belirlenmesidir. Ön işleme için ihtiyaç duyulan yöntemler derlemin diline, içeriğine ve kalitesine göre değişiklik göstermektedir (Wu ve diğerleri, 2020). Bu araştırma kapsamında konu modellemesine hazırlık için belirlenen özel karakter ve sayıların metinden atılması, büyük harflerin küçük harflere çevrilmesi,[9] metni cümle, kelime gibi birimlere ayırma (tokenization), dur kelimelerinin (stop words) atılması ve kelimelerin eklerini kaldırarak morfolojik köklerinin elde edilmesi (stemming) ön işlemleri gerçekleştirilmiştir.

Bir sonraki aşama olan veri temizleme aşaması da konu modellemesinin üreteceği sonuçlar açısından önemlidir. Konu modelleme için başlık ve özet bölümleri bir arada kullanılmıştır. Fakat arXiv'deki fizik makalelerinin özetleri genelde kısadır (başlık+özet için kelime uzunluğu ortalama 119, ortanca 107, maksimum 392, minimum 4). Bazıları ise çalışmanın konusu hakkında ipucu vermeyecek niteliktedir (örneğin, "Başlık: Conference Summary, Özet: Island Universes Conference Summary"). Derlemdeki geri çekilen (retracted) çalışmaların ise özetleri bulunmamakta, sadece çalışmanın geri çekildiğine dair bir not yer almaktadır. Bu tür çalışmalar ile özet ve başlık kelime toplam sayısı 25 kelimeden az olan 3617 makale (%0,8) kapsam dışı bırakılmış, LDA algoritması 431.196 makale üzerinde çalıştırılmıştır.

LDA konu modelleme algoritmasının çalıştırılabilmesi için konu sayısının önceden belirlenmesi gerekmektedir (Blei ve Lafferty, 2009; Ponweiser, 2012). Konu modelleme algoritmasının çalıştırılacağı derlemde kaç konu olduğuna önceden karar verilmesi gerekliliği çelişkili olmakla beraber kritik ve zordur (Gläser ve diğerleri, 2017). En uygun konu sayısını tespit etmek amacıyla farklı skorların üretildiği yaklaşımlar uygulanmaktadır. Bu yaklaşımlar genelde LDA'nın konu-terim, belge-konu vs. dağılımları aracılığıyla konu çiftleri arasındaki mesafeleri hesaplayarak en uygun konu sayısını saptamayı amaçlamaktadır. Bazı yaklaşımlarda

[9] Bu aşamada fizik alanı özelinde oluşturulan büyük ve küçük harfe duyarlı dur listesi (ADS Team, 2008) kullanılmıştır.

skorun yüksek olması, bazılarında ise düşük olması beklenmektedir (Carroll, 2018). iSearch derlemindeki yayınlar için konu sayısını belirlemek amacıyla dört ölçev (metric) temel alınmış (Arun ve diğerleri, 2010; Cao ve diğerleri, 2009; Deveaud ve diğerleri, 2014; Griffiths ve Steyvers, 2004) ve buna göre oluşturulan kod (Nikita, 2020) iSearch derlemine uyarlanmıştır. Konu sayısını belirlemek için kullanılan dört ölçevle ilgili ayrıntılı bilgi Ek 1'de verilmektedir.

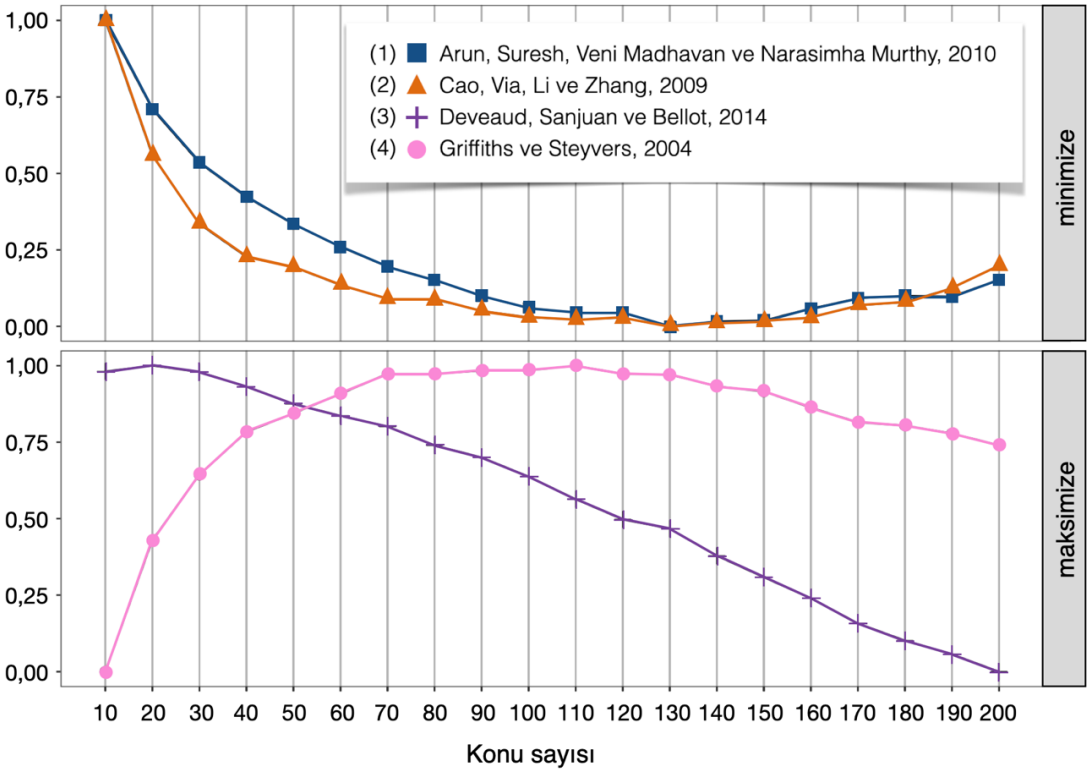
LDA algoritması uygulanırken, özellik sayısı (yani sabit kelime boyutu) α ve β Dirichlet ön parametrelerine ince ayar yapılabilir (George ve Doss, 2017; Pathik ve Shuklai 2020, s. 516). Bu parametrelerde yapılan değişikliklerin amacı, LDA'nın önsellerini (Dirichlet hiperparametreleri) ayarlayarak tahmine dayalı dağılımın entropisini en aza indirmektir (Zhang ve diğerleri, 2016, s. 1763). Fakat bu durum sadece küçük ölçekli ve çarpık kelime sıklıklarının görüldüğü doğal dil kullanılan belgeleri içeren derlemlerde geçerlidir. Derlem büyükse, hiperparametreler tahmin performansının ayarlanmasında önemsizdir (Zhang ve diğerleri, 2016, s. 1772). Bu yüzden konu modelleri uygulamalarında tipik olarak parametrelerin ayarlanmasının çok az pratik etkisi olduğundan sabit konsantrasyon parametreleri ve simetrik Dirichlet önselleri kullanılır (Wallach ve diğerleri, 2009, s. 1763). Makul büyüklükte bir derlem ortalama 1000-2000 belge ve 5000-7000 arası kelime içermektedir (Crossley ve diğerleri, 2017; Deerwester ve diğerleri, 1990, s. 394). Dolayısıyla bu araştırmada kullanılan derlem büyük ölçekli bir derlemdir. Algoritma çalıştırılırken orta düzey model için ön tanımlı parametreler kullanılmıştır.[10]

LDA algoritmasına girilecek konu sayısını belirlemek için kullanılan tüm ölçevlerde skorlar 0 ile 1 aralığındadır. Dört yaklaşımın bir arada gösterildiği Şekil 3'te "maksimize" olarak gruplanan ölçevler skorlarının yüksek olması beklenen, "minimize" olarak gruplananlar ise düşük olması beklenenlerdir. Tüm algoritmaların aynı optimal grup sayısını belirlemesi beklenemez, ancak ortak bölgeye -yüksek maksimize, düşük minimize- bakılarak en uygun konu sayısı belirlenebilir (Holliger, 2018). Ölçevler iSearch derlemine uygulandığında bu derlem için en uygun konu sayısının 110 ile 130 arasında olduğu anlaşılmaktadır (Şekil 3). Üçüncü ölçev hep düşme eğilimindedir. Diğer ölçevlerle uyumsuz olması ve bilgi verici bir örüntüye sahip olmaması sebebiyle konu sayısı belirlenirken bu ölçev göz ardı edilmiştir (Bayer ve Michael, Bonaccorsi, Melluso ve Massucci, 2022; 2019; Guillemette ve diğerleri, 2017; Holliger, 2018). Bu durum muhtemelen üçüncü ölçevin temelde kullanıcı sorgusundaki gizli kavramların sayısını tahmin etmek amacıyla kullanılmasından kaynaklanmaktadır. Bu araştırma kapsamında konu sayısı 130 olarak belirlenmiş ve hesaplamalar da buna göre yapılmıştır.

[10] LDA algoritması uygulanmadan önce hesaplanan istatistikler (tekil kelime sayısı, makale sayısı vs.) orta düzey (medium) model ile uyumludur. Bunun dışında araştırma kapsamında tercih edilen doğal dil işleme kütüphanesi SpaCy'de bazı parametreler makalelerin uzunluğuna göre dinamik olarak ayarlanmaktadır. Bu araştırmada kullanılan bazı parametreler şunlardır: konu sayısı (num_topics)=130, alpha='symmetric', tekrar sayısı (iterations)=50, gamma eşik değeri (gamma_threshold)=0,001, minimum olasılık (minimum_probability)=0,01.

Şekil 3

iSearch derlemine en uygun konu sayısının belirlenmesi



Pennant Erişim Algoritması

Pennant erişim yönteminde makalelerin çekirdek makale ile birlikte atıf alma sıklıkları aşağıda yer alan ağırlık (*tf*idf*) formülüne (Manning ve Schütze, 2000, s. 542) göre hesaplanarak ilgi değerleri belirlenmektedir.

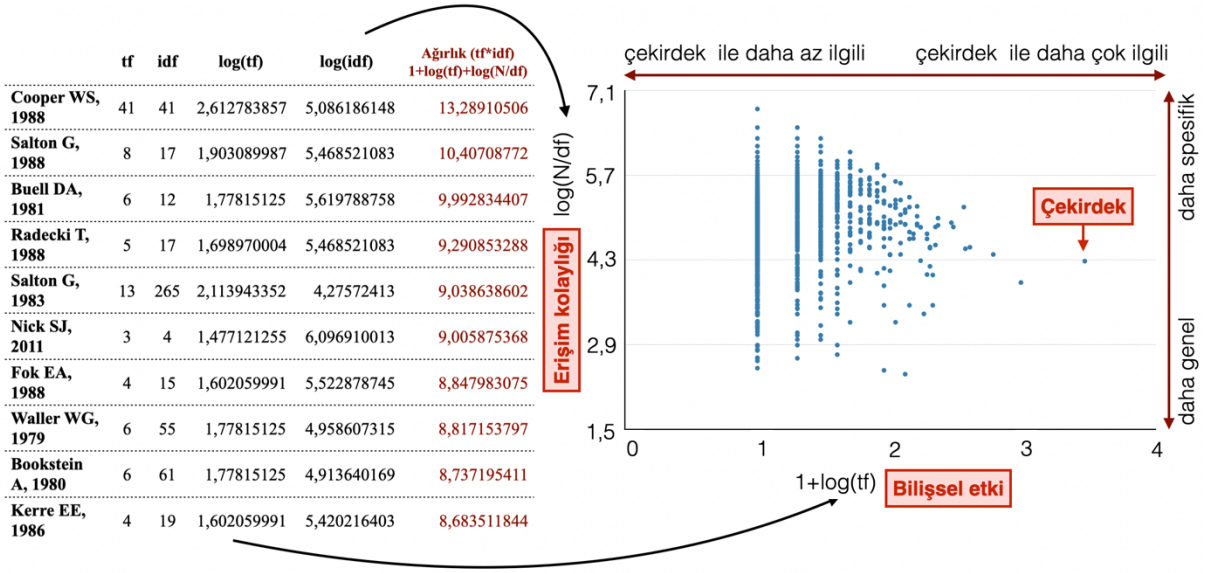
$$tf * idf = ilgi = [1 + \log(tf)] * [\log(N/df)] \quad (3)$$

Formül 3'te *df* ortak atıf alan makalelerin toplam atıf sayısı, *tf* bir makalenin çekirdek makale(ler) ile birlikte aldığı atıf sayısı, *N* ise derlemdeki toplam makale sayısıdır (iSearch derlemi için 434.813). Formüle göre yüksek ilgi puanı *tf* ve *idf* değerlerinin birbirine yakın olması anlamına gelmekte ve ilgili çalışmanın çekirdek[11] makaleye yakın olarak konumlandırılmasını sağlamaktadır (Akbulut, 2016). Örnek olarak Şekil 4'te William S. Cooper'ın (1988) Boole sisteminin sorunlarını tartıştığı makalesi çekirdek olarak belirlenmiştir. Makalenin "literatürdeki diğer çalışmaları nasıl etkilediği her bir yazarın çekirdek yazarlar ile ortak atıfları ve toplam atıflarının logaritmaları alınarak oluşturulmuş, bu etki pennant erişim yöntemi aracılığıyla görselleştirilmiştir" (Akbulut, 2016, s. 35). Pennant erişim yöntemi ile hesaplanan ilgi sıralamasında ilk sıralarda olan makaleler (ilgi puanı en yüksek olanlar) hem bilişsel etki hem de erişim kolaylığı ölçeklerinde (scales) en yüksek puanı almış olan makalelerdir.

