Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

BİLGİ ERİŞİMDE İLGİ SIRALAMALARININ ARTIRIMLI OLARAK GELİŞTİRİLMESİ

Müge AKBULUT

Doktora Tezi

Ankara, 2022

BİLGİ ERİŞİMDE İLGİ SIRALAMALARININ ARTIRIMLI OLARAK GELİŞTİRİLMESİ

Müge AKBULUT

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

Doktora Tezi

Ankara, 2022

KABUL VE ONAY

Müge AKBULUT tarafından hazırlanan "Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi" başlıklı bu çalışma, 1 Haziran 2022 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Doktora Tezi olarak kabul edilmiştir.

Prof. Dr. Fazlı CAN (Başkan)

Prof. Dr. Yaşar TONTA (Danışman)

Prof. Dr. Umut AL (Üye)

Doç. Dr. Yurdagül ÜNAL (Üye)

Doç. Dr. İhsan Tolga MEDENİ (Üye)

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. Uğur ÖMÜRGÖNÜLŞEN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinleri yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

- Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.
  - Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
  - Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ….. ay ertelenmiştir. (2)
  - Tezimle ilgili gizlilik kararı verilmiştir. (3)

……/……/……

Müge AKBULUT

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.
**** Tez danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, **Prof. Dr. Yaşar TONTA** danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Müge AKBULUT

TEŞEKKÜR

Bu çalışma sürecinde desteğini gördüğüm pek çok kişiye teşekkür borçluyum. En başta bana her zaman yol gösteren ve düşünmemi sağlayan, değerli danışmanım, sevgili hocam Prof. Dr. Yaşar Tonta'ya sonsuz emekleri, sabrı ve nezaketi için içtenlikle teşekkür ederim. Kendisiyle çalışabildiğim için kendimi çok şanslı hissediyorum.

Tez İzleme Komitemde yer alarak değerli görüşlerini ileten ve tezin geliştirilmesine önemli katkılar sağlayan Prof. Dr. Fazlı Can ve Prof. Dr. Umut Al'a emeklerinden dolayı teşekkür ederim. Tez savunma sınavımda yer almayı kabul ederek değerli görüşlerini paylaşan Doç. Dr. Yurdagül Ünal'a ve Doç. Dr. İhsan Tolga Medeni'ye de teşekkür borçluyum.

Desteğini her zaman hissettiğim, beni her koşulda gülümseten sevgili arkadaşım Arş. Gör. Nisa Öktem'e ne kadar teşekkür etsem az. Sevgili arkadaşım Öğr. Gör. Pınar Gevheroğlu'ya da sağladığı motivasyon için ve çalışmam sırasında ihtiyaç duyduğum kaynakları çok hızlı bir şekilde temin ettiği için çok teşekkürler. İyi ki varsınız. Umarım bir gün tekrar birlikte çalışma şansımız olur.

Sevgili çalışma arkadaşlarım Arş. Gör. Dr. Erdinç Alaca ve Arş. Gör. Dr. Demet Soylu'ya da sabır ve destekleri için çok teşekkürler.

Çalışmam sırasında takıldığım yerlerde kendilerine danıştığım, sorularımı özenle cevaplayan Prof. Dr. Chaomei Chen ve Prof. Dr. Howard D. White'a teşekkür ederim.

iSearch derlemiyle ilgili yardımları için iSearch Team üyelerine (Peter Ingwersen, Birger Larsen, Haakon Lund ve Marianne Lykke) ve çalışmam sırasında kullandığım kaynakların temin edilmesinde yardımını gördüğüm sevgili Bora Somunoğlu'ya teşekkür ederim.

Çalışmanın önceki sürümünü okuyarak değerli önerilerde bulunan sevgili arkadaşım Dr. Öğr. Üyesi Sümeyye Akça'ya dikkatli okuması ve desteği için teşekkür ederim.

Her zaman ve her koşulda desteklerini hissettiğim sevgili anneme ve babama hep yanımda oldukları için minnettarım.

ÖZET

AKBULUT, Müge. *Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi*, Doktora Tezi, Ankara, 2022.

İlgi sıralaması algoritmaları erişilen belgeleri arama sorgularıyla belgeler arasındaki konusal benzerlik (ilgi) derecelerine göre sıralamaktadır. Fakat bazen sıralamada birbirine çok benzeyen kaynaklara ek olarak sorgulanan konunun çeşitli yönlerini ele alan makalelere de ihtiyaç duyulmaktadır. Bu yüzden özellikle literatür taramalarında erişilen makalelerin konu çeşitliliği de önemlidir. Dahası, ilgi sıralamaları kullanıcıların bilgi ihtiyaçlarına göre kişiselleştirilebilmelidir.

Bu çalışmanın amacı yeni bir ilgi sıralaması yöntemi geliştirmektir. Bu amaçla önce 65 sorgu için arXiv'den alınan iSearch derlemindeki yaklaşık 435 bin fizik makalesinin özetlerine LDA (Latent Dirichlet Allocation – Gizli Dirichlet Ayırımı) olasılıksal konu modelleme algoritması uygulanarak ilgi sıralamaları elde edilmiştir. Daha sonra bu sıralamalar ilgi kuramı, bilgi erişim ve bibliyometriye dayanarak geliştirilen pennant erişim yöntemiyle desteklenerek artırımlı olarak geliştirilmiş yeni ilgi sıralamaları oluşturulmuştur. Bulgular konu modelleme algoritması ile elde edilen ilgi sıralamaları atıf verileriyle bütünleştirildiğinde (1) ilgi düzeyleri daha yüksek ve çeşitli makaleler içeren daha zenginleştirilmiş ilgi sıralamaları oluşturulabileceğini, (2) sıralamaların kullanıcıların ihtiyaçlarına/önceliklerine göre kişiselleştirilerek yeniden sıralanabileceğini ve (3) erişim çıktılarının görselleştirilerek literatürün daha kolay izlenebileceğini göstermektedir.

Bu araştırma LDA konu modelleme algoritması ile elde edilen ilgi sıralamalarının atıf verilerine dayanan pennant erişim teknikleriyle artırımlı olarak geliştirilebileceğini gösteren ilk çalışmadır. İlgi sıralamalarını oluşturmak için kullanılan veriler (özet ve başlıklar, toplam atıf ve ortak atıf sayıları) atıf dizinlerinde mevcuttur. Dolayısıyla geliştirdiğimiz yöntem hesaplama, sağlamlık, tekrarlanabilirlik ve ölçeklenebilirlik sorunları çözümlendiğinde yakın gelecekte, örneğin, Web of Science, Scopus ve TR-Dizin'de kullanılabilir.

Anahtar Sözcükler

İlgi sıralamaları, olasılıksal konu modellemesi, Gizli Dirichlet Ayırımı (LDA) algoritması, pennant erişim, Maksimum Marjinal İlgi (MMR), konusal çeşitlilik.

ABSTRACT

AKBULUT, Müge. *Incremental Refinement of Relevance Rankings in Information Retrieval*, Ph.D. Dissertation, Ankara, 2022.

Relevance ranking algorithms rank retrieved documents based on the degrees of topical similarity (relevance) between search queries and documents. However, in some cases, sources that address various aspects of a queried topic are needed in addition to the articles that demonstrate a high level of similarity with the search query. Therefore, topical diversity of retrieved articles is also essential, especially in literature search results. Moreover, relevance rankings should be personalized based on users' information needs.

The aim of this study is to develop a new relevance ranking method. To that end, firstly, the relevance rankings for 65 search queries were obtained by applying the LDA (Latent Dirichlet Allocation) probabilistic topic modeling algorithm to the abstracts of some 435,000 physics articles in the iSearch corpus taken from arXiv. Then, these rankings were supported by the pennant retrieval method based on relevance theory, information retrieval, and bibliometrics, and incrementally refined new relevance rankings were created. Findings show that when the relevance rankings obtained by the topic modeling algorithm are fused with the citation data: (1) more enriched relevance rankings containing higher relevance levels with more diverse articles can be created; (2) the rankings can be personalized based on users' information needs; and (3) the literature can be followed more easily by visualizing the retrieval outputs.

Our research is the first to show that LDA-based relevance rankings can be incrementally refined with the pennant retrieval techniques based on citation data. The data used to create relevance rankings such as titles, abstracts, and the total number of citations and co-citations are readily available in the citation indexes. Hence, the method we developed can be used in, for instance, Web of Science, Scopus, and TR-Dizin in the near future once the computation, robustness, reproducibility, and scalability issues are resolved.

