

## İlgi Sıralamalarının Artırımı 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 derlemedeki 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 artırımı olarak iyileştirilmesi üzerinde odaklandık. Bu yöntemin hesaplama yükü, sağlamlık, tekrarlanabilirlik ve ölçeklenebilirlik açılarından dinamik derlemler üzerinde sınılandıktan 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](mailto:mugeakbulut@gmail.com)

Ankara Yıldırım Beyazıt University, Department of Information Management, Ankara, Turkey. E-mail: [mugeakbulut@gmail.com](mailto:mugeakbulut@gmail.com)

\*\*\* Hacettepe Üniversitesi, Bilgi ve Belge Yönetimi Bölümü, Ankara, Türkiye. E-posta: [yasartonta@gmail.com](mailto:yasartonta@gmail.com)  
Hacettepe University, Department of Information Management, Ankara, Turkey. E-mail: [yasartonta@gmail.com](mailto:yasartonta@gmail.com)



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



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 algoritmalarından biri LDA'dır. Bu algoritmada, ilgiyi belirleyebilmek için derlemede 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.

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



tümleştirme (data fusion) yoluyla birleştirilmekte ve sürece dâhil olan tüm algoritmalarından 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<sup>3</sup> 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.<sup>4</sup> Daha sonra çeşitli makrolar yazılarak makalelerde atıf yapılan kaynakların her birinin temel konu kategorileri belirlenmiştir.<sup>5</sup> iSearch derlemindeki yayınların %98'inin temel konusu fiziktir.<sup>6</sup> 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

<sup>3</sup> 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.

<sup>4</sup> Kodlar için bkz. [https://mugeakbulut.com/phd/codes/iSearch\\_verilerini\\_Arxivden\\_indirme.py](https://mugeakbulut.com/phd/codes/iSearch_verilerini_Arxivden_indirme.py)

<sup>5</sup> Makaleler arXiv'de birden fazla konu altında listelenebilmektedir. Bu çalışma kapsamında ise her makale için sadece temel (primary category) konu kategorisi esas alınmıştır.

<sup>6</sup> Etkileşimli grafik için bkz. <https://mugeakbulut.com/phd/gorsellestirme/bubble.html>



### 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.

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

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  $\varphi$ , konuların makalede bulunma olasılığı ise  $\theta$  ile temsil edilmektedir. Dirichlet parametreleri de  $\alpha$  ve  $\beta$ 'dir. Konuların makalelerdeki dağılımını  $\alpha$ , kelimelerin konulardaki dağılımını ise  $\beta$  temsil eder (düşük  $\alpha$  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 ( $\theta$ ) makalelere dağılımı olasılığı ( $p$ ) hesaplanır. İkinci aşamada Dirichlet dağılımına göre kelimelerin ( $\varphi$ ) konulara dağılımı olasılığı belirlenir ( $\beta$ ). 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.



skorun yüksek olması, bazılarında ise düşük olması beklenmektedir (Carroll, 2018). iSearch derlemindeki yayınlar için konu sayısını belirlemek amacıyla dört ölçek (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 ölçekle ilgili ayrıntılı bilgi Ek 1’de verilmektedir.

LDA algoritması uygulanırken, özellik sayısı (yani sabit kelime boyutu)  $\alpha$  ve  $\beta$  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 çalışmada kullanılan derlem büyük ölçekli bir derlemidir. Algoritma çalıştırılırken orta düzey model için ön tanımlı parametreler kullanılmıştır.<sup>10</sup>