[11] Çekirdek (seed) terimi, pennant erişim yönteminde literatürdeki etkisi belirlenmek istenen ya da ilgi sıralaması oluşturulan çalışma (veya yazar) anlamında kullanılmaktadır.

Şekil 4

Örnek pennant erişim gösterimi

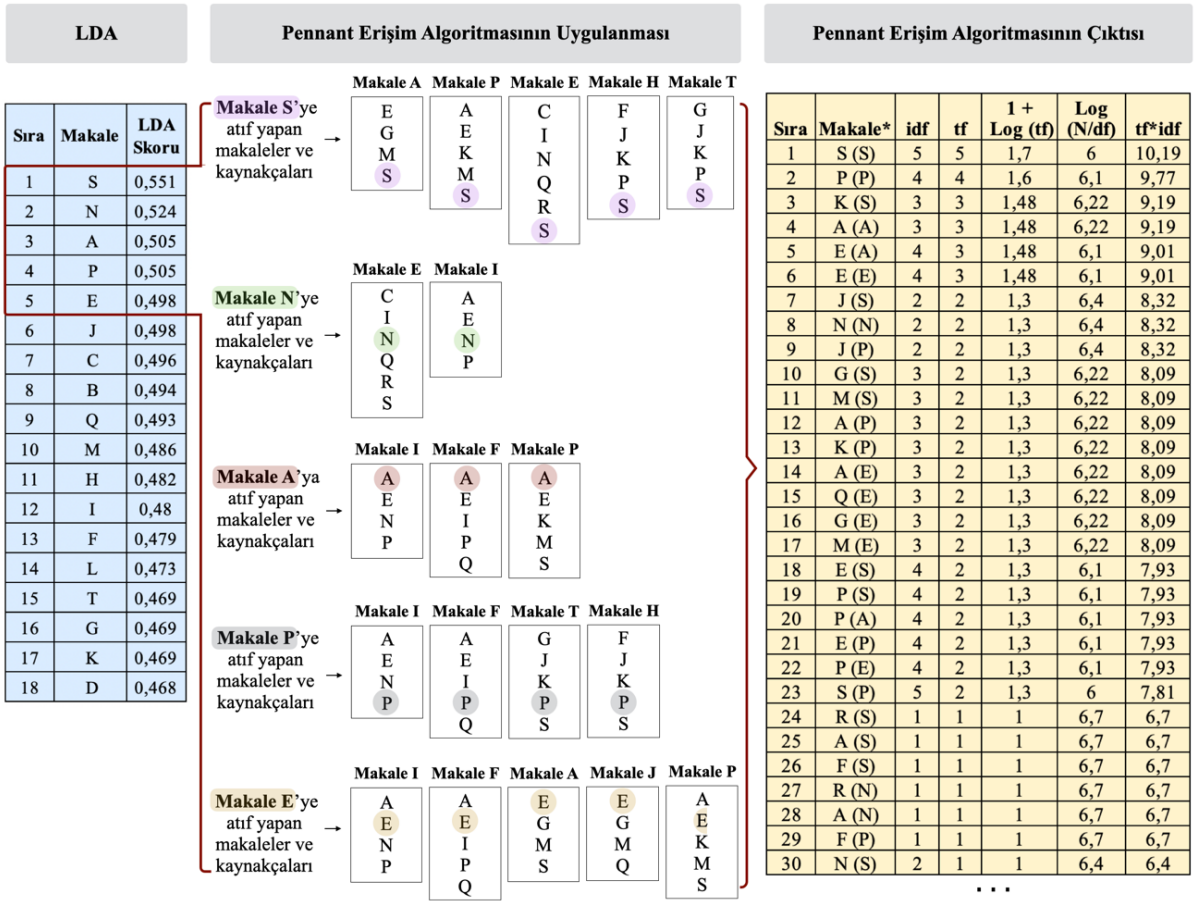


İlgi sıralaması hesaplanırken ortak atıf değerinden elde edilen bilişsel etki ölçüsü de hesaba katıldığı için, erişim kolaylığı ölçeğinde daha aşağıda olan makalelerin ilgi sıralamasında ilk sıralarda olabilmesi için bilişsel etki ölçeğinde de ilk sıralarda olması gerekir. Formüldeki *idf* faktörü çekirdek makale ile daha az ilgili makaleleri ilgi sıralamasında aşağı itmekte ve daha ilgili olanları ise yukarı taşımaktadır.

Pennant erişim algoritmasının uygulama aşamalarının örnek gösterimi Şekil 5'te verilmektedir. Elimizdeki derlemde 20 makale olduğunu (Makale A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S ve T) ve bu makalelerden O ve R'nin özetleri olmadığını varsayalım. Özeti bulunan tüm makalelere LDA algoritması uygulandıktan sonra, herhangi bir sorgu çalıştırıldığında LDA ilgi sıralaması elde edilmektedir. İlgi sıralamasındaki ilk beş makaleye atıf yapan makaleler kaynakçalarıyla birlikte değerlendirilerek toplam atıf ve ortak atıf sayıları hesaplanmakta ve pennant erişim algoritması uygulanmaktadır. Çekirdek makale ya da makalelere atıfta bulunan makalelerin kaynakçalarında yer alan her bir makale pennant sıralamasında (LDA+Pennant sıralaması) yer almaktadır (çekirdek makalelerden en az biriyle bir ve/veya daha fazla ortak atıf aldığı için). Pennant sıralamasında makaleler *tf ve idf* değerlerinin çarpımına göre büyükten küçüğe doğru sıralanmaktadır. *Tf*idf* değeri en yüksek olan makalenin sorguyla en ilgili makale olduğu varsayılmaktadır. Birden fazla çekirdek makale olduğu için yeni sıralamaya eklenen makaleler farklı çekirdek makaleler aracılığıyla sıralamaya eklenmektedir. Listedeki makalelerin hangi çekirdek makale vasıtası ile listeye girdiği parantez içinde belirtilmiştir. Eğer aynı makale birden fazla çekirdek makale aracılığı ile sıralamaya dâhil olduysa *tf*idf* değeri en yüksek olan makale dikkate alınmaktadır.

Şekil 5

Pennant erişim algoritmasının uygulama aşamaları



Pennant erişim algoritması her ne kadar temel kaynakları saptama konusunda başarılı sonuçlar sağlasa da bu algoritmanın bazı dezavantajları bulunmaktadır. Bunlardan en önemlisi, pennant erişim yönteminin doğası gereği bir çekirdek çalışmaya ihtiyaç duymasıdır. Bu araştırma kapsamında ise arama sorgusuyla ilgili kaynaklar LDA algoritması ile belirlenmiş, ardından da bu listedeki ilk beş kaynağa pennant erişim algoritması uygulanmıştır. Bu sayede pennant erişim için çekirdek makale şartı sorunu aşılmıştır. Pennant erişim algoritmasının bir diğer dezavantajı ise ortak atıf sayıları temel alındığı için ilgi sıralaması listesinin uzunluğunun kestirilememesidir. Ama pennant erişim yöntemi LDA algoritmasını destekleyici olarak kullanıldığı için derlemdeki tüm çalışmalar ilgi sıralamasına dâhil olmaktadır. İlgi sıralamalarında kullanıcıların genellikle ilk sıralara odaklandığı düşünüldüğünde liste kısa bile olsa tümleşik listede pennant erişimin etkisi görülmektedir.

iSearch derleminde 3,7 milyondan fazla atıf bulunduğu için, hesaplamanın daha hızlı yapıldığı SQL tabanlı bir platform tercih edilmiştir. Hazırlanan MS Access uygulaması[12] ile pennant erişim hesaplamaları yapılmıştır.

Listelerin Tümleştirilmesi

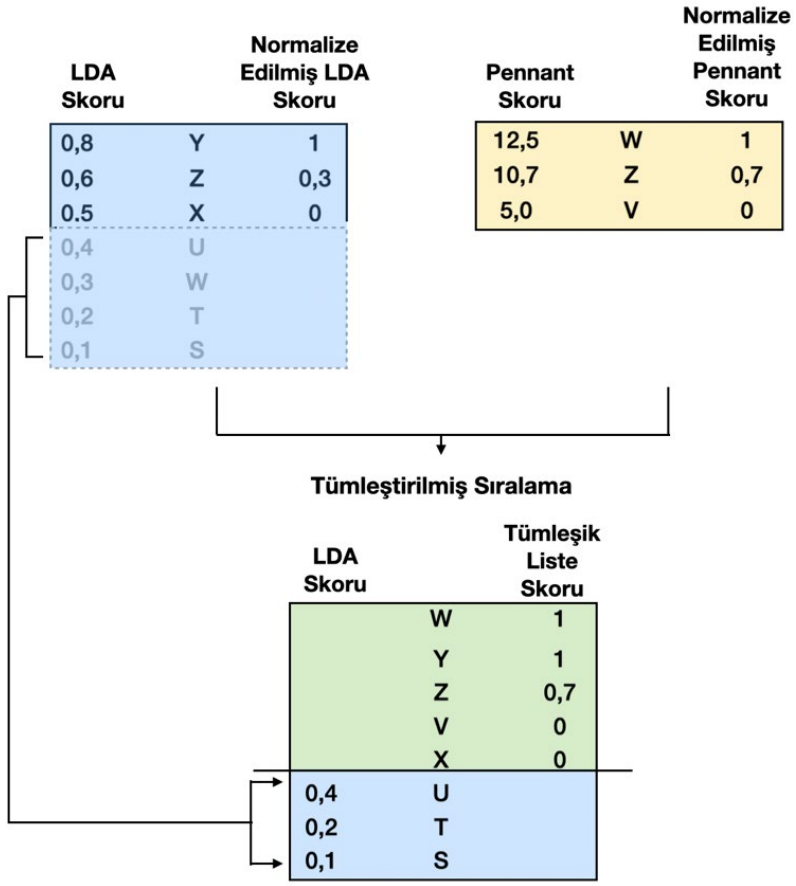
Pennant erişimin LDA sıralamasına katkısı ortak atıf sayıları ile doğrudan ilgilidir. Ortak atıf sayısı her çekirdek makale için farklı olduğundan pennant erişim algoritmasının ürettiği ilgi sıralamalarının uzunluğu da kestirilememektedir. Öte yandan LDA algoritması ise derlemdeki özeti olan tüm çalışmalar için bir sorgu-belge benzerlik değeri hesaplamaktadır. Bu iki listenin

[12] Bkz. https://mugeakbulut.com/phd/codes/pennant/

birleştirilmesi sırasında önce pennant sıralama listesi esas alınarak iki listenin uzunlukları eşitlenmiş, ardından eşitlenmiş uzunluktaki listelerdeki değerler normalize edilip[13] (min-max normalizasyon) tümleştirilmiştir (bkz. Şekil 6).

Şekil 6

Sıralama tümleştirme



Normalizasyon işlemi Formül 4'e göre yapılmıştır. Herhangi bir değerin normalize edilmiş yeni değeri hesaplanırken, o değerden en küçük değer çıkarılmakta ve en büyük değer ve en küçük değer arasındaki farka bölünmektedir.

$$x_{normalize} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Tümleştirme aşamasında her iki sıralama listesinde de olan ortak makaleler, normalize edilmiş değerlerinin en yüksek olduğu skorla tümleşik listeye dâhil olmaktadır (örneğin Şekil 6, Makale Z için normalize edilmiş pennant skoru 0,7). Derlemde özet bilgisi olmayan makaleler varsa (Şekil 6, örneğin Makale V) LDA skoru hesaplanamamaktadır. Fakat bu makaleler için ortak atıf verisi varsa pennant erişim yöntemi ile tümleştirilmiş listede bu makaleler de yer almaktadır. Tümleşik liste uzunluğu tamamlandığında ise LDA sıralamasında en başta kesilen çalışmalar skoru en yüksek olandan başlayarak (Şekil 6'daki LDA sıralamasındaki U makalesinden itibaren) tümleşik listenin arkasına eklenmektedir.