Keywords

Relevance rankings, probabilistic topic modeling, the Latent Dirichlet Allocation (LDA) algorithm, pennant retrieval, Maximal Marginal Relevance (MMR), topical diversity.

İÇİNDEKİLER

KISALTMALAR DİZİNİ

DCG	İndirimli Birikimli Kazanç (Discounted Cumulative Gain)
df	Belge sıklığı (document frequency)
idf	Ters belge sıklığı (inverse document frequency)
LDA	Gizli Dirichlet Ayırımı (Latent Dirichlet Allocation)
MMR	Maksimum Marjinal İlgi (Maximal Marginal Relevance)
NDCG	Normalleştirilmiş DCG (Normalized Discounted Cumulative Gain)
tf	Terim sıklığı (Term frequency)
WoS	Web of Science

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

1. BÖLÜM: GİRİŞ

1.1. KONUNUN ÖNEMİ VE KAVRAMSAL ARKAPLAN

Bilgi erişim kullanıcının bilgi ihtiyacını tanımladığı sorgudaki terimler[1] ile belgelerde geçen terimlerin eşleştirilmesine dayanmaktadır. Fakat bu süreçte belge ve sorgu temsilinde aynı kavramın farklı biçimlerde temsil edilebilme olasılığından kaynaklanan belirsizlikler söz konusudur (Ganguly ve Jones, 2018). Temsil için belirlenen terimler öznel olduğu için kişiye, zamana ve duruma göre değişebilir (Cormack ve Grossman, 2017; Jiang, Liu ve Gao, 2015; Swanson, 1986a; Voorhees, 2000). Dolayısıyla belgenin konusu, ilgi düzeyi, bilgi ihtiyacının ne olduğu gibi hususlarda ortak bir anlayış bulunmamaktadır (Wilson, 1978). Bilgi erişimin mantıksal organizasyonundan kaynaklanan bu erişim problemleri nedeniyle sistemdeki tüm ilgili belgelere ve sadece ilgili belgelere erişim sağlayacak ideal bilgi sistemi tasarlamak mümkün değildir (Mizzaro, 1997; Wilson, 1978). Ancak 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). Bu noktada ilgi sıralamaları (relevance rankings) kullanıcı tatmini açısından önemli rol oynamaktadır (Lei, Wang, Chen ve Li, 2001). Çünkü kullanıcılar genellikle birkaç ilgili belgeye fazla çaba harcamadan eriştiklerinde tatmin olmaktadırlar (Tonta, 1992, 1995).

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 oranında), bir başka konudaki bilgi ihtiyacını ise daha az (%50 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). Benzer şekilde konu-sorgu, belge-konu ve belge-belge benzerlikleri de hesaplanabilir. Olasılıksal yöntemler ikili (binary) sınıflandırmadan (ilgili-ilgisiz) daha detaylı bilgi sağladığı için ilgi sıralamaları oluşturulmasında da sıklıkla tercih edilmektedir. Örneğin, konu modelleme (topic modeling) algoritmaları herhangi bir terim için benzer ya da eş anlamlı terimlerin de geçtiği belgeleri listeler (Hornik ve Grün, 2011). 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, Ng ve Jordan, 2003; Li ve McCallum, 2006). Konu modelleme sırasında ilgiyi belirleyebilmek için derlemde (collection) yer alan belgeler hem belli bir belgede geçen

[1] Bu tezde "terim", "sözcük" ve "kelime" sözcükleri eş anlamlı olarak kullanılmıştır. Aynı şekilde "belge", "kaynak", "çalışma" ve "makale" sözcükleri de eş anlamlı olarak kullanılmıştır.

kelimeler hem de farklı belgelerde geçen kelimeler birlikte geçiş sıklıkları açısından incelenmektedir. Böylece her belgenin bir veya birden fazla konuya ait olabileceği sonucunu veren model oluşturulur ve konu sayısı algoritmaya girdi olarak verildikten sonra LDA tarafından her belge için saptanan konuların olasılık dağılımı elde edilmiş olur.[2] LDA algoritmasının çıktısı, belli konular altında sınıflandırılmış belgeler (documents in topics), bir konu altında sınıflandırılmış kelimeler (topic words) ve belli belgelerde yer alan konular (topics in documents) olmak üzere üç ana sınıftan oluşmaktadır.

İlgi sıralaması oluşturma problemi çoğunlukla belge ve sorgu arasında eşleşme (token matching) problemine indirgenmektedir (Montazeralghaem, Rahimi ve Allan, 2020). Çünkü tam eşleşme (exact matching), sorgunun belge ile ilgisini değerlendirmek için kullanılan en önemlli sinyallerden biridir (Boualili, Moreno ve Boughanem, 2022, s. 1). LDA algoritmasında da ilgi sıralaması oluşturma problemi belge-belge ve sorgu-belge benzerliği olarak tanımlanmaktadır (Vorontsov ve Potapenko, 2014). 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 konusal (tematik) bağlamları ve bu bağlamlar kullanılarak ilgi olasılıklarının tahmin edilmesi de gerekmektedir (Guo, Fan, Ai ve Croft, 2016; Ren ve diğerleri, 2018; Wu, Luk, Wong ve Kwok, 2007).

Konu modellemesi belgelerin hangi konulardan hangi oranda bahsettiğini anlamak ve konuların hangi oranda hangi kelimeleri potansiyel olarak içerebileceğini ortaya çıkarmak açısından önemlidir. Fakat temelde kelime sıklıklarına dayalı olan bu model tek başına bağlam ve tematik ilgi yakalama konusunda yeterince başarılı değildir.

İlgi kararının verilmesi sürecinde konusal bağlantılar ile ilgili en önemli ipuçları atıflardan elde edilmektedir (Carevic ve Schaer, 2014). Bu durum, makalesinde belli kaynaklara atıf yapan bir yazarın bu çalışmaları kendi çalışması ile ilgili bulduğu varsayımına dayanır (Akbulut, 2016). Buradan hareketle örneğin, atıf yapan ve atıf yapılan yayın arasında bir anlamsal ilişki olabileceği düşünülerek geliştirilen ortak atıf analizi ile konusal ilgi örüntüleri ortaya çıkarılabilmektedir (Han, 2020; Knoth ve diğerleri, 2017; Küçüktunç, Saule, Kaya ve Çatalyürek, 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

[2] LDA algoritmasının işleyişi ile ilgili ayrıntılı bilgi için bkz. Bölüm 2.2.

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, Ju, Wong, Shmulevich ve Chiang, 2017).

Bibliyometri ile bilgi erişimin ilişkilendirilmesi ilk olarak bibliyografik eşleştirme (bibliographic coupling) ile başladı (Kessler, 1963). Bibliyografik eşleştirmede atıf yapan ve atıf yapılan çalışmaların kaynakça benzerliğinden faydalanılmaktadır (Carevic ve Mayr, 2014). İki çalışmanın kaynakçası ne kadar fazla ortak kaynak içeriyorsa bu iki çalışmanın aynı konuda olma olasılıkları o kadar yüksek ve dolayısıyla bu iki çalışma birbiriyle o kadar ilgilidir (Akbulut, 2016, s. 18). Örneğin, Web of Science (WoS) ilgili kayıtlar (related records) özelliği ile arama yapılan makalenin kaynakçası ve veri tabanındaki diğer çalışmaların kaynakçalarına bakılarak hangi çalışmaların kaynakçalarının arama yapılan çalışmanın kaynakçasıyla daha çok örtüştüğü belirlenmektedir. İlgi sıralaması, kaynakçası en çok örtüşen kaynaktan başlayarak listelenmektedir. Buna ek olarak algoritma çoğunlukla güncel kaynakları ilk sıralarda listelemektedir. Bu nedenle örneğin belli bir teorinin ilk defa ortaya atıldığı çekirdek makalenin sıralamada yer alma olasılığı düşmektedir. Oysa listedeki uç örnekleri kapsayan farklı konulardaki makaleleri içeren sıralamalar kullanıcılar için bazen daha faydalı olabilmektedir (Rafols, Leydesdorff, O'Hare, Nightingale ve Stirling, 2012). Bu yüzden ilgi sıralamasının uç örnekleri kapsayacak şekilde oluşturulması önemlidir.