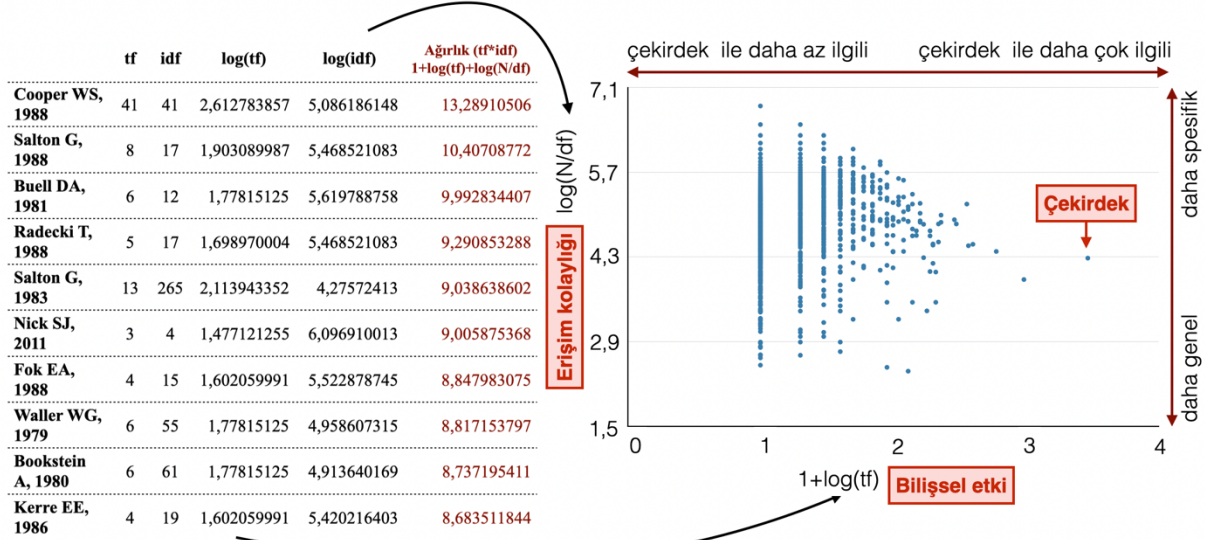
LDA algoritmasına girilecek konu sayısını belirlemek için kullanılan tüm ölçeklerde skorlar 0 ile 1 aralığındadır. Dört yaklaşımın bir arada gösterildiği Şekil 3’te “maksimize” olarak gruplanan ölçekler skorlarının yüksek olması beklenen, “minimize” olarak gruplananlar ise düşük olması beklenenlerdir. Tüm algoritmaların aynı optimal grup sayısını belirlemesi beklenemez, ancak ortak bölgeye -yüksek maksimize, düşük minimize- bakılarak en uygun konu sayısı belirlenebilir (Holliger, 2018). Ölçekler iSearch derlemine uygulandığında bu derlem için en uygun konu sayısının 110 ile 130 arasında olduğu anlaşılmaktadır (Şekil 3). Üçüncü ölçek hep düşme eğilimindedir. Diğer ölçeklerle uyumsuz olması ve bilgi verici bir örüntüye sahip olmaması sebebiyle konu sayısı belirlenirken bu ölçek göz ardı edilmiştir (Bayer ve Michael, Bonaccorsi, Melluso ve Massucci, 2022; 2019; Guillemette ve diğerleri, 2017; Holliger, 2018). Bu durum muhtemelen üçüncü ölçeğin 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.

<sup>10</sup> 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 çalışmada 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 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.