[13] Normalize etme işleminde minimum değer 0'a, maksimum değer 1'e ve diğer tüm değerler 0 ile 1 arasında ondalık sayıya dönüştürülmektedir (Thara ve diğerleri, 2019).

İlgi Sıralamalarının Kişiselleştirilmesi

Kişiselleştirme aşamasında ise ilgi sıralaması kullanıcının ihtiyacına göre liste en ilgilileri önceleyecek ya da interdisipliner yapıda olacak şekilde yeniden sıralanmaktadır (re-ranking). Bu adımda tümleştirme fonksiyonunda ağırlıklar pennant ya da LDA algoritması ağırlıklı olacak şekilde ayarlanarak sorguyla en ilgili olan ve farklı alanlardan kaynakların tümleşik listede üst sıralarda yer alması sağlanmış ya da her ikisini de maksimize eden bir sıralama oluşturulmuştur. Kişiselleştirme aşamasında her iki sıralama listesinde de bulunan ortak makaleler, tümleştirme öncesinde hesaplanan ve normalize edilmiş değerlerde en yüksek olan skor değil, ağırlıklı olması istenen algoritmadaki değeri ile listeye dâhil edilmektedir (örneğin Şekil 6, Makale A için normalize edilmiş LDA skoru 0,3).

Performans Değerlendirme

İlgi ve Çeşitlilik. iSearch derlemi her ne kadar uzmanlar tarafından derecelendirilmiş ilgi değerlendirmeleri içeriyorsa da, derlemde ilgi değerlendirmesi içeren yayınların oranı çok düşük (yaklaşık %2) olup bu oran listeleri ilgi açısından değerlendirmek için yeterli değildir. Derlemdeki her makaleye karşılık bir ilgi değeri olması gerekmektedir. Bu bağlamda listelerin karşılaştırılması için merkezilik kavramına ilişkin ölçevlerden derece merkeziliğinin (degree centrality) kullanılmasına karar verilmiştir. Derece merkeziliği, kabaca, ağdaki bir noktanın (node) diğer noktalara olan bağlantı (tie) sayısıdır. Bir noktanın bağlantı sayısı arttıkça derece merkezilik değeri de artar. İlgi olarak merkezilik (centrality-as-relevance), paragraf özetlemek (Marujo ve diğerleri, 2017; Ribeiro ve de Matos, 2011) ve ağ analizini geliştirmek (Giustolisi ve diğerleri, 2020) için kullanılmıştır. İlgi olarak merkezilik hesaplamasında belli bir kümedeki terimleri en iyi yansıtan ağırlık merkezi (centroid) değeri esas alınmakta ve belli bir küme içindeki merkez terim temel alınarak diğer terimlerin buna uzaklığı hesaplanmaktadır. Bu çalışma kapsamında da bütün ağ dikkate alınarak makalelerin belli bir sınıftaki (belli bir arXiv temel konu kategorisindeki) ağırlığı hesaplanmıştır. Böylece iSearch derlemindeki her makale için ilgi değerleri elde edilmiştir. Merkezilik değerleri NetworkX Python paketi[14] kullanılarak hesaplanmıştır. Öte yandan listelerin çeşitlilik oranı ise içerdikleri makalelerin arXiv temel konu kategorileri incelenerek hesaplanmıştır.

LDA algoritmasında her ne kadar kelimelere dayanan konu bazlı bir analiz yapılsa da, bu konular arXiv'de yer alan konu kategorilerinden bağımsızdır. Bu çalışmada ise merkezilik değerleri hesaplanırken arXiv sistemindeki konu başlıklarından yararlanıldığından, bu durum LDA algoritması ya da pennant erişim yöntemiyle elde edilen sıralamalar açısından herhangi bir ön yargı oluşturmamaktadır. Bu çalışmada derece merkeziliği değerleri "ilgi değeri", arXiv konuları ise "çeşitlilik" olarak değerlendirilmiştir. Böylece kelimeler arasındaki ilişkilere dayanan konu modelleme algoritması ile ortak atıf değerlerini dikkate alan pennant erişim algoritmasının benzer ve farklı yönleri ortaya çıkarılmıştır.

Maksimum Marjinal İlgi Algoritmasının İlgi Sıralamalarına Etkisi. Çeşitlilik oranı yüksek listelerin oluşturulması için ilgi yeniliği (relevance novelty) değerinin ölçülmesi gerekmektedir. Maksimum Marjinal İlgi (Maximal Marginal Relevance - MMR) yaklaşımı ilgi düzeyini ve yeniliği bağımsız olarak ölçmeye ve sonuçları doğrusal bir biçimde birleştirmeye olanak sağlamaktadır (Carbonell ve Goldstein, 1998). MMR yaklaşımında doğrusal kombinasyon "marjinal ilgi" olarak adlandırılır. Erişilen makalenin yüksek marjinal ilgi düzeyi ölçütünü sağlaması için hem sorguyla alakalı olması hem de önceden seçilen makalelerle *minimum* benzerlik göstermesi gerekmektedir. MMR yaklaşımı bilgi erişim performansını %8 ile %17 oranında artırmaktadır (Yang ve diğerleri, 2007).

[14] https://networkx.org. Örnek kodlar için bkz. https://tinyurl.com/pebwtkjr.

MMR algoritması benzer cümleleri ya da kaynakları elediği için (başka bir deyişle konunun çeşitli yönlerini içeren farklı cümle veya kaynaklara eriştiği için) metin özetleme algoritması marjinal ilgi değeri daha yüksek sıralamalar oluşturmak için kullanılmaktadır. MMR algoritması kavramsal olarak bu çalışmada önerilen ve marjinal ilgili makaleleri sıralayan tümleşik liste sıralama algoritmasına benzemektedir (bkz. Tablo 1). Her iki algoritma da sorguyla ilgili ama marjinal kaynakların da yer aldığı bir liste oluşturmak amacıyla kullanılmaktadır. MMR algoritması sistematik bir şekilde önce sorgu ile en ilgili makaleyi listenin ilk sırasına yerleştirip ardından o makalenin konusuyla ilgili ama ona en az benzeyen makaleleri listeye eklemektedir. Bu sayede MMR, hemen hemen aynı bilgileri içeren ve birbirinin tekrarı olan çalışmaları (belge-belge benzerlik oranı yüksek olan çalışmaları) lüzumsuz (redundant) olarak işaretleyerek sıranın sonlarına doğru itip yeniden sıralamaktadır. Tümleştirme algoritması ise kelime sıklıklarına göre hesapladığı ilgi skorları ile atıflardan yola çıkarak hesapladığı belge-belge benzerliklerini kullanarak her iki listeden de en yüksek olan skorları üst sıralarda gösterecek şekilde bir sıralama yapmaktadır.

Tablo 1

MMR ve tümleştirme fonksiyonunun karşılaştırılması

	Tümleştirme Algoritması	**MMR Algoritması**
Amaç	LDA ve pennant erişimde en yüksek ilgi değerleri alan belgeleri önceleyerek sıralamayı artırımlı olarak geliştirme (incremental refinement)	Marjinal ilgili belgeleri önceleyecek şekilde sıralamayı artırımlı olarak geliştirme
Yöntem	Skorlara göre birleştirme	Sistematik birleştirme
Girdi	Terim sıklıkları (LDA), atıflar (pennant)	Belge-belge benzerliği ve sorgu-belge benzerliği
Kişiselleştirme	Pennant ve LDA algoritmalarının ağırlığı	λ (lambda) değeri

Bu araştırmada performans değerlendirme ve karşılaştırma (benchmarking) amacıyla MMR algoritmasından yararlanılmıştır.[15] MMR algoritmasının LDA, pennant erişim ve tümleştirme algoritmaları uygulanarak ayrı ayrı elde edilen ilgi sıralaması listelerine etkileri incelenmiştir. Etki oranlarına bakılarak hangi algoritmanın hangi özellikleri öne çıkardığı saptanmıştır.

Bulgular ve Yorum

Aşağıda araştırmadan elde edilen temel bulgular sunulmaktadır.

Algoritmaların Tümleşik Sıralama Listelerine Katkısı

Tümleştirme algoritmasının sistematik bir yapısı olmadığı için LDA ve pennant algoritmalarının tümleşik listeye olan katkıları her sorgu için farklılık göstermektedir. Tümleştirme algoritmasında makaleler hem atıflarına hem de özet ve başlıklarındaki kelime sıklıklarına göre değerlendirilmektedir. Fakat tümleştirme algoritmasında MMR algoritmasındaki gibi bir sistematik birleştirmeden ziyade skorlar ön plandadır. Örneğin, bir sorguya karşılık erişilen makalelerde kelime sıklıklarının olasılıksal dağılımlarında çok baskın

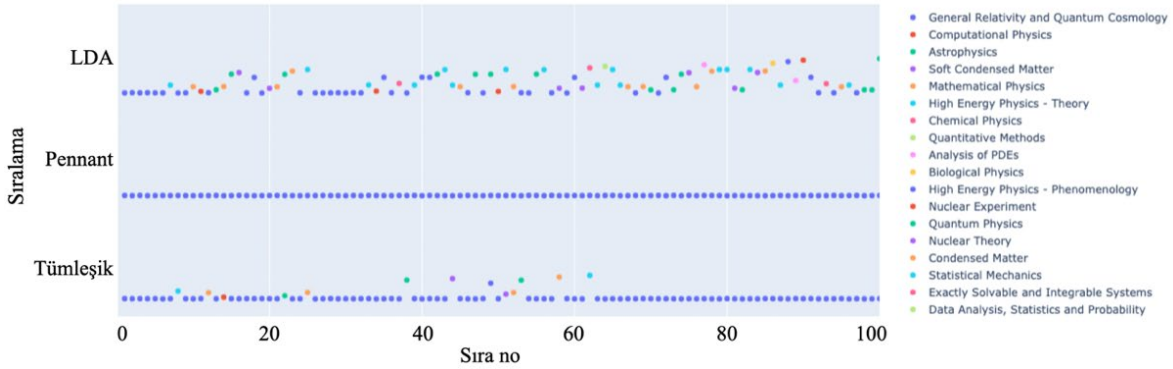
[15] Tümleştirilmiş sıralama listesine uygulanan MMR algoritması kodlarına https://colab.research.google.com/drive/1dESqgDRL6WfyCSDgHxAa1kPDoHKBFK54 adresinden erişilebilir.

bir örüntü varsa tümleştirilmiş listede de "LDA baskın" bir yapı gözlenmektedir.[16] Toplam atıf sayısı az ama ortak atıf sayısının fazla olduğu durumlarda ise bunun tam tersi bir durum söz konusu olmakta ve tümleşik listede "pennant baskın" bir yapı gözlenmektedir. Makalenin toplam atıf sayısı ve ortak atıf sayısının birbirine yakın olması o makalenin çoğunlukla çekirdek makale ile birlikte anıldığı anlamına gelmektedir. Dolayısıyla o makale derlemdeki diğer makalelere göre, çekirdek makale(ler) ile daha ilgilidir. Şekil 7, pennant baskın tümleşik sıralama listesine bir örnektir (sorgu no. 66). Bu sorgu için LDA algoritmasının eriştiği ilk 100 makale çok farklı konulardadır. Öte yandan, söz konusu makalelere pennant erişim algoritması uygulandığında ise ilk sıralarda erişilen makalelerin tutarlı bir şekilde aynı konuda olduğu görülmektedir.

Şekil 8'de LDA ve pennant erişim algoritmalarının skorları, Şekil 9'da ise tümleşik sıralama listesine LDA ve pennant erişim algoritmalarının katkısı gösterilmektedir (66. sorgu için). İlk 12 makale için kelime sıklıklarına göre olan sonuçlar daha ön planda olmasına karşın 13. sıradan sonra ortak atıf sayılarının ön plana çıktığı izlenmektedir. Skorlarda 13. sıradaki çakışmanın yansıması tümleşik listeye katkı grafiğinde de gözlenebilmektedir. Kesme noktası 25'ten sonra listede pennant ağırlıklı bir sıralama göze çarpmaktadır.

Şekil 7

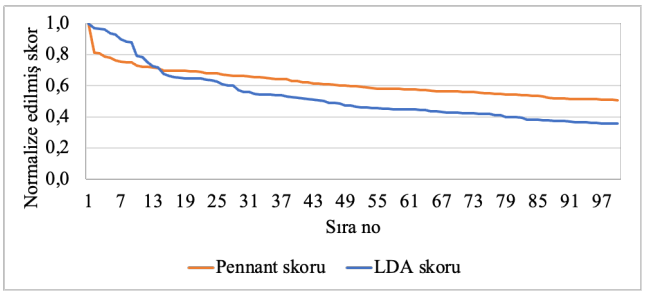
Pennant baskın tümleşik sıralama listesi (Sorgu no: 66)



Not. Listedeki ilk 100 makale gösterilmiştir.