Bibliyometriyle bilgi erişimi ilişkilendiren çalışmalar ortak atıf analizi (co-citation analysis) çalışmalarıyla devam etmiştir (Small, 1973). Belli iki belgeye başka çalışmalarda ne kadar çok birlikte atıf yapılmışsa bu iki belge konusal açıdan o kadar çok birbirine benzemektedir. Belgeler arasındaki benzerliğe karar vermek için bibliyografik eşleştirme ve ortak atıf analizi birlikte de kullanılmaktadır (Bichteler ve Eaton, 1980). Ortak atıf sayıları dinamik ve ileriye dönük, bibliyografik eşleştirme bağlantıları ise durağan (statik) ve geriye dönüktür. Başka bir deyişle, iki belge gelecekte de ortak atıf almaya devam edebilir, ama aynı şey bibliyografik eşleştirme için söz konusu değildir (Bichteler ve Eaton, 1980, s. 279; Garfield, 2001, s. 3; Sugimoto ve Larivière, 2018, s. 67). Her iki yöntem de belgelerin benzerliklerini ve ilgisini hesaplamak için öneri sistemlerinde (recommendation systems) ilgi sıralaması oluşturmak amacıyla kullanılmaktadır (Beel, Gipp, Langer ve Breitinger, 2016; Strohman, Croft ve Jensen, 2007). Fakat ortak atıf analizine dayanan bilgi erişim algoritmaları bibliyografik eşleştirmeye göre daha yüksek performans göstermektedir (Waltman ve Van Eck, 2012; Zarrinkalam ve Kahani, 2012).

Sperber ve Wilson'ın (1995) ilgi teorisine (relevance theory) göre bir girdinin ilgisini belirleyen şeyler, o girdinin *bilişsel etki*si (cognitive effect) ile o girdiyi işlemek için gereken *işleme*

*kolaylığı*dır (ease of processing). Dolayısıyla bir girdinin ilgisini belirleyen şey o girdinin *bilişsel etki*si ile *işleme kolaylığı*nın birbirine oranıdır ve sıralama (ordinal) ölçeği ile ölçülebilir (Clark, 2013). Diğer yandan Salton'ın vektör uzayı (vector space) modelindeki *tf***idf* (terim sıklığı*ters belge sıklığı) formülü bilgi erişim sistemlerinde sorgu terimlerinin (*tf*) ve dizin (index) terimlerinin (*idf*) ağırlıklandırılması amacıyla yaygınlıkla kullanılmaktadır (Manning, Raghavan ve Schutze, 2008). Pennant erişim yaklaşımı ise temelde ilgi teorisine, vektör uzayı modeline ve bibliyometriye dayanmaktadır. Ters belge sıklığı yöntemi ile yakından ilgilidir. Pennant erişimde ters belge sıklığındaki konusal isim öbeklerinin ağırlıklandırılması yerine bibliyometrik dağılımlar kullanılmaktadır. Belgeler ve yazarlar arasındaki ilgi bilişsel etki ve erişim kolaylığı ile ilişkilendirilerek konusal ilgi (*tf*) ve bilgiyi elde etmek için harcanan çaba (*idf*) olarak sırasıyla *x* ve y eksenlerine yansıtılır (White, 2007a, 2007b, 2009, 2010, 2015, 2016). Dolayısıyla *tf***idf* formülü farklı bir biçimde yorumlanıp, bilişsel etki ve çaba bilgisi de kullanılarak ilgi sıralamaları elde edilebilmektedir.[3]

Pennant erişim yöntemi ile hesaplanan ilgi sıralamasında ilk sıralarda olan çalışmalar hem bilişsel etki hem de işleme kolaylığı ölçeklerinde (scales) en yüksek puanı almış olan çalışmalardır. İlgi sıralaması hesaplanırken ortak atıf değerinden elde edilen bilişsel etki ölçüsü de hesaba katıldığından, işleme kolaylığı ölçeğinde daha aşağıda olan ögelerin ilgi sıralamasında ilk sıralarda olabilmesi için bilişsel etki ölçeğinde ilk sıralarda olması gerekir (ortak atıf ve toplam atıf değerlerinin birbirine yakın ve görece yüksek olması gerekir).

Araştırma kapsamında en uygun olarak tanımlanan sorguyla ilgili ve çeşitli konulardaki makalelerin üst sıralarda yer aldığı sıralamalar için sınırları genişleten (boundary spanning) makaleler değerlidir. Bu tür marjinal[4] makalelerin pennant erişim algoritmasındaki karşılığı ise ortak atıf sayısı nispeten az olan ve başka disiplinlerle ilişki kurulmasını sağlayan makalelerdir.

Kelime tabanlı konu modelleme yaklaşımı LDA ve atıf tabanlı yaklaşım olan pennant erişim algoritmaları ayrı ayrı değerlendirildiğinde ikisinin de bazı eksik yanlarının olduğu bilinmektedir. Örneğin, kelime tabanlı yaklaşımlar farklı alanlardaki özdeş kavramların değişik kullanımlarının neden olduğu belirsizlikten (ambiguity) etkilenmektedir (bazen "yapay öğrenme" ile "makine öğrenmesi" eş anlamlı olarak kullanılmaktadır). Öte yandan, iki farklı kavram farklı alanlarda aynı adla kullanılabilir (Küçüktunç, Saule, Kaya ve Çatalyürek, 2012, s. 1; Zarrinkalam ve

---

[3] Pennant erişim algoritmasıyla ilgili ayrıntılı bilgi için bkz. Bölüm 2.3.

[4] "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.

Kahani, 2012). Bu durum ilgili yayınların göz ardı edilmesine ya da listede ilgisiz yayınların yer almasına yol açabilir (Küçüktunç ve diğerleri, 2015, s. 2). Konusal bağlam yakalamada başarılı olan atıf tabanlı yaklaşımlarda ise bir çalışmanın atıf alması için belli bir zaman geçmesi gerekmektedir (Ke, Ferrara, Radicchi ve Flammini, 2015).

Konularla atıflar arasındaki ilişki genel olarak kabul edilenden daha belirsizdir (Ballester ve Penner, 2022; Harter, Nisonger ve Weng, 1993). Atıf tabanlı yaklaşımlarda genellikle bir çekirdek makaleye ihtiyaç vardır (White, 2018b, s. 758). LDA algoritmasının atıflarla desteklendiği uygulamalarda LDA'nın performansı arttığı için önemli ve etkili çalışmalara erişim sağlanmaktadır (Guo, Zhang, Zhu, Chi ve Gong, 2013; Huang, Liu, He ve Du, 2016, 2018; Li, He ve Liu, 2017; Nguyen ve Do, 2018; Wang, Zhai ve Roth, 2013; Xia, Li, Tang ve Moens, 2012; Zhou, Yu ve Hu, 2017; Zou, Liu, Buntine ve Liu, 2021).

İyi bir bilgi erişim sisteminin kullanıcının sorgusuna göre derlemdeki hangi belgelerin daha ilgili olduğunu öngörmesi ve bu belgeleri olasılık sıralama ilkesine (probability ranking principle) göre sıralaması beklenir (Robertson, 1977; Sparck Jones, Walker ve Robertson, 2000). 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). Aynı konuda bilgiye ihtiyaç duyan iki kullanıcının ilgili bulduğu kaynaklar birbirinden farklı olabilir. İdeal ilgi sıralaması, arama sonuçlarını mevcut duruma ve ihtiyaca göre uyarlamak için kullanıcı tercihlerini içermelidir (El-Arini, Veda, Shahaf, Guestrin, 2009).

Farklı erişim algoritmalarının farklı ilgili belgelere eriştikleri bilinmektedir (Croft, 2002). Bilgi erişimde veri tümleştirme (data fusion) ile birden fazla algoritmadan elde edilen farklı özellikteki (örneğin ilgili kaynakların ya da çeşitli kaynakların üst sıralarda olduğu) sıralamalar birleştirilerek daha ilgili sıralamalar elde edilmektedir (Baeza-Yates ve Ribeiro-Neto, 1999; Nuray ve Can, 2006). Tümleştirilmiş sıralama sürece dâhil olan tüm algoritmalardan daha iyi performans göstermektedir (Meng, Yu ve Liu, 2002). Sıralama tümleştirme problemi *polinom sınırlı*dır (Nabiyev, 2013). Diğer bir deyişle, tümleştirilecek sıralamaların sayısı ve sıralamaların içerdiği kaynak sayısı ya da sıralamanın uzunluğu örneğin iki katına çıktığında hesaplama miktarı da artar, fakat "çılgınca" artmaz (Say, 2018, s. 68). Her ne kadar üstel bir artış söz konusu olmasa da birleştirme aşamasında hesaplamanın süresini kısaltmak için artırımlı geliştirme (incremental refinement, incremental boosting) yöntemleri kullanılmaktadır. Hafızaya alma ve yeniden kullanım olarak tanımlanabilecek artırımlı yöntemler bu noktada yüksek hesaplama maliyetini önlemektedir (Jin, Valizadegan ve Li, 2008; Rao, 2004). Bu tür artırımlı yapılar, her biri bir

öncekinin ürettiği çıktıyı geliştiren bir dizi aşamadan oluşacak şekilde tanımlanmakta ve işlevselliği de önemli ölçüde artırmaktadır (Winograd, 1997).