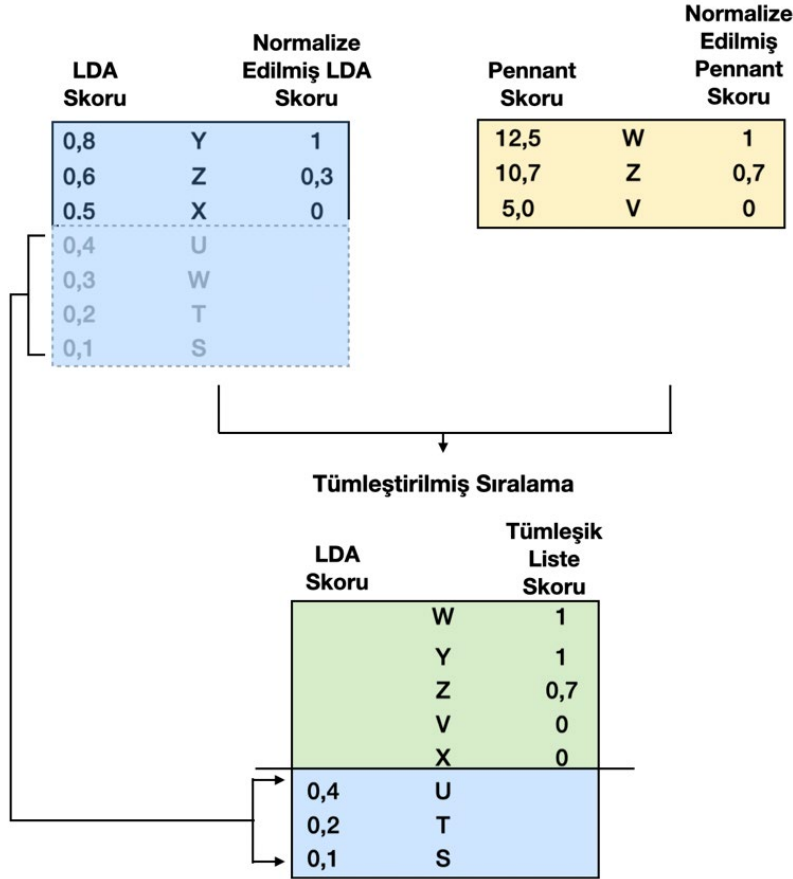
Pennant erişim algoritmasının uygulama aşamalarının örnek gösterimi Şekil 5'te verilmektedir. Elimizdeki derlemde 20 makale olduğunu (Makale A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S ve T) ve bu makalelerden O ve R'nin özetleri olmadığını varsayalım. Özeti bulunan tüm makalelere LDA algoritması uygulandıktan sonra, herhangi bir sorgu çalıştırıldığında LDA ilgi sıralaması elde edilmektedir. İlgi sıralamasındaki ilk beş makaleye atıf yapan makaleler kaynaklarıyla 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.



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

### Şekil 6

Sıralama tümleştirme



Normalizasyon işlemi Formül 4'e göre yapılmıştır. Herhangi bir değer 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). Derlemede 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.

<sup>13</sup>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).



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*

	<b>Tümleştirme Algoritması</b>	<b>MMR Algoritması</b>
Amaç	LDA ve pennant erişimde en yüksek ilgi değerleri alan belgeleri önceleyerek sıralamayı artırmalı olarak geliştirme (incremental refinement)	Marjinal ilgili belgeleri önceleyecek şekilde sıralamayı artırmalı 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	$\lambda$ (lambda) değeri

Bu araştırmada performans değerlendirme ve karşılaştırma (benchmarking) amacıyla MMR algoritmasından yararlanılmıştır.<sup>15</sup> 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

<sup>15</sup> Tümleştirilmiş sıralama listesine uygulanan MMR algoritması kodlarına <https://colab.research.google.com/drive/1dESqgDRL6WfyCSDgHxAa1kPD0HKBFK54> adresinden erişilebilir.