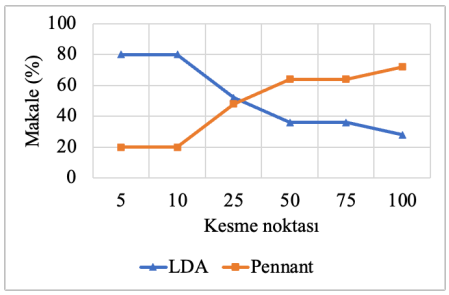
Şekil 8

Algoritmaların ilgi skorları (Sorgu no: 66)



Şekil 9

Algoritmaların çeşitli kesme noktalarında tümleşik listeye katkıları (Sorgu no: 66)



Daha önce de belirtildiği gibi toplam atıf ve ortak atıf verilerine göre işletilen pennant erişim çıktısının kaç makaleye erişeceği kestirilememekte, çok uzun ya da çok kısa listeler olabilmektedir. Pennant sıralama listesinin kısa olduğu durumlarda tümleşik sıralama listesinde

[16] Bkz. https://mugeakbulut.com/phd/gorsellestirme/liste_konu_grafik_100.pdf ve https://mugeakbulut.com/phd/gorsellestirme/kesme_noktalari.pdf, Sorgu no: 23.

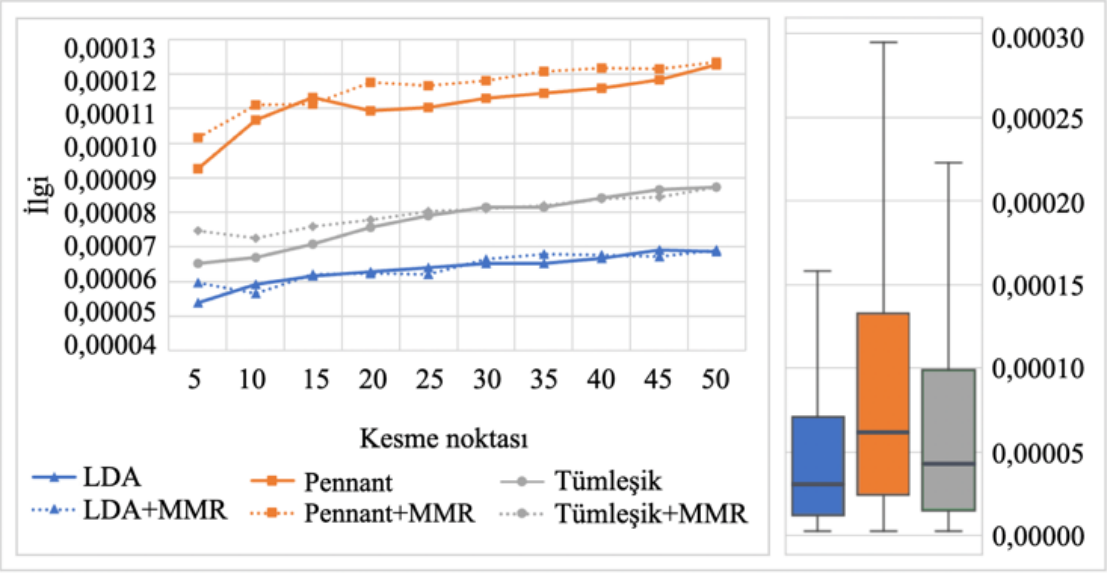
bir noktadan sonra sadece LDA kaynakları yer almaktadır. Örneğin, 65. sorguda pennant algoritması işletildiğinde sadece 22 makaleye erişilmektedir (dolayısıyla LDA çıktısı da 22'ye sabitlenmiştir).[17] Tümleşik listede ise 39. sıradaki makaleden sonra sadece LDA listesinden gelen kaynaklar yer almaktadır.

Algoritmaların İlgi Değerleri ve Konu Çeşitliliğine Göre Karşılaştırılması

Şekil 10 algoritmaların 65 sorgu için çeşitli kesme noktalarında ilgi değerlerinin ortalamalarını vermektedir. Şeklin *y* eksenindeki "ilgi" değerleri iSearch derlemindeki makalelerin 65 sorgu için derece merkeziliğine dayanan (centrality-as-relevance) ilgili olma olasılıklarının dağılımına göre hesaplanmış olup, tüm makalelerin ilgi değerlerinin toplamı 1'dir (Ribeiro ve de Matos, 2011, s. 280). Şekil 11 ise farklı algoritmaların çeşitli kesme noktalarında eriştikleri makalelerin ortalama kaç farklı konu içerdiklerini göstermektedir. Şekil 10'da pennant erişim algoritmasının merkezilik değerlerine göre tüm kesme noktalarında açık ara daha ilgili kaynakları sıraladığı gözlenmektedir. LDA ve pennant erişiminin birleştirilmesi ile elde edilen tümleşik sıralama listesi ise merkezilik değerleri açısından pennant erişime göre daha az ilgili kaynaklara erişmektedir. Fakat araştırmanın amacı algoritmik olarak tüm yapısal sıralamayı, kelime sıklıkları ve atıfları birlikte değerlendirerek yaratmak olduğundan ilgi (Şekil 10) ve çeşitliliği (Şekil 11) birlikte yorumlamakta fayda vardır. LDA algoritması farklı konulardan çalışmaları sıralarken, pennant erişim algoritması tutarlı bir şekilde ilgi değeri yüksek ve benzer konudaki çalışmalara ilk sıralarda erişmektedir. İki algoritmanın birleştirilmesi ile hem çeşitlilik hem de ilgi oranı yüksek bir liste (tümleşik) elde edilmektedir.

Şekil 10

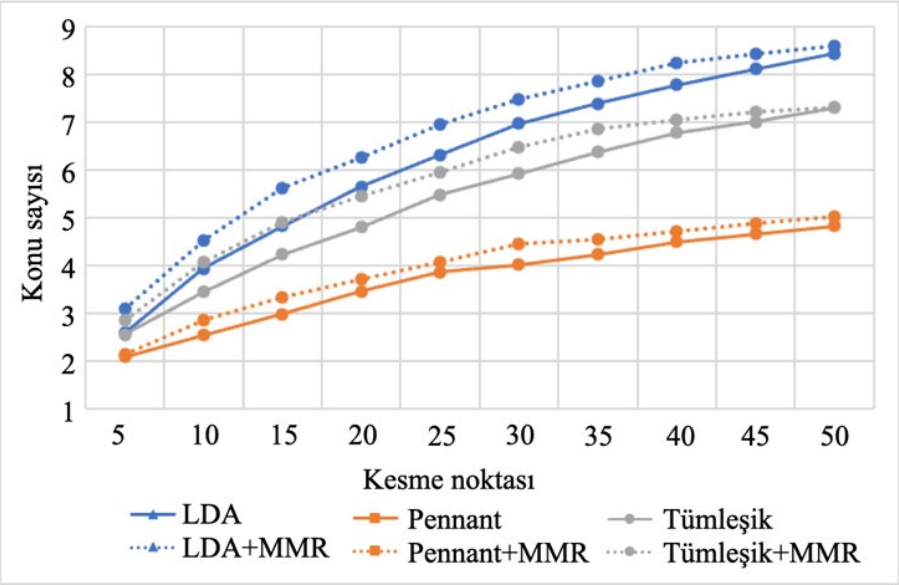
Algoritmaların ilgi değerlerine göre karşılaştırılması (N=65)



[17] Bkz https://mugeakbulut.com/phd/gorsellestirme/liste_konu_grafik_100.pdf (Sorgu no. 65).

Şekil 11

Algoritmaların konu çeşitliliğine göre karşılaştırılması (N=65)



Daha önce de belirtildiği gibi, çalışma kapsamında önerilen tümleşik sıralama listesi kavramsal olarak MMR algoritmasının çıktısına benzemektedir. MMR algoritması, ilgi sıralamalarını hem sorguyla ilgili hem de çeşitlilik oranı yüksek olacak şekilde yeniden sıralamak için de kullanıldığından, bu çalışmada LDA, pennant ve tümleşik erişim algoritmalarının sıralama listelerine MMR algoritması uygulandıktan sonra erişilen makaleler ilgi ve çeşitlilik açısından karşılaştırılmış ve farklı algoritmaların hangi özellikleri öne çıkardığı saptanmıştır. Başka bir deyişle MMR algoritmasındaki ilgi yeniliği (λ=0,5) bu çalışmada bir tür sağlama işlevi görmüştür.

Şekil 10 ve 11'deki noktalı çizgiler ilgili algoritmaların MMR algoritması (λ=0,5) uygulanmış halini temsil etmektedir. LDA algoritmasına kıyasla pennant erişim çıktısına uygulanan MMR'nin daha etkili olduğu gözlenmektedir (Şekil 10). MMR uygulandığında kesme noktası 15 hariç tüm kesme noktalarında daha ilgili kaynaklara üst sıralarda erişilmektedir. Çünkü pennant erişim algoritması benzer konudaki makaleleri üst sıralara yerleştirirken, MMR algoritması ise benzer makaleleri, ilgi sıralamasının sonuna doğru itip farklı konularda ama gene de sorguyla ilgili olan makaleleri üst sıraya yerleştirmektedir. Pennant listesindeki makaleler tutarlı bir şekilde aynı konuda olduğu için kesme noktası arttıkça ilgi oranlarının düşmesi normaldir. MMR uygulandığında ise birbirine çok benzeyen makaleler alt sıralara itildiği ve fakat farklı konularda olan ama çekirdek çalışmayla en ilgili makaleler üst sıralara yerleştirildiği için neredeyse tüm kesme noktalarında MMR'nin daha etkili olduğu gözlenmektedir.

Tümleşik sıralama listesinde ise hem kelime sıklıkları (LDA) hem ortak atıf bağlantıları (pennant erişim) kullanılarak sorguyla doğrudan ilgili makaleler -benzer olanları alt sıralara itme gibi bir endişe olmadan- üst sıralara yerleştirilmektedir. Bu listeye MMR uygulandığında da üst sıralarda daha ilgili kaynaklara erişilmektedir. Fakat kesme noktası 25'ten sonra ilgi değerleri tümleşik algoritmanın eriştiği makalelerin ilgi değerleri ile neredeyse aynıdır (diğer bir deyişle MMR etkisi yok olmaktadır). Bu durum pennant algoritmasının tümleşik listeye yansıması olarak değerlendirilebilir. Pennant algoritması kesme noktası 25'e kadar genellikle benzer konudaki çalışmaları ilk sıralara yerleştirmiş ve kesme noktası 25'ten sonra ise marjinal kaynakları sıralamaya eklemeye başlamıştır. iSearch derlemi yerine atıf yoğunluğu nispeten daha yüksek olan bir derlem üzerinde pennant erişim algoritması çalıştırılsaydı, marjinal kaynakların listeye eklendiği kesme noktası muhtemelen 25'ten daha büyük olacaktı.

Farklı sıralama algoritmalarının erişilen belgelerin hangi özelliklerini öne çıkardıklarının anlaşılabilmesi için sıralama listelerini bir de çeşitlilik açısından incelemek gerekir. LDA algoritması sorgularla ilgili çeşitli konulardan belgelere tümleşik ve pennant algoritmalarına kıyasla daha sık erişmektedir (Şekil 11). Bu, beklenen bir sonuçtur. Çünkü farklı belgelerdeki kelimeler arasındaki ilişkileri dikkate alan LDA algoritması sorgularla ilgili belgeleri daha çok sayıda alt kümelere ayırabilmektedir. Oysaki belgeler arasındaki ortak atıflara dayanan pennant algoritmasındaki ilgili konu çeşitliliği nispeten daha düşüktür. Başka bir deyişle, salt kelimeler arasındaki ilişkiler dolayısıyla LDA algoritması sorguyla marjinal ilgisi olan belgelere erişebilir. Ama sorguyla ilgisi olmayan farklı konulardaki belgelerde aynı kaynaklara ortak atıf yapılması daha düşük bir olasılıktır.