Daha ilgili sıralamalar oluşturmak amacıyla belli bir sıralamadaki kaynaklar çeşitli özelliklerine göre ağırlıklandırılarak ve artırımlı olarak geliştirilerek yeniden sıralanabilir. Bu sayede kullanıcılara anlık olarak istedikleri türde (örneğin popülerlik ya da çeşitlilik oranı yüksek makalelerin üst sıralarda olduğu) sıralama elde etme şansı da sağlanmaktadır. Fakat, algoritmalar kendilerini oluşturan bileşenlere ve bileşenlerin her birine verilen ağırlıklara karşı duyarlıdır. Özellikle farklı konulardaki makaleleri içeren ilgi sıralamaları, erişim sonuçlarının standarttan veya çoğunluktan sapmasını ölçen ön yargı (bias) kavramı ile yakından ilgilidir (Mowshowitz ve Kawaguchi, 2002). Oluşturulan sıralamada bazı özellikteki belgelerin gereğinden fazla dâhil edilmesi ya da hariç tutulması ön yargı anlamına gelmektedir. Bu yüzden hem sıralama tümleştirme hem de yeniden sıralama aşamasında ön yargı kavramına dikkat edilmesi gerekmektedir (Dwork, Hardt, Pitassi, Reingold ve Zemel, 2012).

1.2. ARAŞTIRMANIN AMACI

Bu çalışmanın temel amacı, konu modellemesi ile atıflara dayanan pennant erişim algoritması yaklaşımını bir arada kullanarak ilgi ve çeşitlilik oranı yüksek, sorgudaki terimlerin ya da yöntemin farklı uygulamalarının gözlenebildiği, marjinal ve sorguyla ilgili kaynakları içeren ilgi sıralamaları oluşturmaktır.

Bu amaçla 2009 yılına kadar arXiv'e eklenen tüm fizik konulu makaleleri içeren iSearch derlemi üzerinde bir uygulama yapılmıştır. Derlemde yer alan çalışmaların özet ve başlık bölümlerine LDA algoritması uygulanarak oluşturulan ilgi sıralamaları, pennant erişim algoritması ile desteklenerek alternatif ilgi sıralamaları oluşturulmuştur. Çalışmaların içeriği hakkında genel bilgi içeren özetler üzerinde konu modellemesi yapılarak sonuçlar atıf bilgileri ile desteklendiğinde tam metin konu modellemesine benzer ya da daha iyi bir performans elde edilebileceği öngörülmüştür. Bir çalışmanın tam metninde farklı bağlamlarda geçen bazı terimler LDA kapsamında değerlendirmeye alınmasa bile, o çalışmanın kaynakçasında geçen diğer çalışmalar hesaplamaya dâhil edildiğinde, ortak atıf ve toplam atıf sayıları kullanılarak belgelerin bağlamı ve ilgisi konusunda önemli ipuçları elde edileceği varsayılmıştır. Böylece hem algoritma belgelerin sadece özet bölümlerine uygulandığı için konu modellemesinin hızlı yapılması hem de istenen özelliklerin ön planda olduğu ilgi sıralamalarının ortaya çıkarılması planlanmıştır.

Bu çalışmanın bir diğer önemli amacı da çeşitlilik derecesi yüksek sıralamalar oluşturmaktır. Özellikle farklı bağlamlardaki kullanımları yakalayan pennant erişim yönteminin hesaplamaya dâhil edilmesiyle çeşitlilik oranı yüksek sıralamalar oluşturulabileceği öngörülmüştür.

Önerilen yöntemle konu açısından en kapsayıcı sıralama oluşturulduktan sonra, kullanıcılara arama yaptıkları konu ya da terimle ilgili spesifik makaleleri ya da öncül yazarların çalışmalarını ön plana çıkaran sıralamalar oluşturma fırsatı da sağlanabilir. Ek olarak, yukarıda anılan yöntemlerle hesaplanan ağırlıklar algoritmaya dâhil edilerek sıralamalar kullanıcıların isteğine göre çeşitlilik ve ilgi derecelerine göre şekillendirilebilir. Böylece önerilen yöntemle çeşitlilik ve ilgi oranları daha yüksek ve kullanıcıların ihtiyaçlarına göre kişiselleştirilebilen ilgi sıralamaları elde edilmesi amaçlanmıştır.

1.3. ARAŞTIRMA SORULARI VE HİPOTEZLERİ

Terimlerin birlikte geçiş haritaları ve konu modelleme yöntemleri ayrı ayrı incelendiğinde başarılı sonuç vermemektedir (Leydesdorff ve Nerghes, 2017). Bu araştırmada ise atıf yapılan kaynaklara gömülü (embedded) konusal ilişkilerin bütünleşik bir şekilde analiz edilerek ilgi sıralamalarına yansıtılması ile daha başarılı sıralamalar ve araştırmacılara faydalı etkileşimli grafikler oluşturulabileceği fikrinden hareket edilmektedir. Bu bağlamda belgelerin sadece özetlerine uygulanan konu modellemesi ile oluşturulan ilgi sıralaması performansının atıf bilgileri kullanılarak geliştirilmesi hedeflenmektedir.

Araştırma kapsamında aşağıdaki araştırma sorularına yanıt aranmaktadır:

1. LDA ve pennnat erişim sıralamaları ile tümleşik sıralamalarda hangi özellikler öne çıkmaktadır?
2. Çalışmaların özetlerine uygulanan LDA konu modelleme algoritmasıyla oluşturulan sıralama ile, toplam atıf ve ortak atıf verileri de dâhil edilerek hesaplanan pennant erişim çıktıları artırımlı olarak tümleştirilerek (fusion) arama yapılan konuyu daha geniş kapsamda içeren ilgi sıralaması elde edilebilir mi?
3. Kelime dağılımına dayalı olasılıksal konu modelleme algoritması LDA kullanılarak oluşturulan sıralamaya pennant erişim yönteminin katkıları nelerdir?
4. Pennant erişimin artırımlı olarak geliştirilen ilgi sıralamasına katkısı ortak atıf ve toplam atıf sayıları ile doğrudan ilgili midir? Ortak atıf sayısı az olan derlemlerde de bu yönteme başvurulabilir mi?
5. Arama sırasında araştırmacının önceliğine göre belirli özelliklerin (örneğin arama yapılan konu üzerine inşa edilen makaleler) ön planda çıktığı ilgi sıralamaları anlık olarak oluşturulabilir mi?
6. LDA konu modelleme algoritması pennant erişim ile desteklenerek, LDA modelinin kaçırdığı temel kaynaklara ve ilgili terimin kullanıldığı diğer alanlardaki çalışmalara erişim sağlanabilir mi?

Bu bağlamda temel hipotez (H1) "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 konuyu tüm yönleri ile ele alan artırımlı olarak geliştirilmiş ilgi sıralamaları oluşturulabilir" şeklinde belirlenmiştir. Bir diğer hipotez (H2) ise "İlgi sıralamaları kullanıcının ihtiyacına göre; öncül çalışmalar, spesifik makaleler, makalelerin alana etkisi gibi özellikleri öncelenecek şekilde yeniden sıralanabilir" şeklinde oluşturulmuştur.

Araştırma hipotezlerine ve sorularına yanıt bulunması ilgi ve çeşitlilik değerleri daha yüksek artırımlı olarak geliştirilen tümleşik ilgi sıralamalarının oluşturulmasının mümkün olup olmadığını ortaya çıkaracaktır. Başarılı sonuçlar alındığı takdirde bu tezde önerilen konu modellemesi ile atıflar ve ortak atıflara dayanan artırımlı ilgi sıralaması geliştirme yönteminin uygulanabilirliği dinamik derlemler üzerinde ölçeklenebilirlik, hesaplama yükü, vb. gibi açılardan test edilebilir ve bilgi erişim sistemlerinde (örneğin, öneri sistemleri) ve atıf dizinlerinde kullanılabilir.