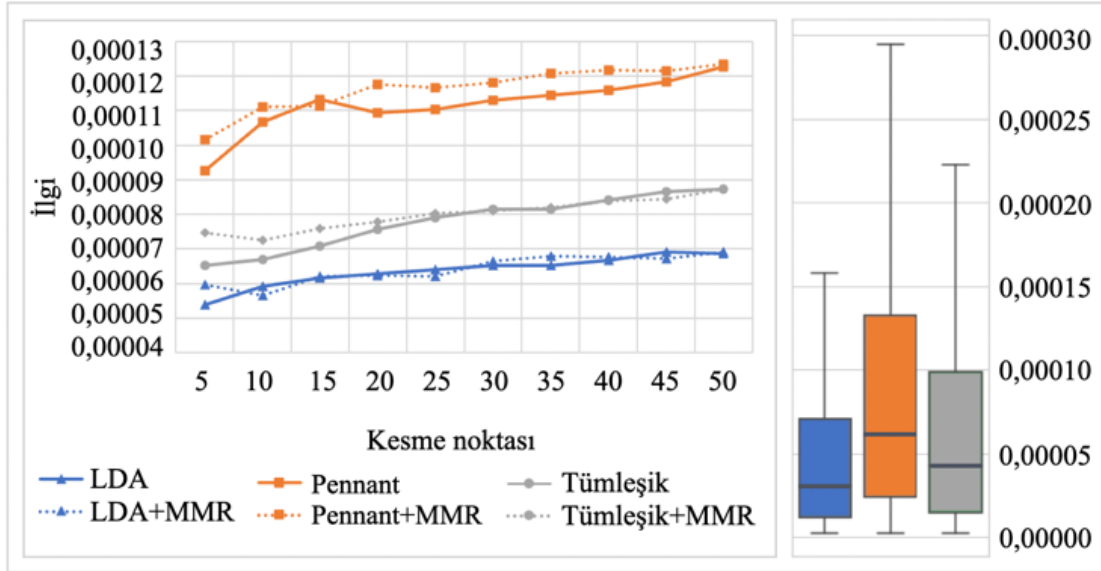
bir noktadan sonra sadece LDA kaynakları yer almaktadır. Örneğin, 65. sorguda pennant algoritması işletildiğinde sadece 22 makaleye erişilmektedir (dolayısıyla LDA çıktısı da 22'ye sabitlenmiştir).<sup>17</sup> Tümüleşik listede ise 39. sıradaki makaleden sonra sadece LDA listesinden gelen kaynaklar 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 (tümüleşik) elde edilmektedir.

### Şekil 10

Algoritmaların ilgi değerlerine göre karşılaştırılması (N=65)



<sup>17</sup> Bkz [https://mugeakbulut.com/phd/gorsellestirme/liste\\_konu\\_grafik\\_100.pdf](https://mugeakbulut.com/phd/gorsellestirme/liste_konu_grafik_100.pdf) (Sorgu no. 65).



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 belgelere 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 listelerinkine 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



(Leydesdorff ve Nerghes, 2017). Fakat teoride konu sayısının fazla olması, konuların ayrını 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 çalışmada 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 artırımı 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 artırımı 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 artırımı 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.



kanısındaız. WoS gibi büyük ölçekli atıf dizinlerinde sürekli yeni makale girişı olduđu için ilgi deđerlerinin atıf ilişkileri dikkate alınarak anlık olarak hesaplanması söz konusudur. LDA algoritmasında yeni gelen çalışmalar için yeniden hesaplama işlemleri 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.

Sađlamlık konusunda algoritmalara girdi olarak kullanılan derlemlerde kasıtlı eksiklikler ve farklılıklar yaratılarak farklı senaryolar ile deneyler yapılabilir. Örneđin, çok dilli ya da atıf yoğunluđu farklı bir derlem üzerinde hangi algoritmanın daha iyi işlediđi, önerilen yöntemin başarılı bir biçimde çalışıp çalışmadıđı test edilebilir. Ölçeklenebilirlik ve tekrarlanabilirlik konularında ise iSearch derleminden farklı örnekler seçilerek algoritmalar tekrar çalıştırılabilir ve önerilen yöntemin farklı büyüklükteki derlemler üzerinde benzer performans gösterip göstermediđi test edilebilir.



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

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{(k, k') \in \mathbb{T}_K} D(k \| k') \quad (4)$$

Formül 4'te LDA'ya girdi olarak verilen konu sayısı ve  $\mathbb{T}_K$ , LDA tarafından modellenen  $K$  konu kümesidir. Başka bir deyişle,  $\hat{K}$ , 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 örnekleme yöntemini 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.<sup>19</sup>

## 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 geçişleme; 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](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>

<sup>19</sup> Bkz. [https://www.mugeakbulut.com/phd/codes/LDA\\_Konu\\_sayisi\\_belirleme/code.R](https://www.mugeakbulut.com/phd/codes/LDA_Konu_sayisi_belirleme/code.R).



- 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 Décarry, 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](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](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-ASII%3E3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASII%3E3.0.CO;2-9)



- 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).  $\lambda$ -diverse nearest neighbors browsing for multidimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 25(3), 481-493. <https://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 Nergheş, 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](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](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](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](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](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](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](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](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](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](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](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>