LDA, pennant ve tümleşik sıralama listelerine MMR algoritması uygulandığında hemen hemen tüm kesme noktalarında daha çeşitli konulardaki kaynaklara en üst sıralarda erişildiği gözlenmektedir (Şekil 11). Fakat pennant sıralama listelerine uygulanan MMR algoritması LDA ve tümleşik sıralama listelerininkine oranla çok daha az etkili olmuştur. Bunun başlıca nedeni ortak atıf analizine dayanan pennant erişim algoritmasının konuyla doğrudan ilgili olan ve ortak atıf yapılan (yüksek kesin isabet) makalelerin yanı sıra, konunun sınırlarını genişlettiği (boundary spanning) için seyrek ortak atıf yapılan (düşük kesin isabet) makalelere de erişmesidir. Bu tarz marjinal ilgili makaleler genellikle ortak atıf yoluyla başka disiplinlerle ilişki kurulmasını sağlayan makalelerdir. Bunun MMR'deki karşılığı ise konusal olarak ilgili makaleler sıralamaya girdikten sonra, ilgili olarak işaretlenenlere daha az benzeyen makalelerin de sıralamaya eklenmesidir.

Öte yandan pennant erişimde ortak atıf yapılan makaleler sorguyla ilgiliyse, konuları aynı olsa bile sıralamanın en başına eklenmektedir. Dolayısıyla pennant erişimde MMR algoritmasındaki gibi daha önce erişilen ilgili makalelere benzeyen makaleleri "cezalandırma" (daha alt sıralara itme) kaygısı söz konusu değildir. Şekil 10'daki pennant erişim çizgisinde kesme noktası 15'teki ilgi değeri artışı da bu yüzdendir. Aynı konuda olan ve daha çok ortak atıf alan makaleler listenin ilk sıralarına eklenmektedir. Pennant erişimde bir makaleye diğer disiplinlerde yayımlanan makalelerden sık atıf yapıldığı zaman bu makaleler de sıralama listesine üst sıralardan girebilmektedir.

Ortak Atıf Sayılarının Algoritmaların İşleyişine Etkisi

Ortak atıf (*tf*) ve toplam atıf (*df*) sayıları pennant erişim algoritmasının performansını doğrudan etkilediği için sıralama listeleri ortak atıf eşiği belirlenerek ilgi ve çeşitlilik açısından incelenmiştir. Eşik değerleri her bir sorgu için pennant erişim algoritması uygulandıktan sonraki sıralama listesi kullanılarak belirlenmiştir. Bütün sorgular arasında pennant ilgi sıralaması en kısa olan listede (sorgu 60) toplam 22 makale listelenmiştir. Bu nedenle tüm sorgular için ilk 22'şer makale incelenmiş ve sıralama listelerindeki ilk 22 makalenin ortak atıf ortalamaları (minimum 1, maksimum 67) dikkate alınmıştır. Tüm sorgular, erişilen makalelerin ortak atıf sayısı beş ve daha küçük olanlar (37 sorgu) ve beşten büyük olanlar (28 sorgu) olarak sınıflandırılarak MMR algoritmasının tümleşik sıralama listelerine olan etkisi ilgi ve çeşitlilik açısından değerlendirilmiştir.

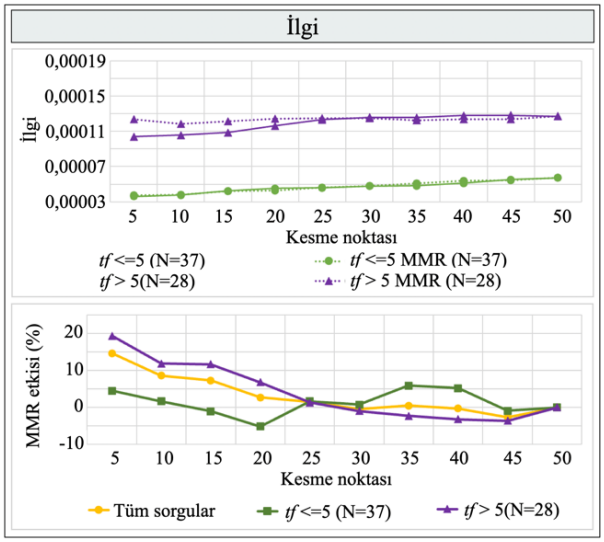
Genelde ortak atıf sayısı arttıkça ilgi oranları da artmaktadır. Örneğin, tümleşik sıralama listelerinde ortak atıf sayısı beşten büyük olan sorguların ilgi değeri, beş ve daha küçük olanların ilgi değerinin ortalama 2,6 katıdır (Şekil 12). MMR algoritmasının pennant sıralama listelerine etkisi kesme noktası 25'e kadar yüksek iken 25'ten sonra azalmaktadır. Etkinin neredeyse yok olduğu kesme noktası 25'te tümleşik sıralama listelerine marjinal olarak tanımlanabilecek ortak atıf aracılığıyla farklı disiplinlerle bağlantısı keşfedilen makaleler eklenmektedir. Ortak atıf sayısı beşten büyük sorgular için de aynı durum söz konusudur. Ortak atıf ortalaması beş ve daha küçük olan sorgular için ise pennant erişim algoritmasının ilgili

kaynaklara erişme oranı daha düşük olduğu için MMR algoritması kesme noktası 25'ten sonra da ilgili kaynaklara erişmeye devam etmektedir.

Çeşitlilik açısından ise ortak atıf sayısı arttıkça benzer konulardaki çalışmalar ilk sıralarda yer aldığı için çeşitlilik azalmaktadır. Tümleşik sıralama listelerinde ortak atıf sayısı beş ve beşten küçük olan sorguların eriştiği makalelerdeki tekil konu sayıları beşten büyük olan sorgularınkinin 1,4 katıdır (Şekil 13). Ortak atıf sayısı beşten büyük makaleler için kesme noktası 10 ve 20 arasında MMR algoritmasının etkisinin yüksek olması, muhtemelen pennant erişim algoritmasının benzer konudaki çalışmalara erişmesinden ve bu durumun tümleşik sıralama listelerine yansımasından kaynaklanmaktadır. İlgi açısından MMR'nin etkisinde tam tersi bir örüntünün gözlenmesi bu yorumu desteklemektedir. Söz konusu kesme noktalarında MMR algoritmasının etkisi çeşitlilik açısından yüksektir. Çünkü pennant erişim algoritması ortak atıf sayısı yüksek olduğu zaman aynı konulardaki ilgili çalışmaları tutarlı bir şekilde üst sıralarda listelemektedir. Ortak atıf sayısı beş ve beşten daha küçük olan sorgular için ise etki daha düşük olmasına karşın, özellikle kesme noktası 20'den sonra, benzer bir örüntü gözlenmektedir. Tüm sorgular söz konusu olduğunda da, kolayca tahmin edileceği gibi, ortalama bir etki söz konusudur.

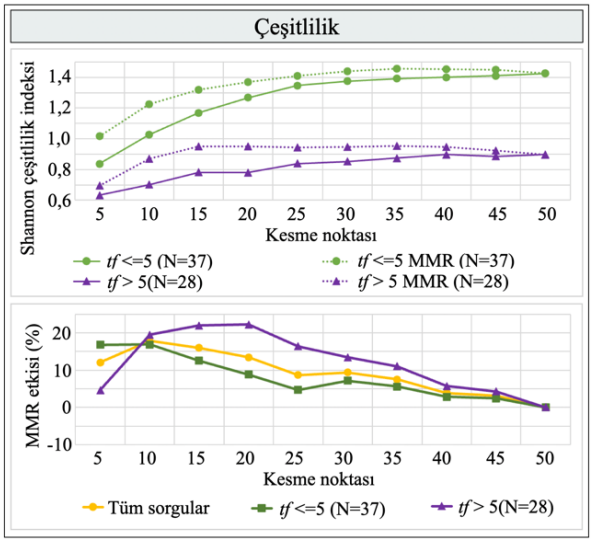
Şekil 12

Tümleşik sıralama listesi için ilgi değerleri ve MMR etkisi (ortak atıf eşiği <=5)



Şekil 13

Tümleşik sıralama listesi için konu sayıları ve MMR etkisi (ortak atıf eşiği >5)



Genel olarak hem ilgi hem de çeşitlilik açısından MMR algoritmasının etkisi sıralama listelerinin ilk sıralarında gözlenmektedir. Pennant erişim algoritması kesme noktası 25'e kadar hem ilgili hem de çeşitli makalelere erişmektedir. Kesme noktası arttıkça MMR algoritmasının etkisi de azalmaktadır. Bu noktada MMR algoritmasının bir yeniden sıralama algoritması olduğu ve sıralamalardaki ilk 50'şer makale üzerinde işletildiği unutulmamalıdır. Dolayısıyla kesme noktası arttıkça MMR algoritmasının etkisinin giderek azalması ve kesme noktası 50'de hiç etkisinin kalmaması beklenen bir durumdur.

Konu modelleme algoritmasının performansını etkileyen en önemli kriterlerden birisi olan konu sayısı derlem büyüklüğü ile doğrudan ilişkilidir. Derlem büyüklüğü arttıkça algoritmaya girdi olarak verilmesi gereken en uygun konu sayısı da artmaktadır (Griffiths ve Steyvers, 2004). Konu sayısı birkaç düzineyi aştığında ise LDA algoritması daha az başarılı olmakta ve konularda tutarsızlıklar gözlenmektedir (Hecking ve Leydesdorff, 2018). Bu bulgu LDA ile tutarlı konular oluşturmak ve güvenilir istatistikler sağlamak için büyük miktarda veriye (1000 ve üzeri makale) ihtiyaç duyulduğu yönündeki bulgularla çelişmektedir

(Leydesdorff ve Nerghes, 2017). Fakat teoride konu sayısının fazla olması, konuların ayrıntı düzeyini de artırdığı için tutarsızlık sorunlarının ortaya çıkması normaldir. Bu çalışma kapsamında ihtiyaç duyulan bibliyometrik veriler de atıf dizinlerinde yer aldığından büyük derlemler söz konusudur. Büyük derlemlerde konu sayısı da fazla olacağı için tutarsızlık sorunları yaşanacağı açıktır. Bu araştırmada iSearch derlemi üzerinde yaptığımız uygulamalardan elde ettiğimiz bulgular Hecking ve Leydesdorff'un (2018) bulguları ile örtüşmektedir. LDA için en uygun konu sayısı (130) belirlendikten sonra, LDA algoritması çalıştırıldığında ilgi sıralamasında ilk sıralarda erişilen makalelerin çok çeşitli konularda olduğu saptanmıştır. Burada söz edilen konular arXiv konu başlıkları olup, LDA algoritması tarafından denetimsiz olarak oluşturulan konulardan bağımsızdır. Ardından pennant erişim algoritması ile erişim çıktıları daha tutarlı hale getirilmiştir. Diğer bir deyişle LDA algoritmasının performansı artırımlı olarak iyileştirilmiştir (incremental improvement).

Pennant erişim algoritması çekirdek makalenin literatüre olan etkisini gözlemeye ve çekirdek makale ile ilgili olarak en etkili araştırmaları ortaya çıkarmaya da imkân vermektedir. Bu çalışmada derlem fizik makalelerinden oluştuğu için etki ile ilgili uzman yorumu yapılamamıştır. Bunun yerine Danilov'un "Experimental Review on Pentaquarks" başlıklı derleme makalesi (Danilov, 2005) üzerinde pennant erişim algoritması çalıştırılmış ve 202 ilgili makaleye erişilmiştir. Danilov'un makalesine arXiv'deki kaynaklardan 13 kez atıf yapılmıştır. Makalenin kaynakçasında 49 referans vardır. Ancak bu referansların sadece 38'i iSearch derleminde "dâhili referans" olarak yer almaktadır. Pennant erişim algoritması 38 makaleden 37'sine (%97), LDA algoritması ise 15'ine (%39) erişmiştir. Diğer bir deyişle 38 makaleyi temel kaynak olarak kabul ettiğimizde, pennant erişim ile artırımlı olarak geliştirilen sıralamada LDA'nın kaçırdığı 22 makaleye (kaynak makalelerin %58'i) erişilmiştir. Benzer bir biçimde Maron ve Kuhns'un (1960) çalışmasına uygulanan pennant erişim çıktısı, Maron ve Kuhns'un çalışmasının bilgi erişim literatürünü nasıl etkilediğini inceleyen Thompson'ın (2007) kaynakçası ile karşılaştırmış ve %82'lik bir çakışma saptanmıştır (Akbulut, 2016). Bu bulgu, çalışmamız kapsamında elde edilen bulgularla benzerlik göstermektedir.

Ortak atıfa dayalı pennant erişim algoritmasının derlemdeki en ilgili makalelere eriştiği yönündeki bulgular, atıfların önemli ve etkili çalışmaları bulmada çok önemli rol oynadığının saptandığı daha önceki araştırmaların bulgularını desteklemektedir (Huang ve diğerleri, 2016; Huang ve diğerleri, 2018; Li ve diğerleri, 2017; Zhou ve diğerleri, 2017). Öte yandan, pennant erişim algoritması ile elde edilen bağlamların otomatik olarak etiketlenmesi, görselleştirmelerin kapsamlılığını büyük ölçüde artırabilir (Chang ve diğerleri, 2009; Nolasco ve Oliveira, 2016; Rüdiger ve diğerleri, 2021). Görselleştirme için pennant erişim diyagramlarından yararlanılması kullanıcıların literatürü izlemelerini kolaylaştırılabilir (Akbulut ve diğerleri, 2020). Bunun dışında yayın yılı bilgileri de hesaplamalara dâhil edilerek örneğin popülerlik ile ilgili özellikler de eklenebilir.