1.4. ÖZGÜN DEĞER

İlgi sıralamaları araştırmacıların literatür izlemelerini kolaylaştıran en önemli unsurdur. Bu bağlamda ilgi sıralamaları atıf dizinleri özelinde bugüne kadar birçok araştırmaya konu olmuş, yerli ve yabancı araştırmacılar tarafından değişik amaçlarla incelenmiştir. Ancak her ne kadar bibliyometrik bilgilerin (atıf ve ortak atıf değerleri) erişim performansını ciddi düzeyde artırdığı kanıtlanmış olsa da bu bilgilerin bileşen olarak değerlendirildiği algoritmalar henüz atıf dizinlerine entegre edilmemiştir (Akbulut, Tonta ve White, 2020). Hâlihazırda atıf dizinlerinde yer alan bibliyometrik bilgiler ilgi sıralaması oluşturulması amacıyla ya çok temel düzeyde (örneğin, kaynakça çakıştırma) kullanılmakta ya da hiç kullanılmamaktadır (Vergoulis ve diğerleri, 2019). İlgi sıralamalarının iyileştirilmesi için, konu modellemesi LDA ve atıflara dayanan penannt erişim bileşenlerinden oluşan bir algoritma ise bugüne kadar önerilmemiştir. Bu araştırma kapsamında önerilen algoritma ile atıf veri tabanlarındaki mevcut bilgiler kullanılıp kullanıcılara daha isabetli ilgi sıralamaları sunularak bu eksikliğin bir ölçüde giderilmesi amaçlanmaktadır. Bunun yanı sıra, önerilen algoritma ile araştırmacıların arama yaptıkları terimle (ya da konu) ilgili temel kaynakları ve ilgili terimin başka hangi alanlarda kullanıldığını izlemelerinin kolaylaştırılması hedeflenmektedir.

1.5. YÖNTEM

Çalışma kapsamında Lykke, Larsen, Lund ve Ingwersen (2010) tarafından oluşturulan iSearch derlemi tercih edilmiştir. Söz konusu derlem 65 sorgu, 434.813 fizik makalesi ve bu makalelere ait 3,7 milyondan fazla dâhili referanstan[5] oluşmaktadır. Neredeyse tüm fizik makaleleri bilimsel dergilerde yayımlanmadan önce arXiv'e yüklendiğinden, arXiv'in böyle bir araştırma için iyi bir derlem olduğu düşünülmektedir (Ginsparg, 2016).

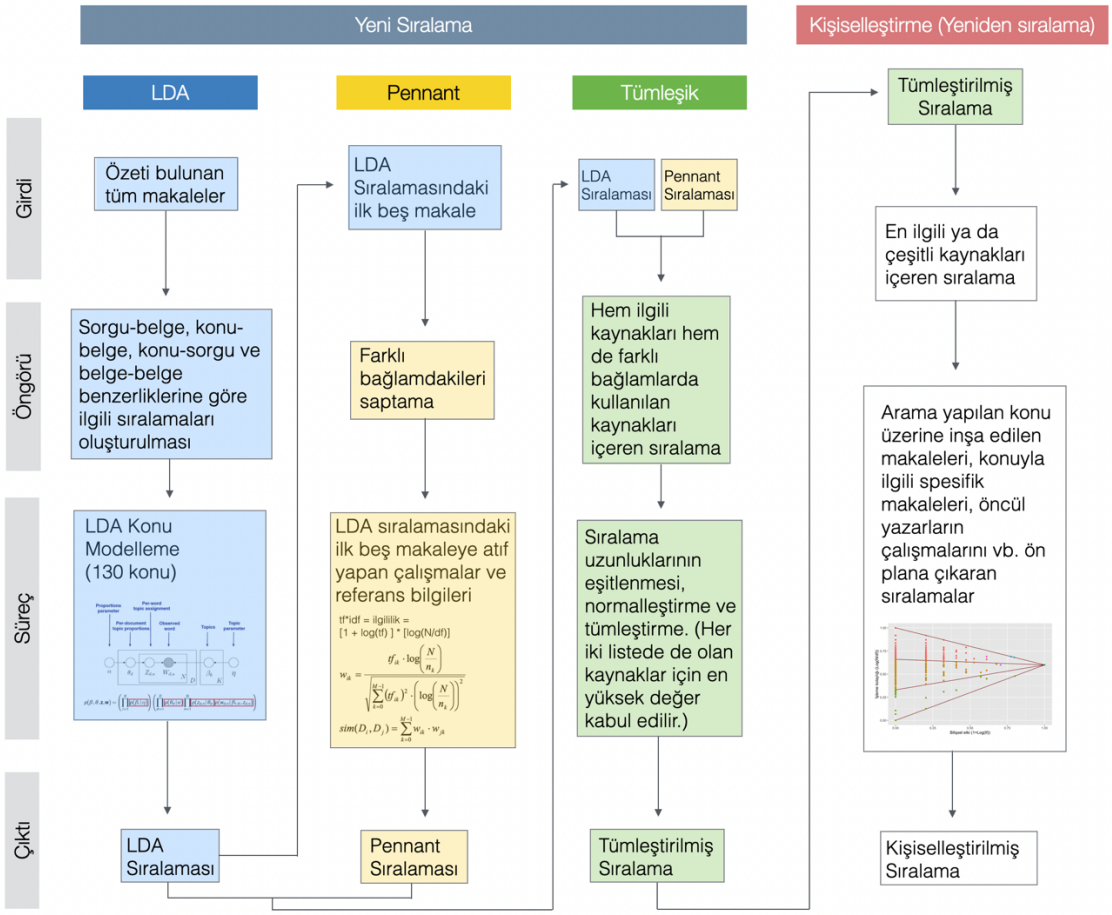
Bu araştırmada ilgi sıralamaları oluşturulurken hem kelime sıklıklarından hem de atıf ilişkilerinden yararlanılmıştır. Sıralamaların ilgi oranlarının karşılaştırılması için derece merkeziliği (degree centrality) ölçevi (metric) kullanılmıştır. Her bir çalışma için derece merkeziliği değerleri hesaplanmış ve sıralamalar bu değerlere göre analiz edilmiştir. Araştırma kapsamında "çeşitlilik oranı yüksek sıralama" ise farklı konulardaki makaleleri içeren sıralama olarak kavramsallaştırılmıştır.[6] Çeşitlilik oranlarının hesaplanması için Shannon çeşitlilik indeksi ve konu sayısı değerleri dikkate alınmıştır (Lande, 1996; Bonaccorsi, Melluso ve Massucci, 2022). Önerilen algoritmada makalelerin hem entellektüel içerikleri hem de belge bağlamları dikkate alınmıştır. İzlenen yol Şekil 1'deki gibidir. Süreçte tümleşik sıralamanın oluşturulması (LDA+Pennant erişim) ve sıralamanın kişiselleştirilmesi (yeniden sıralama) olmak üzere iki ana adım bulunmaktadır.

İlk aşamada olasılıksal konu modellemesine hazırlık için ön işlemler (noktalama işaretlerinin temizlenmesi, konu sayısının belirlenmesi vb.) gerçekleştirildikten sonra belgelerin özetleri üzerinde LDA olasılıksal konu modelleme algoritması çalıştırılarak ilgi sıralaması elde edilmiştir. Ardından sorgu ile ilgili olan konu(lar)daki belgelere atıf yapan belgelerin referans bilgileri de hesaplamaya dâhil edilerek pennant erişim algoritması uygulanmıştır. Son olarak bu iki sıralama birleştirilerek tümleştirilmiş sıralama elde edilmiştir. Kişiselleştirme aşamasında ise kullanıcının ihtiyacına göre liste yeniden sıralanmaktadır (re-ranking). Bu adımda iki algoritmadan birine ağırlık verilmesinin yanı sıra pennant diyagramlarındaki sektör bilgileri de kullanılmaktadır.

---

[5] "Dâhili referans" sadece iSearch derleminde yer alan makaleler anlamına gelmektedir. Diğer bir deyişle iSearch derlemindeki makalelerin kaynakçalarında yer alan kaynaklar eğer arXiv'de yer almıyorsa veri setine dâhil değildir.

[6] Çeşitlilik oranı yüksek olan sıralamalar interdisiplinerlik derecesi (degree of interdisciplinarity) yüksek sıralamalar olarak da nitelendirilebilir (Bonaccorsi, Melluso ve Massucci, 2022).



Şekil 1. İlgi sıralaması oluşturulması sırasında uygulanan işlemler

LDA ve Pennant erişim algoritmaları uygulanarak oluşturulacak yeni ilgi sıralamasının farklı konudaki makaleleri içermesi, ama sorguyla ilgili kaynakları hâlâ üst sıralarda listelemesi öngörülmektedir. Sıralamalar için genel ilgi puanları belirlemek yerine sıralamanın özelliklerini ortaya çıkaracak ölçümler yapılmıştır.