Bu araştırma olasılıksal konu modellemesi ile oluşturulan ilgi sıralamalarının atıflara dayanan pennant erişim yöntemiyle artırımlı olarak iyileştirilmesine yönelik olarak yapılan bildiğimiz kadarıyla ilk çalışmadır. Çalışmanın bulguları, LDA algoritması ile pennant erişim yönteminin bütünleşik olarak kullanıldığında, ilgi ve çeşitlilik oranı maksimize edilmiş erişim çıktıları elde edilebileceğini göstermektedir. Çalışmanın en önemli sınırlılığı ise sadece fizik makalelerinin bulunduğu bir derlem üzerinde yapılmış olmasıdır. Oysaki atıf örüntüleri, özet uzunlukları, yazar sayısı gibi etmenler bilimsel alanlara göre farklılık göstermektedir (Samraj, 2005; Liu ve Fang, 2020). Bu yüzden aynı yöntemin farklı alanlarda uygulanması; atıf bilgileri, tam metin ve özet bilgileri, konu sınıfları ve uzman değerlendirmesi (sorgu-makale ilgisi) içeren derlemler üzerinde çalışılması gerekmektedir.

Araştırmanın bulguları önemli ölçüde MMR algoritmasının farklı sıralama algoritmaları üzerindeki etkileri incelenerek yorumlanmıştır. Farklı sıralama listelerinin ilgi ve çeşitlilik oranlarının fizikçiler tarafından yorumlandığı benzer çalışmaların yapılmasında fayda vardır.

Sonuç ve Öneriler

Araştırma sonuçlarına göre, kelimeler arası ilişkilere odaklanan LDA konu modelleme algoritması ile oluşturulan ilgi sıralamalarında ilk sıralarda çoğunlukla marjinal ilgili belgeler yer almaktadır. Öte yandan, ortak atıf analizine dayanan pennant erişim algoritması makalelerin bağlamı ve ilgi düzeyi hakkında önemli ipuçları sağladığı için, ilgi sıralamalarının başlarında tutarlı bir şekilde benzer konudaki ve sorguyla yakından ilgili çalışmalar yer almaktadır. Pennant erişim ile sonradan eklenen marjinal ilgili makaleler genellikle daha az ortak atıfı olan ve başka disiplinlerle ilişki kurulmasını sağlayan makalelerdir. Ama bu tarz çalışmalar arama yapılan konu ya da terimle hâlâ ilgilidir. Dolayısıyla LDA algoritması pennant erişim algoritması ile desteklendiğinde hem sorguyla en ilgili makalelerin hem de seyrek ortak atıf yapılan marjinal makalelerin yer aldığı ilgi sıralamaları elde edilmiştir. Başka bir deyişle erişim çıktısında hem arama yapılan konunun sınırları genişlemiş hem de erişilen makalelerin sorguyla ilgi oranları artmıştır (bkz. Şekil 10 ve 11).

Önerilen sıralama yönteminde tümleştirme aşamasında her iki algoritmadan elde edilen skorlar normalize edilerek kullanılmaktadır. Dolayısıyla bir sorguya karşılık erişilen makalelerde kelime sıklıklarının olasılıksal dağılımlarında baskın bir örüntü varsa tümleştirilmiş listede "LDA baskın", ortak atıf yoğunluğu yüksek bir örüntü varsa "pennant baskın" bir yapı gözlenmektedir. Bu açıdan önerilen yöntem kelime sıklığı ve atıf bilgilerini orijinal yapıyı koruyacak şekilde bütünleştirmektedir. Önerilen yöntemde sistematik birleştirme yapısı olmadığı için bu yöntem benzer amacı olan diğer algoritmalardan (örneğin, MMR) ayrılmaktadır.

Bulgular makalelerin başlık ve özet bilgileri ile bibliyometrik bilgilerin bir arada kullanılarak alternatif ve kişiselleştirilebilir özellikte ilgi sıralamaları oluşturulabileceğini göstermektedir. Çalışma kapsamında konu modellemesi makalelerin sadece özet ve başlık bölümlerine uygulandığı için süreç hızlandırılmış, seyrek bir atıf ağında bile istenen özellikte listeler elde edilebilmiştir. Bu sonuçlar önerilen yöntemin atıf veri tabanlarına entegre edilerek literatür taramaları için uygun listeler elde edilebileceğini göstermektedir.

Çalışma kapsamında sadece ilgi sıralamalarının daha da geliştirilmesi üzerine odaklanılmıştır. Bunun ötesinde önerilen yöntem için farklı senaryolarla geçerlik ve uygunluk değerlendirmeleri yapılmasında fayda vardır. Algoritmaların değerlendirmesi ile ilgili en önemli kavramlar hesaplama (computation), dinamizm (dynamism), sağlamlık (robustness), tekrarlanabilirlik (replicability) ve ölçeklenebilirlik (scalability) olarak sıralanabilir (Ballester ve Penner, 2022). LDA da dahil olmak üzere metin işleme algoritmalarında hesaplama açısından "hesaplama zamanı" ve "bellek alanı" olmak üzere iki önemli zorluk bulunmaktadır. Problemin boyutu büyüdükçe bunu çözmek için gerekli hesaplama zorluğu da çok hızlı olarak artmaktadır. İşlemin üstelliği problemini aşmak için sorunu basite indirgeyerek daha az hesaplama gücüyle çözebilmek ve işlem sonuçlarını sonradan da kullanabilmek için daha az bellek gerektiren bir ortamda saklamak önemlidir (Mahajan ve diğerleri, 1999). Bu açıdan, önerilen yöntemde LDA'nın tam metin yerine özet ve başlıklara uygulanması hesaplama süresini azaltmıştır. Ayrıca her iki algoritmanın ilgili belgelerin farklı özelliklerini ön plana çıkardığı gösterilmiştir. Dolayısıyla yeniden sıralama açısından tek tek makaleler için hesaplama yapılması yerine, bir kez sıralama değerleri hesaplandıktan sonra sadece iki algoritmadan birine ağırlık verilerek istenen yapıda ilgi sıralamaları oluşturulabilir.

Önerilen ilgi sıralaması yöntemini durağan bir derlem (iSearch) üzerinde test etmemize karşın, bu yöntemin zamanla dinamik yapıdaki atıf dizinlerinde de kullanılabileceği

kanısındayız. WoS gibi büyük ölçekli atıf dizinlerinde sürekli yeni makale girişi olduğu için ilgi değerlerinin atıf ilişkileri dikkate alınarak anlık olarak hesaplanması söz konusudur. LDA algoritmasında yeni gelen çalışmalar için yeniden hesaplama işlemi ortak atıf verilerinin anlık olarak hesaplanmasından daha da maliyetli olabilir. Önerilen yöntemin dinamik bir yapıda nasıl işleyeceğinin test edilmesi hem dinamizm hem de hesaplama yükü açısından önemlidir.

Sağlamlık konusunda algoritmalara girdi olarak kullanılan derlemlerde kasıtlı eksiklikler ve farklılıklar yaratılarak farklı senaryolar ile deneyler yapılabilir. Örneğin, çok dilli ya da atıf yoğunluğu farklı bir derlem üzerinde hangi algoritmanın daha iyi işlediği, önerilen yöntemin başarılı bir biçimde çalışıp çalışmadığı test edilebilir. Ölçeklenebilirlik ve tekrarlanabilirlik konularında ise iSearch derleminden farklı örneklemler seçilerek algoritmalar tekrar çalıştırılabilir ve önerilen yöntemin farklı büyüklükteki derlemler üzerinde benzer performans gösterip göstermediği test edilebilir.

Ek 1. LDA Konu Modelleme Algoritmasında Konu Sayısını Belirlemek İçin Kullanılan Ölçevler

LDA konu modelleme algoritmasını çalıştırmadan önce konu sayısının parametre olarak girilmesi gerekir. Bu amaçla araştırmada dört farklı ölçev kullanılmıştır. Bu ölçevlerden ilkinde LDA'nın konu-terim ve belge-terim matrisi çıktılarından oluşturulan dağılımlara bakılarak uygun konu sayısı belirlenir (Arun ve diğerleri, 2010). Uygun konu sayısına ulaşıldığında varyansın önemli ölçüde azaldığından hareketle dağılımlar simetrik Kullback-Leibler ıraksaklığı (KL divergence)[18] cinsinden hesaplanır. Burada LDA algoritması Belge-Terim Sıklık Matrisi *M*'yi $T * W$ düzeyinde bir konu-terim (Topic-Word) matrisi M_1'e ve Belge-Konu (Document-Topic) matrisi M_2'ye bölen negatif olmayan bir matris çarpanlara ayırma mekanizması olarak işlev görür. M_1 matrisinin tek değer dağılımlarının (singular value distribution) simetrik KL ıraksaklığı ve $L * M_2$ vektörünün dağılımı hesaplanır. Derlemdeki (C) belge sayısı d, terim (kelime) dağarcığının boyutu ise w ile temsil edilmektedir (Ek 1, Formül 1).

$$C_{d*w} = M1_{d*t} \times Q_{t*w} \quad (1)$$

Bölmenin kalitesi seçilen optimum konu sayısına (t) bağlıdır. Ölçev, bu matris faktörlerinden türetilen dağılımların simetrik KL ıraksaklığı cinsinden hesaplanır. Optimum olmayan sayıda konu için sapma değerleri daha yüksektir (Arun ve diğerleri, 2010, s. 391-392).

İkinci ölçevde de benzer biçimde konular arasındaki ortalama kosinüs mesafesi minimuma ulaştığında LDA modelinin en iyi performansı gösterdiği varsayılmaktadır (Cao ve diğerleri, 2009). Konular (T_i, T_j) arasındaki mesafe konular üzerine kelime ataması yapılmasının anlamlı olup olmadığı hakkında bilgi vermektedir. Konular arasındaki korelasyonu ölçmek için standart kosinüs mesafesi kullanılmaktadır (Ek 1, Formül 2) (Cao ve diğerleri, 2009, s. 1778).

$$korelasyon(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \quad \sqrt{\sum_{v=0}^{V} (T_{jv})^2}} \quad (2)$$

Formül 2'de $korelasyon(T_i, T_j)$ değeri küçüldükçe konular da birbirinden daha bağımsız ve ayrık olur. Konu yapısının kararlılığını ölçmek için her konu çifti arasındaki ortalama kosinüs mesafesi kullanılır. İlk iki ölçev için skorlar minimum olduğunda ilgili derlem için en uygun konu sayısına ulaşılır (Holliger, 2018).

Üçüncü ölçev olan gizli kavram modellemede [Latent Concept Modeling (LCM)] konular arasındaki farklılık değerlerinin maksimuma ulaşması esastır (Deveaud ve diğerleri, 2014). Bu yüzden konu çiftleri arasındaki uzaklık maksimize edilir. LDA'nın konularının en yüksek olasılıklı n kelimeden oluştuğunu göz önünde bulundurarak, verilen bir fonksiyon için en büyük n değeri elde eden top$-n$ argümanı üreten bir $\mathrm{argmax}[n]$ operatörü tanımlanır (Ek 1, Formül 3). Bu operatör kullanılarak, k konusunda $P_{TM}(w|k) = \phi_{k,w}$ olasılığı en yüksek olan n kelimenin W_k kümesi elde edilir (Deveaud ve diğerleri, 2014, s. 66-67).

$$W_k = \underset{w}{argmax}[n]\, \phi_{k,w} \quad (3)$$

[18] Kullback-Leibler ıraksaklığı iki olasılık dağılımı arasındaki farkı ölçmektedir.

LDA konularının tüm çiftleri (k_i, k_j) arasındaki bilgi sapmasını (D) maksimize ederek sorgunun gizli kavramlarının sayısı tahmin edilir. Tahmin edilen $\widehat{K}$ kavramlarının sayısı aşağıdaki Ek 1, Formül 4'e göre hesaplanır (Deveaud ve diğerleri, 2014, s. 67)

$$\widehat{K} = \underset{K}{argmax} \frac{1}{K(K-1)} \sum_{(k,k^i)\in\mathbb{T}_K} D(k\|k') \tag{4}$$

Formül 4'te LDA'ya girdi olarak verilen konu sayısı ve $\mathbb{T}_K$, LDA tarafından modellenen K konu kümesidir. Başka bir deyişle, $\widehat{K}$, LDA'nın en dağınık konuları modellediği konu sayısıdır.

Dördüncü ölçevde ise konu sayısı çıkarımı için Markov zinciri Monte Carlo algoritması ve Bayes modeli birlikte kullanılır (Griffiths ve Steyvers, 2004). Ölçevde kelimelerin konulara atanmasında sonsal dağılım (posterior distribution) dikkate alınır ve tahminleme yapılır. Bu süreçte yinelenen rastgele örneklemeyi kullanan performansı yüksek ve hızlı bir hesaplama türü olan Monte Carlo algoritması tercih edilmiştir. Bu ölçevin teknik ayrıntıları için bkz. Griffiths ve Steyvers (2004).

LDA konu modelleme algoritmasına girilecek konu sayısını belirlemek için kullanılan ve yukarıda kısaca açıklanan dört ölçev ve her ölçev için verilen formüller kullanılarak bu araştırmada konu sayısının nasıl hesaplandığı R ile yazılan kodlarda daha ayrıntılı olarak verilmektedir.[19]

İzin ve Katkı Bildirimleri

Etik Kurul İzni: *Çalışma etik kurul kararı gerektirmemektedir.*

Yazarlık Katkısı: *Yazarlar katkılarını şu şekilde beyan etmişlerdir.*

M. Akbulut: Kavramsallaştırma, yöntem, veri toplama, veri analizi, görselleştirme, yazılım geliştirme ve geçerleme; özgün ve gözden geçirilmiş taslakların yazılması ve düzenlenmesi

Y. Tonta: Kavramsallaştırma, yöntem, taslakları gözden geçirme ve düzeltme, son taslak, denetleme

Kaynakça

ADS Team (2008). SAO/NASA ADS Abstract Service Stopword List. https://adsabs.harvard.edu/abs_doc/stopwords.html

Abramo, G., D'Angelo, C. A. ve Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics*, *12*(4), 1182-1193. https://doi.org/10.1016/j.joi.2018.09.001

Adomavicius, G. ve Kwon, Y. (2011). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, *24*(5), 896-911. https://doi.org/10.1109/TKDE.2011.15

Akbulut, M. (2016). *Atıf klasiklerinin etkisinin ve ilgililik sıralamalarının pennant diyagramları ile analizi* [Yayımlanmamış Yüksek lisans tezi] Hacettepe Üniversitesi. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/3529

[19] Bkz. https://www.mugeakbulut.com/phd/codes/LDA_Konu_sayisi_belirleme/code.R.

Akbulut, M., Tonta, Y. ve White, H. D. (2020). Related records retrieval and pennant retrieval: An exploratory case study. *Scientometrics, 122*(2), 957-987. https://doi.org/10.1007/s11192-019-03303-9

Alpaydın, E. (2017). *Yapay öğrenme*. Boğaziçi Üniversitesi Yayınevi.

Arun, R., Suresh, V., Madhavan, C. V. ve Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* içinde (s. 391-402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Baeza-Yates, R. ve Ribeiro-Neto, B. (1999). *Modern information retrieval.* ACM Press.

Ballester, O. ve Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics, 16*(1). https://doi.org/10.1016/j.joi.2021.101224

Bayer, D. ve Michael, S. (2019). *Exploring the Daschle Collection using text mining.* arXiv. https://arxiv.org/pdf/1904.12623.pdf

Beel, J. ve Gipp, B. (2009). Google Scholar's ranking algorithm: An introductory overview. B. Larsen ve J. Leta (Yay. haz.). *Proceedings of the 12th International Conference on Scientometrics and Informetrics* içinde (s. 230-241). International Society for Scientometrics and Informetrics. https://www.issi-society.org/proceedings/issi_2009/ISSI2009-proc-vol1_Aug2009_batch2-paper-1.pdf

Beel, J., Gipp, B., Langer, S. ve Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, *17*(4), 305-338. https://doi.org/10.1007/s00799-015-0156-0

Belter, C. W. (2017). A relevance ranking method for citation-based search results. *Scientometrics*, *112*(2), 731-746. https://doi.org/10.1007/s11192-017-2406-y

Bichteler, J. ve Eaton III, E. A. (1980). The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American Society for Information Science*, *31*(4), 278–282. https://doi.org/10.1002/asi.4630310408

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://dl.acm.org/doi/pdf/10.1145/2133806.2133826

Blei, D. M. ve Lafferty, J. D. (2009). Topic models. A. Srivastava ve M. Sahami (Yay. haz.). *Text Mining: Classification, Clustering and Applications* içinde (s. 71-94). CRC Press, Taylor & Francis. http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf

Blei, D. M., Ng, A. Y. ve Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8

Bonaccorsi, A., Melluso, N. ve Massucci, F. A (2022). Exploring the antecedents of interdisciplinarity at the European Research Council: A topic modeling approach. *Scientometrics*. https://doi.org/10.1007/s11192-022-04368-9

Bornmann, L., Haunschild, R. ve Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications, 8*(1), 1-15. https://doi.org/10.1057/s41599-021-00903-w

Boyd-Graber, J. ve Blei, D. M. (2010). *Syntactic topic models.* arXiv. https://arxiv.org/pdf/1002.4665.pdf

Bradley, K. ve Smyth, B. (2001). Improving recommendation diversity. D. O'Donoghue (Yay. haz.). *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science* içinde (s. 141-152). NUIM Department of Computer Science. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.5232&rep=rep1&type=pdf

Cambria, E. ve White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48-57. https://doi.org/10.1109/MCI.2014.2307227

Cao, J., Xia, T., Li, J., Zhang, Y. ve Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781. https://doi.org/10.1016/j.neucom.2008.06.011

Carbonell, J. ve Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 335-336). Association for Computing Machinery. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.3982&rep=rep1&type=pdf

Carevic, Z. ve Mayr, P. (2014). *Recommender systems using pennant diagrams in digital libraries.* arXiv. https://arxiv.org/pdf/1407.7276v1.pdf

Carevic, Z. ve Schaer, P. (2014). On the connection between citation-based and topical relevance ranking: Results of a pretest using iSearch. *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014)* içinde (s. 37-44). Springer-Verlag. https://ceur-ws.org/Vol-1143/paper5.pdf

Carroll, M. (2018). *Changes in media coverage of GCSEs from 1988 to 2017.* Cambridge. https://www.cambridgeassessment.org.uk/Images/504456-changes-in-media-coverage-of-gcses-from-1988-to-2017.pdf

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. ve Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* içinde (s. 288-296). MIT Press. https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Chen, M. ve Décary, M. (2018). A cognitive-based semantic approach to deep content analysis in search engines. *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* içinde (s. 131-139). IEEE. https://doi.ieeecomputersociety.org/10.1109/ICSC.2018.00027

Chen, Z. ve Liu, B. (2014). Mining topics in documents: Standing on the shoulders of big data. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* içinde (s. 1116-1125). ACM. https://dl.acm.org/doi/pdf/10.1145/2623330.2623622

Cooper, W. S. (1988). Getting beyond boole. *Information Processing & Management*, *24*(3), 243-248. https://doi.org/10.1016/0306-4573(88)90091-X

Croft, W. B. (2002). Combining approaches to information retrieval. W.B. Croft (Yay. haz.). *Advances in Information Retrieval. The Information Retrieval Series, Vol 7*. içinde (s. 1-35). Springer, https://doi.org/10.1007/0-306-47019-5_1

Crossley, S., Dascalu, M. ve McNamara, D. (2017). How important is size? An investigation of corpus size and meaning in both latent semantic analysis and latent dirichlet allocation. *The Thirtieth International Flairs Conference* içinde (s. 293-296). The AAAI Press. https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15441/14942

Danilov, M. (2005). *Experimental review on pentaquarks.* arXiv. https://arxiv.org/abs/hep-ex/0509012

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. ve Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Deveaud, R., SanJuan, E. ve Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61-84. https://doi.org/10.3166/dn.17.1.61-84

Ekinci, E. ve İlhan Omurca, S. (2020). Concept-LDA: Incorporating Babelfy into LDA for aspect extraction. *Journal of Information Science*, *46*(3), 406-418. https://doi.org/10.1177/0165551519845854

Ganguly, D. ve Jones, G. J. (2018). A non-parametric topical relevance model. *Information Retrieval Journal*, *21*(5), 449-479. https://doi.org/10.1007/s10791-018-9329-y

George, C. P. ve Doss, H. (2017). Principled selection of hyperparameters in the Latent Dirichlet Allocation model. *Journal of Machine Learning Research, 18*(1), 5937-5974. https://www.jmlr.org/papers/volume18/15-595/15-595.pdf

Giustolisi, O., Ridolfi, L. ve Simone, A. (2020). Embedding the intrinsic relevance of vertices in network analysis: The case of centrality metrics. *Scientific Reports*, *10*(3297). https://doi.org/10.1038/s41598-020-60151-x

Gläser, J., Glänzel, W. ve Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981-998. https://doi.org/10.1007/s11192-017-2296-z

Griffiths, T. L. ve Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Guillemette, J., Simms, B., Zhou, D. ve Mills, S. (2017). Applying latent dirichlet allocation to yelp reviews. https://people.math.carleton.ca/~smills/2017-18/STAT4601-5703/Research%20Projects/2018%20Submissions/GuillemetteSimmsZhouD/Applying%20LDA.pdf

Guo, J., Fan, Y., Ai, Q. ve Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* içinde (s. 55-64). ACM. https://doi.org/10.1145/2983323.2983769

Guo, Z., Zhang, Z. M., Zhu, S., Chi, Y. ve Gong, Y. (2013). A two-level topic model towards knowledge discovery from citation networks. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 780-794. https://doi.org/10.1109/TKDE.2013.56

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent dirichlet allocation topic model. *Scientometrics*, *125*(3), 2561-2595. https://doi.org/10.1007/s11192-020-03721-0

Hecking, T. ve Leydesdorff, L. (2018). *Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps.* arXiv. https://arxiv.org/pdf/1806.01045.pdf

Herlocker, J. L., Konstan, J. A., Terveen, L. G. ve Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5-53. https://doi.org/10.1145/963770.963772

Holliger, T. S. (2018). *Strategic sourcing via category management: Helping air force installation contracting agency eat one piece of the elephant* (Yüksek lisans tezi, Air Force Institute of Technology). https://apps.dtic.mil/sti/pdfs/AD1056353.pdf

Huang, L., Liu, H., He, J. ve Du, X. (2016). Finding latest influential research papers through modeling two views of citation links. F. Li, K. Shim, K. Zheng ve G. Liu (Yay. haz.). *Web Technologies and Applications APWeb 2016* içinde (s. 555-566). Springer, Cham. https://doi.org/10.1007/978-3-319-45814-4_45

Huang, X., Chen, C., Peng, C., Wu, X., Fu, L. ve Wang, X. (2018). Topic-sensitive influential paper discovery in citation network. D. Phung, V. Tseng, G. Webb, B. Ho, M. Ganji ve L. Rashidi (Yay. haz.). *Advances in Knowledge Discovery and Data Mining* içinde (s. 16-28). Springer, Cham. https://doi.org/10.1007/978-3-319-93037-4_2

Jin, R., Valizadegan, H. ve Li, H. (2008). Ranking refinement and its application to information retrieval. *Proceedings of the 17th International Conference on World Wide Web* içinde (s. 397-406). ACM. http://doi.org/10.1145/1367497.1367552

Ke, Q., Ferrara, E., Radicchi, F. ve Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426-7431. https://doi.org/10.1073/pnas.1424329112

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10-25. https://doi.org/10.1002/asi.5090140103

Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N. ve Bayer, V. (2017). *Towards effective research recommender systems for repositories*. arXiv. https://arxiv.org/abs/1705.00578

Kucuktunc, O. ve Ferhatosmanoglu, H. (2011). λ-diverse nearest neighbors browsing for multidimensional data. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 481-493. httpls://doi.org/10.1109/TKDE.2011.251

Küçüktunç, O., Saule, E., Kaya, K. ve Çatalyürek, Ü. V. (2015). Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology*, *5*(4), 1-21. https://doi.org/10.1145/2668106

Lei, M., Wang, J., Chen, B. ve Li, X. (2001). Improved relevance ranking in WebGather. *Journal of Computer Science and Technology*, *16*(5), 410-417. https://doi.org/10.1007/bf02948958

Leydesdorff, L. ve Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N< 1,000). *Journal of the Association for Information Science and Technology, 68*(4), 1024-1035. https://doi.org/10.1002/asi.23740

Li, C., Feng, H. ve Rijke, M. D. (2020). Cascading hybrid bandits: Online learning to rank for relevance and diversity. *Fourteenth ACM Conference on Recommender Systems* içinde (s. 33-42). ACM. https://doi.org/10.1145/3383313.3412245