Araştırmada MMR (Maximal Marginal Relevance - Maksimum Marjinal İlgi) algoritması sağlama amaçlı kullanılmıştır. MMR algoritmasının LDA, pennant erişim ve tümleştirme algoritmaları uygulanarak ayrı ayrı elde edilen ilgi sıralamalarına etkileri incelenmiştir. Etki oranlarına bakılarak hangi algoritmanın hangi özellikleri öne çıkardığı saptanmıştır. İlgi sıralamalarının genel değerlendirmesi için ise ağ merkezilik değerleri temel alınmış, DCG (Discounted Cumulative Gain - İndirimli Birikimli Kazanç) ve NDCG (Normalized Discounted Cumulative Gain - Normalleştirilmiş İndirimli Birikimli Kazanç) değerleri ile kapsama (coverage) ve yenilik (novelty) oranları kullanılmıştır.

Yöntemle ilgili ayrıntılı bilgi Bölüm 3'te verilmiştir.

1.6. ARAŞTIRMANIN DÜZENİ

Araştırma raporu beş bölümden oluşmaktadır:

Birinci bölümde araştırmanın kavramsal arkaplanına, ilgi sıralamalarının önemine, konusal ilgi belirleme yaklaşımlarına ve bu yaklaşımların eksik yönlerine değinilmiştir. Bunun yanı sıra araştırmanın amacına, araştırma sorularına, araştırmanın özgünlüğüne ve yöntemine yer verilmiştir.

İkinci bölümde ilgili literatürün kritik bir değerlendirmesi yapılmıştır.

Üçüncü bölümde sıralamaların oluşturulması (ya da yeniden sıralanması) için kullanılan verilerin toplanması, analize uygun hale getirilmesi sürecinde izlenen yöntem ve teknikler hakkında detaylı bilgi verilmiştir.

Dördüncü bölümde araştırma sorularını cevaplamak ve hipotezleri test etmek amacıyla oluşturulan ya da yeniden sıralanan listelerin karşılaştırılmasına ilişkin ayrıntılı bulgular sunulmuş ve tartışılmıştır.

Beşinci ve son bölümde ise araştırma kapsamında elde edilen bulgular değerlendirilmiştir. Atıf dizinlerindeki veriler kullanılarak anlık olarak kişiselleştirilebilen sıralamalar oluşturulması için geliştirilen algoritmanın bilgi erişim sistemlerine ve atıf dizinlerine eklenebilmesi yönünde bazı önerilerde bulunulmuş ve konuyla ilgili gelecekte yapılabilecek çalışmalara değinilmiştir.

2. LİTERATÜR DEĞERLENDİRMESİ

2.1. GİRİŞ

Yayın sayılarının hızla artması araştırmacıların ilgili kaynaklara erişmelerini giderek zorlaştırmaktadır (Bornmann, Haunschild ve Mutz, 2021). 1950'lerin başından beri ilgi ve dolayısıyla ilgi sıralamaları bilgi erişim sistemlerinin tasarımı, optimizasyonu ve değerlendirilmesinde odak noktası olmuştur (Pang ve diğerleri, 2017; Saracevic, 2021; Verma, Yılmaz ve Craswell, 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; Leonhardt, Rudra, Khosla, Anand ve Anand, 2022; White, 2007b). Diğer bir deyişle, kullanıcılar için en az çaba gerektiren ve en kolay çıkarımlar terim eşleşmesine dayananlardır.

2.2. KELİME TABANLI YAKLAŞIM

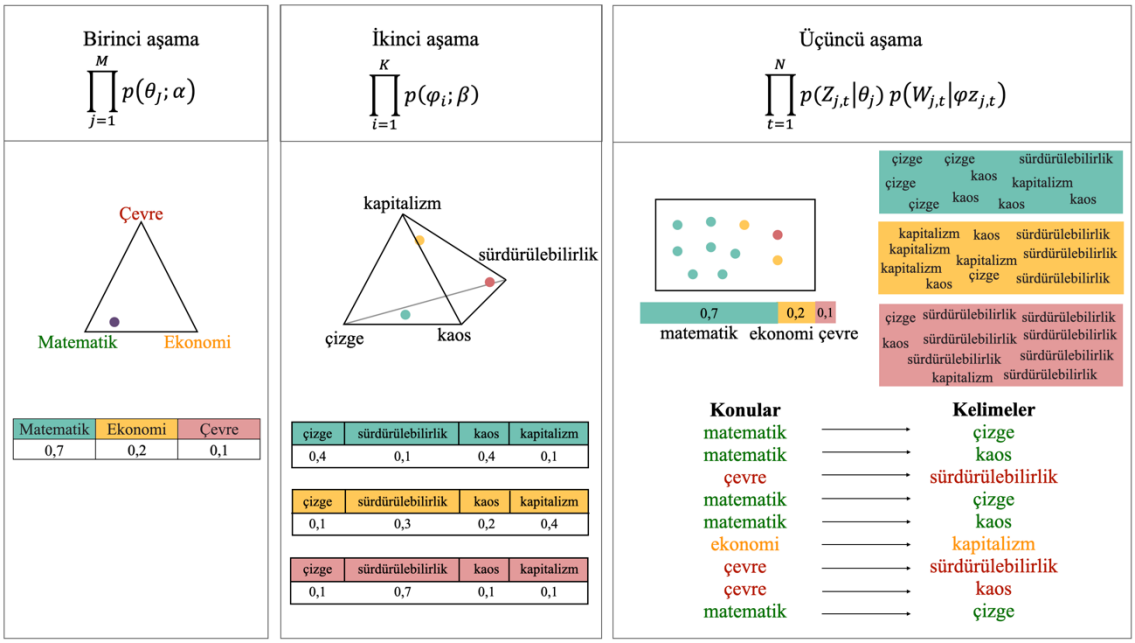
Sorgu ve belgelerde geçen terimler arasında tam eşleşme olmadığında belgeleri işlemek daha zordur. Çünkü çakışma oranları yüksek olmasa bile sorgu ve belgeler arasında konusal ilgi olabilir (Akbulut, 2016, s. 9). Bunun için belgelerin tam metinleri üzerinde doğal dil işleme yöntemleri kullanılarak ilgi düzeyleri 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 gizli (latent) tematik bilgileri belirlemek mümkündür (Boyd-Graber ve Blei, 2010). Konu modellemede belgeler, her konunun kelimelerin dağılımına göre karakterize edildiği gizli konular üzerine rastgele karışımlar olarak temsil edilir (Blei ve diğerleri, 2003, s. 996). Diğer bir deyişle konu modeli, gizli konular aracılığıyla terimler ve belgeler arasındaki ilişkiyi temsil etmektedir. Konu modelleme metinsel verilerdeki gizli anlamsal yapıyı ortaya çıkardığı için belge sınıflama ve ilgi sıralaması oluşturmada da sıklıkla tercih edilmektedir (Wang, Cao, Xu ve Li, 2012). 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, Gerrish, Wang, Boyd-Graber ve Blei, 2009).

LDA algoritması bir makalenin sınırlı sayıda konunun karışımından oluştuğu ve her kelimenin de makalenin konularından birisi ile ilişkilendirilebileceği varsayımına dayanır (Zhang, Luo, Wang ve Liu, 2015). LDA kabaca üç aşamalı bir hiyerarşik Bayes modelidir (Zhang ve diğerleri, 2015).

Bayes yaklaşımında parametreler önsel (prior) bir dağılımdan gelen rastsal değişkenler olarak görülmektedir (Alpaydın, 2017, s. 291). Diğer bir deyişle Bayes kuralı, önsel olasılık ve olabilirliği birleştirip sonsal olasılık dağılımını hesaplamamızı sağlar (Chan, 2021; Vorontsov ve Potapenko, 2014). Üç aşamalı modelde çıkarsanan dağılım, yeni bir öngörü dağılımı için girdi olarak kullanılır.

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

Formül 1'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.



Matematik	Ekonomi	Çevre
0,7	0,2	0,1

çizge	sürdürülebilirlik	kaos	kapitalizm
0,4	0,1	0,4	0,1

çizge	sürdürülebilirlik	kaos	kapitalizm
0,1	0,3	0,2	0,4

çizge	sürdürülebilirlik	kaos	kapitalizm
0,1	0,7	0,1	0,1

Şekil 2. LDA algoritmasının aşamaları

Ancak LDA algoritmasının bazı dezavantajları da 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; Nguyen ve Do, 2018; Xie, Liang, Li ve Tan, 2019). 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.