Li, W. ve McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd International Conference on Machine Learning* içinde (s. 577-584). Springer. https://doi.org/10.1145/1143844.1143917

Li, Y., He, J. ve Liu, H. (2017). Topic analysis and influential paper discovery on scientific publications. *2017 14th Web Information Systems and Applications Conference (WISA)* içinde (s. 68-73). IEEE. https://doi.org/10.1109/WISA.2017.69

Liu, X., Wang, G. ve Bhuiyan, M. Z. A. (2022). Re-ranking with multiple objective optimization in recommender system. *Transactions on Emerging Telecommunications Technologies*, *33*(1): e4398. https://doi.org/10.1002/ett.4398

Liu, X. Z. ve Fang, H. (2020). A comparison among citation-based journal indicators and their relative changes with time. *Journal of Informetrics, 14*(1), 1-17. https://doi.org/10.1016/j.joi.2020.101007

Lykke, M., Larsen, B., Lund, H. ve Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. *European Conference on Information Retrieval* içinde (s. 627-630). Springer. https://doi.org/10.1007/978-3-642-12275-0_63

Ma, Z., Liu, Y., Yang, Z., Yang, J. ve Li, K. (2022). A parameter-free approach to lossless summarization of fully dynamic graphs. *Information Sciences*, *589*, 376-394. https://doi.org/10.1016/j.ins.2021.12.116

McNee, S. M., Riedl, J. ve Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. *CHI'06 extended abstracts on human factors in computing systems* içinde (s. 1097-1101). ACM. https://doi.org/10.1145/1125451.1125659

Mahajan, M., Beeferman, D. ve Huang, X. D. (1999). Improved topic-dependent language modeling using information retrieval techniques. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* içinde (s. 541-544). IEEE. https://doi.org/10.1109/ICASSP.1999.758182

Manning, C. ve Schütze, H. (2000). *Foundations of statistical natural language processing.* MIT Press. https://ics.upjs.sk/~pero/web/documents/pillar/Manning_Schuetze_Statistical NLP.pdf

Maron, M. E. ve Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM, 7*(3), 216-244. https://doi.org/10.1145/321033.321035

Marujo, L., Ribeiro, R., Gershman, A., De Matos, D. M., Neto, J. P. ve Carbonell, J. (2017). Event-based summarization using a centrality-as-relevance model. *Knowledge and Information Systems*, *50*, 945–968. https://doi.org/10.1007/s10115-016-0966-4

Mayr, P. ve Mutschke, P. (2013). Bibliometric-enhanced retrieval models for big scholarly information systems. *2013 IEEE International Conference on Big Data* içinde (s. 5-8). IEEE. https://doi.org/10.1109/BigData.2013.6691762

Meng, W., Yu, C. ve Liu, K. L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys (CSUR)*, *34*(1), 48-89. https://doi.org/10.1145/505282.505284

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*, 810-832. https://doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U

Nguyen, T. ve Do, P. (2018). CitationLDA++: an extension of LDA for discovering topics in document network. *Proceedings of the Ninth International Symposium on Information and Communication Technology* içinde (s. 31-37). ACM. https://doi.org/10.1145/3287921.3287930

Nikita, M. (2020, 20 Nisan). Select number of topics for LDA. https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html

Nolasco, D. ve Oliveira, J. (2016). Detecting knowledge innovation through automatic topic labeling on scholar data. *2016 49th Hawaii International Conference on System Sciences (HICSS)* içinde (s. 358-367). IEEE. https://doi.org/10.1109/HICSS.2016.51

Pao, M. L. (1993). Term and citation retrieval: A field study. *Information Processing & Management. 29*(1), 95-112. https://doi.org/10.1016/0306-4573(93)90026-A

Pathik, N. ve Shukla, P. (2020). Simulated annealing based algorithm for tuning LDA hyper parameters. M. Pant, T. Kumar Sharma, R. Arya, Sahana B., H. Zolfagharinia (Yay. haz.). *Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing*, vol 1154 içinde (s. 515-521). Springer. https://doi.org/10.1007/978-981-15-4032-5_47

Ponweiser, M. (2012). *Latent dirichlet allocation in R.* (Yüksek lisans tezi, Viyana Üniversitesi). https://epub.wu.ac.at/id/eprint/3558

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. ve Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262-1282. https://doi.org/10.1016/j.respol.2012.03.015

Ribeiro, R., ve de Matos, D. M. (2011). Revisiting Centrality-as-relevance: Support sets and similarity as geometric proximity. *Journal of Artificial Intelligence Research*, *42*, 275-308. https://www.researchgate.net/publication/259764702_Centrality-as-Relevance_Support_Sets_and_Similarity_as_Geometric_Proximity

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, *33*(4), 294-304. https://doi.org/10.1108/eb026647

Rüdiger, M. S., Antons, D. ve Salge, T. O. (2021). The explanatory power of citations: A new approach to unpacking impact in science. *Scientometrics, 126*, 9779-9809. https://doi.org/10.1007/s11192-021-04103-w

Salton, G., Yang, C. ve Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613-620. https://doi.org/10.1145/361219.361220

Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, *24*(2), 141-156. https://doi.org/10.1016/j.esp.2002.10.001

Saracevic, T. (2021). Relevance: In search of a theoretical foundation. D. H. Sonnenwald (Yay. haz.). *Theory Development in the Information Sciences* içinde (s. 141-163). University of Texas Press. https://doi.org/10.7560/308240-011

Sperber, D. ve Wilson, D. (1995). *Relevance: Communication and cognition*. Blackwell. https://monoskop.org/images/e/e6/Sperber_Dan_Wilson_Deirdre_Relevance_Communica_and_Cognition_2nd_edition_1996.pdf

Swanson, D. R. (1986a). Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, *56*(4), 389-398. https://doi.org/10.1086/601800

Swanson, D. R. (1986b). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1):7-18. https://doi.org/10.1353/pbm.1986.0087

Thara, D. K., PremaSudha, B. G. ve Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, *128*, 544-550. https://doi.org/10.1016/j.patrec.2019.10.029

Thompson, P. (2007). Looking back: On relevance, probabilistic indexing and information retrieval. *Information Processing & Management*, *44*(2), 963-970. https://doi.org/10.1016/j.ipm.2007.10.002

Tonta, Y. (1995). Bilgi erişim sistemleri. *Türk Kütüphaneciliği, 9*(3), 302-314. https://eprints.rclis.org/9571/

Tonta, Y. ve Akbulut, M. (2021). Uluslararası dergilerde yayımlanan Türkiye adresli makalelerin atıf etkisini artıran faktörler. *Türk Kütüphaneciliği*, *35*(3), 388-409. https://doi.org/10.24146/tk.933159

Vergoulis, T., Chatzopoulos, S., Kanellos, I., Deligiannis, P., Tryfonopoulos, C. ve Dalamagas, T. (2019). BIP! finder: Facilitating scientific literature search by exploiting impact-based ranking. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* içinde (s. 2937-2940). ACM. https://doi.org/10.1145/3357384.3357850

Verma, M., Yılmaz, E. ve Craswell, N. (2016). On obtaining effort based judgements for information retrieval. *Proceedings of the 9th ACM International Conference on Web Search and Data Mining* içinde (s. 277-286). ACM. https://doi.org/10.1145/2835776.2835840

Wallach, H., Mimno, D. ve McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems*, *22*. https://proceedings.neurips.cc/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf

Wang, X., Zhai, C. ve Roth, D. (2013). Understanding evolution of research themes: A probabilistic generative model for citations. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* içinde (s. 1115-1123). ACM. https://doi.org/10.1145/2487575.2487698

White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory. Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, *58*, 536-559. https://doi.org/10.1002/asi.20543

White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory. Part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology*, *58*, 583-605. https://doi.org/10.1002/asi.20542

White, H. D. (2009). Pennants for Strindberg and Persson. *Celebrating scholarly communication studies: A festschrift for Olle Persson at his 60th birthday.* Special volume of the *E-newsletter of the International Society for Scientometrics and Informetrics*, *5*, 71-83. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.2055&rep=rep1&type=pdf#page=73

White, H. D. (2010). Some new tests of relevance theory in information science. *Scientometrics*, *83*, 653-667. https://doi.org/10.1007/s11192-009-0138-3

White, H. D. (2015). Co-cited author retrieval and relevance theory: Examples from the humanities. *Scientometrics, 102*(3), 2275-2299. https://doi.org/10.1007/s11192-014-1483-4

White, H. D. (2016). Bag of works retrieval: TF*IDF weighting of co-cited works. *Proceedings of the 3rd Workshop on Bibliometric-Enhanced Information Retrieval (BIR2016)* içinde (s. 63-72). https://ceur-ws.org/Vol-1567/paper7.pdf

White, H. D. (2018). Bag of works retrieval: TF*IDF weighting of co-cited works with a seed. *International Journal of Digital Libraries, 19*, 139-149. https://doi.org/10.1007/s00799-017-0217-7

White, H. D. ve McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, *49*(4): 327-355. https://doi.org/10.1002/(SICI)1097-4571(19980401)49:4%3C327::AID-ASI4%3E3.0.CO;2-4

Wilson, D. ve Sperber, D. (2002). Relevance theory. G. Ward ve L. Horn (Yay. haz.). *Handbook of pragmatics* içinde (s. 1-55). Blackwell. https://jeannicod.ccsd.cnrs.fr/ijn_00000101/document

Wilson, P. (1978). Some fundamental concepts of information retrieval. *Drexel Library Quarterly, 14*(2), 10-24.

Wu, H. C., Luk, R. W., Wong, K. F. ve Kwok, K. L. (2007). A retrospective study of a hybrid document-context based retrieval model. *Information Processing & Management*, *43*(5), 1308-1331. https://doi.org/10.1016/j.ipm.2006.10.009

Wu, J., Son, G. ve Wang, S. (2020). A competency mining method based on Latent Dirichlet Allocation (LDA) model. *Journal of Physics: Conference Series (Vol. 1682, No. 1, p. 012059)* içinde (s. 1-6). IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/1682/1/012059/meta

Xia, H., Li, J., Tang, J. ve Moens M. F. (2012). Plink-LDA: Using link as prior information in topic modeling. S. Lee, Z. Peng, X. Zhou, Y. S. Moon, R. Unland ve J. Yoo (Yay. haz.). *Database Systems for Advanced Applications* içinde (s. 213-227). Springer. https://doi.org/10.1007/978-3-642-29038-1_17

Xie, X., Liang, Y., Li, X. ve Tan, W. (2019). CuLDA_CGS: Solving large-scale LDA problems on GPUs. *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming* içinde (s. 435-436). ACM. https://doi.org/10.1145/3293883.3301496

Yang, H. T., Ju, J. H., Wong, Y. T., Shmulevich, I. ve Chiang, J. H. (2017). Literature-based discovery of new candidates for drug repurposing. *Briefings in Bioinformatics*, *18*(3), 488-497. https://doi.org/10.1093/bib/bbw030

Yang, L., Ji, D. ve Leong, M. (2007). Document reranking by term distribution and maximal marginal relevance for Chinese information retrieval. *Information Processing & Management*, *43*(2), 315-326. https://doi.org/10.1016/j.ipm.2006.07.011

Yılmaz, E., Verma, M., Craswell, N., Radlinski, F. ve Bailey, P. (2014). Relevance and effort: An analysis of document utility. *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management* içinde (s. 91-100). ACM. https://doi.org/10.1145/2661829.2661953

Zarrinkalam, F. ve Kahani, M. (2012). A new metric for measuring relatedness of scientific papers based on non-textual features. *Intelligent Information Management, 4*(4), 99-107. https://www.scirp.org/pdf/IIM20120400001_98298896.pdf

Zhang, D., Luo, T., Wang, D. ve Liu, R. (2015). *Learning from LDA using deep neural networks.* arXiv. https://arxiv.org/pdf/1508.01011.pdf

Zhang, J., Zeng, J., Yuan, M., Rao, W. ve Yan, J. (2016). LDA revisited: Entropy, prior and convergence. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* içinde (s.1763-1772). ACM. https://dl.acm.org/doi/abs/10.1145/2983323.2983794

Zhou, H. K., Yu, H. M. ve Hu, R. (2017). Topic discovery and evolution in scientific literature based on content and citations. *Frontiers of Information Technology & Electronic Engineering*, *18*(10), 1511-1524. https://doi.org/10.1631/FITEE.1601125

Zou, L., Liu, X., Buntine, W. ve Liu, Y. (2021). Citation context-based topic models: Discovering cited and citing topics from full text. *Library Hi Tech*, *39*(4), 1063-1083. https://doi.org/10.1108/LHT-01-2021-0041