2.3. ATIF TABANLI YAKLAŞIM

Kelime tabanlı yöntemler ile ortaya çıkarılamayan belgeler arasındaki anlamsal ilişkiler kaynakça benzerliği ya da atıflar yoluyla açığa çıkarılabilir (Börner, Chen ve Boyack, 2003). Araştırmacılar atıf yaparak hem söz konusu çalışmaların entellektüel ve bilişsel katkısını kayıt altına almış hem de yazarlarına kredi vermiş olurlar (Tonta ve Akbulut, 2021, s. 391). Atıflar yazarların çalışmalarındaki iddiaları desteklemenin yanı sıra modern bilimin geçmişine ilişkin de fikir verir (Bornmann ve Mutz 2015, Comins ve Leydesdorff, 2016a; Heibi, Peroni ve Shotton, 2019; Marx, Bornmann, Barth ve Leydesdorff, 2014).

Bu süreçte farklı alanlardaki araştırmalar ile kurulan bağlantılar makalelerin kavramsal ve anlamsal içerikleri ve bağlamları 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 kullanılmaktadır (Beel ve Gipp, 2009; Beel ve diğerleri, 2016; Portenoy, 2021; Portenoy ve West, 2020).

Arama yapılan makaleye benzer makaleler araştırmacılara sunulurken çoğunlukla makalelerin kaynakça benzerliğinden faydalanılmaktadır (Carevic ve Mayr, 2014; Kessler, 1963; Vergoulis ve diğerleri, 2019). Atıf dizinleri bağlamında, temel düzeyde de olsa atıf bilgileri ilgi sıralaması oluşturmak amacıyla kullanılmaktadır (Belter, 2017). Örneğin, WoS'un ilgili kayıtlar özelliği makalelerin kaynakçaları arasındaki bibliyografik eşleşmeye 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 Griffith, 1981; White ve McCain, 1998). Kaynakça benzerliğiyle ortak atıflar birlikte kullanıldığında ise daha yüksek bilgi erişim performansı elde edilmektedir (Bichteler ve Eaton, 1980; Zarrinkalam ve Kahani, 2012).

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 ilgisi o girdinin *bilişsel etki*si ile o girdiyi işlemek için gereken *işleme kolaylığı*nın, yani *erişim kolaylığı*nın birbirine oranı olarak tanımlanmaktadır.

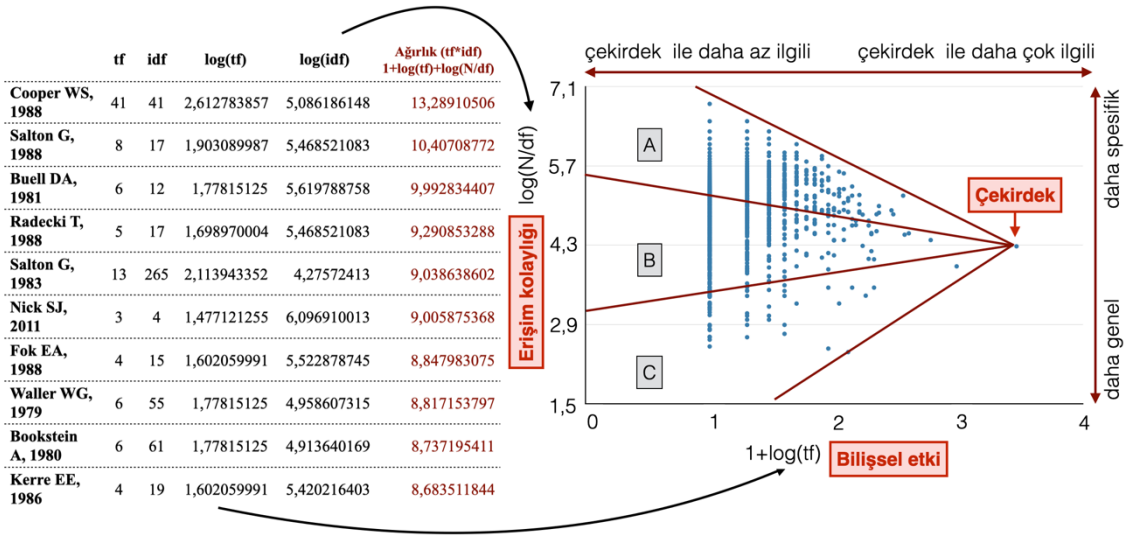
$$ilgi = bilişsel\ etki/erişim\ kolaylığı\ (işleme\ kolaylığı) \qquad (2)$$

İlgi, bir belgenin ilgi düzeyini ve bağlamını saptamak için gereken çaba (effort) ile doğrudan ilişkilidir (White, 2017). 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, 2018a, 2018b). White'ın *pennant erişim* olarak adlandırdığı bu yaklaşımın temeli *ilgi teorisi*ne, Salton'un *vektör uzayı bilgi erişim modeli*ne (Salton, Yang ve Wong, 1975) ve *bibliyometri*ye dayanmaktadır. White, vektör uzayı modelindeki *tf***idf* (terim sıklığı*ters belge sıklığı) formülünü yeniden tanımlamıştır. Bilişsel etkiyi (konusal ilgi) ve erişim kolaylığını (bilgiyi elde etme kolaylığı) hesaplamak için sırasıyla belgelerin ortak atıf (*tf*) ve toplam atıf (*idf*) sayılarından yararlanmaktadır. Pennant erişim algoritmasında 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). Formüldeki *idf* faktörü çekirdek makale ile daha az ilgili makaleleri ilgi sıralamasında aşağı itmekte ve daha ilgili olanları ise yukarı taşımaktadır. Öte yandan yüksek ilgi skoru *tf* ve *idf* değerlerinin birbirine yakın olması anlamına gelmekte ve ilgili çalışmanın çekirdek[7] makaleye yakın olarak konumlandırılmasını sağlamaktadır (Akbulut, 2016). 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).[8]

Şekil 3'teki pennant diyagramı gösteriminde 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 ve diğerleri 2020, s. 978). Pennant erişim yöntemi ile hesaplanan ilgi sıralamasında ilk sıralarda olan makaleler (ilgi puanı en yüksek olanlar) hem bilişsel etki hem de erişim kolaylığı ölçeklerinde en yüksek puanı almış olan makalelerdir.



	tf	idf	log(tf)	log(idf)	Ağırlık (tf*idf) 1+log(tf)+log(N/df)
Cooper WS, 1988	41	41	2,612783857	5,086186148	13,28910506
Salton G, 1988	8	17	1,903089987	5,468521083	10,40708772
Buell DA, 1981	6	12	1,77815125	5,619788758	9,992834407
Radecki T, 1988	5	17	1,698970004	5,468521083	9,290853288
Salton G, 1983	13	265	2,113943352	4,27572413	9,038638602
Nick SJ, 2011	3	4	1,477121255	6,096910013	9,005875368
Fok EA, 1988	4	15	1,602059991	5,522878745	8,847983075
Waller WG, 1979	6	55	1,77815125	4,958607315	8,817153797
Bookstein A, 1980	6	61	1,77815125	4,913640169	8,737195411
Kerre EE, 1986	4	19	1,602059991	5,420216403	8,683511844

Şekil 3. Örnek pennant erişim gösterimi

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

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

Pennant erişim yönteminde *x* ve *y* eksenlerine yerleştirilen ortak atıf ve toplam atıf değerleri, *tf*idf* ağırlığına göre sıralandığında gözlenmesi mümkün olmayan ilişkilerin saçılım grafiklerinde izlenmesine olanak sağlamaktadır. Bu açıdan pennant diyagramları ile disiplinlerin entellektüel yapıları izlenebilmekte ve yazarlar, çalışmalar, dergiler ya da terimler arasındaki konusal bağlantılar gözlenebilmektedir.

Makaleler ortak atıf aldıklarında pozitif kavramsal etki üretmektedirler. İki makalenin ortak atıf değerleri ne kadar yüksekse makaleler arasındaki bağlantı da o kadar kuvvetlidir. Benzer şekilde iki yazarın ortak atıf sayısı yüksekse bu yazarların aynı alanda çalışma olasılıkları da yüksektir. Ortak atıf verileri pennant diyagramları bağlamında yorumlanacak olursa, bilişsel etki (*x*) eksenindeki noktalar çekirdek makale ya da yazara yaklaştıkça ilgi de artmaktadır.

Erişim kolaylığı (*y*) eksenindeki değerler ise toplam atıf değerleridir ve diyagramdaki makalelerle çekirdek makale arasındaki bağlantının ne kadar kolay görünebildiği ile ilgilidir. Diyagramın alt kısımlarında yer alan noktalar çekirdek makaleyle göreceli olarak daha az ya da dolaylı olarak ilgili makalelerdir. Diyagramın en tepesindeki noktalar ise çekirdek makale ile ilgi açısından erişmesi kolay, yani göreli olarak daha çok ilgili olanlardır.

Pennant diyagramı yorumlanırken öncelikle çekirdek çalışma ya da yazarlardan yola çıkılarak, diyagramdaki noktalar genellikle üç gruba (sektör) ayrılmaktadır (bkz. Şekil 3). Çekirdek makalenin ardıllarının *A*, akranlarının *B*, öncüllerinin ise *C* sektöründe olması beklenmektedir.

Pennant erişim yöntemi ile arama sorgusu olarak kullanılan çekirdek makalenin ya da yazarın önceden yayımlanan çalışmalar ve yazarlar ile ilişkilerini ortaya çıkaran ve bu çalışmanın hangi modellerin ya da yapıların oluşmasında etkili olduğunu gözlemeye olanak sağlayan uygulamalar yapılmıştır (Akbulut, 2016; Akbulut ve diğerleri, 2020; Larsen, 2008; Schneider ve diğerleri, 2007; Tonta ve Özkan Çelik, 2013; White 2007a, 2007b, 2009, 2010, 2015). Ayrıca, pennant diyagramları araştırmacıların bir konu hakkındaki ilgili literatürü belli bir kavramın veya yöntemin ortaya çıkışı ve gelişimiyle birlikte daha kolay takip edebilmelerine de yardımcı olmaktadır. Örneğin, bilgi erişim literatüründe atıf klasiği[9] haline gelmiş olan ve kaynakçasında sadece iki kaynak yer alan 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 yüksek bir performans gösterdiği saptanmıştır (Akbulut ve diğerleri, 2020). WoS'un ilgili kayıtlar özelliği ile oluşturulan ilgi

[9] En az 100 atıf alan çalışmalar atıf klasiği sayılmaktadır (Fenton, Roy, Hughes ve Jones, 2002, s. 494).

sıralamasındaki makalelerin çoğu çekirdek makale ile ilgili değilken, pennant erişim yöntemi ile erişilen makalelerin tümünün ilgili olduğu ortaya çıkmıştır.

AuthorWeb sistemi ve *sowiport* dijital kütüphanesi uygulaması pennant erişim yönteminin etkileşimli uygulamalarını içeren sistemlerdir. Bu sistemlere her ne kadar günümüzde erişilemese de bu sistemler aracılığıyla çalışmalar ve yazarlar arasındaki yeni ilişkilerin keşfedildiği ilginç bulgular kayıt altına alınmıştır (Carevic ve Mayr, 2014; White, 2015; White, Buzydlowski ve Lin, 2000). Pennant erişim yöntemi ile oluşturulan ilgi sıralamaları için gerekli veriler (toplam atıf ve ortak atıf sayıları) atıf dizinlerinde yer aldığı halde atıf dizinlerinde pennant erişim yönteminin pratikte hayata geçirildiği aktif bir platform bulunmamaktadır.

2.4. İLGİ SIRALAMALARINDA ÇEŞİTLİLİK

İlk zamanlarda bilgi erişim performans değerlendirme çalışmalarının çoğu ilgi düzeyini belirlemeye odaklanmıştır (Liu, 2009). Fakat ilgi tek başına yeterli bir performans göstergesi değildir (Bradley ve Smyth, 2001; Herlocker, Konstan, Terveen ve Riedl, 2004; McNee, Riedl ve Konstan, 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 çeşitli kaynakları listeleyen ilgi sıralamaları kullanıcılar için daha faydalı olabilmektedir (Abramo, D'Angelo ve Zhang, 2018; Akbulut, 2016; Rafols ve diğerleri, 2012; Rousseau ve Hu, 2019). İlgi, azalan ilgi düzeyine göre sıralanmış bir sıralama oluşturmayı amaçlarken, erişim çıktısının çeşitlendirilmesi geniş bir konu yelpazesini kapsayan sıralanmış bir yayın listesi oluşturmaya odaklanmaktadır (Li, Feng ve Rijke, 2020; Ren ve diğerleri, 2013, 2017).

Farklı alt konuların hızlı bir şekilde kapsanması için kümeleme (clustering) yöntemi ilgi sıralamasındaki kaynakların çeşitlendirilmesinin hızlı ve etkili bir çözüm yoludur (Carpineto, D'Amico ve Romano, 2012). Bu doğrultuda ilgi sıralamalarındaki kaynaklar çeşitlendirilirken fazlalıklardan kaçınılması (sıralamada yer alan bir kaynağa çok benzeyen makalelerin göz ardı edilmesi ya da sıralamanın sonlarına ötelenmesi) kullanıcıların sıralamadan tatmin olma olasılığını artırmaktadır. İlgi sıralamalarındaki çeşitlilik bilgi erişimdeki bazı belirsizliklerin giderilmesi açısından da önemlidir. Çünkü kullanıcılar çoğu zaman tam olarak bilmedikleri konular hakkında sorgu oluşturmak durumundadırlar (Clarke ve diğerleri, 2008; Radlinski, Bennett, Carterette ve Joachims, 2009). Bilgi ihtiyacının farklı yönlerini kapsayan bir sıralama elde etmek için ise belge kümeleri arasındaki olası karmaşık ilişkilere dayanan yenilik ve çeşitlilik özelliklerini dikkate alan çözümler gerekmektedir (Clarke ve diğerleri, 2008).

2.5. TÜMLEŞTİRME VE YENİDEN SIRALAMA

Tümleştirme ve yeniden sıralama (re-ranking) algoritmaları ilginin yanında çeşitlilik, popülerlik (popularity) gibi özellikleri de algoritmalara dâhil ederek erişim çıktısının sorguyla ilgisini daha da artırmakta, kullanıcı ihtiyaçları açısından daha dengeli ilgi sıralamaları oluşturulmasını sağlamaktadır (Li, Wang ve Bhuiyan, 2022; Maslov ve Redner, 2008; Pinski ve Narin, 1976). Bu amaçla birden fazla algoritmadan elde edilen sıralamalar veri tümleştirme (data fusion) yoluyla birleştirilmekte ve sürece dâhil olan tüm algoritmalardan daha yüksek performans gösteren ilgi sıralamaları elde edilmektedir (Baeza-Yates ve Ribeiro-Neto, 1999; Meng ve diğerleri, 2002). Tümleştirmede temel amaç bilgi kaynaklarının değer yaratacak şekilde birleştirilmesidir (Grant, 1996a, 1996b). Bu bağlamda çok farklı veri yapıları üzerinde çalışan algoritmalar bile bir arada kullanılarak katma değerli bilgi sağlanmaktadır. Örneğin, doğal dil işleme ve görüntü işleme yaklaşımları bir arada kullanılarak görseller ve metin verileri üzerinden bağlamsal ve faydalı bilgiler elde edilmiştir (Oral ve Eryiğit, 2022). Benzer şekilde ortaklaşa filtreleme algoritmasıyla elde edilen sıralamaya kullanıcıların etiketlerinden elde edilen ilgi geribildirimi (relevance feedback) eklenerek geliştirme yapılmıştır (Jin ve diğerleri, 2008).

BIP! Finder arama motorunda ise arama sonucu listelenen makalelerin popülerlik ve etkilerine (influence) göre ağırlıkları da hesaplanmaktadır (Vergoulis ve diğerleri, 2019). BIP! Finder arama motorundaki etki değeri PageRank algoritmasıyla elde edilmiştir. Popülerlik derecesi ise makalelerin atıf alma olasılığının zamanla ilgili olmasından hareketle PageRank algoritmasına içeriğin güncellik bilgisinin de eklendiği TAM-RAM sıralama (time-aware ranking) yöntemi ile hesaplanmıştır (Berberich, Vazirgiannis ve Weikum, 2005).

Birleştirme aşamasında artırımlı (incremental) hesaplamalar kullanılması ise yüksek hesaplama maliyetini önlemektedir (Jin ve diğerleri 2008; Ma, Liu, Yang, Yang ve Li, 2022).

Farklı arama yöntemleri ile erişilen belge kümelerini bir arada kullanmak bilgi erişimin belirsizlik derecesini (degree of uncertainty) azaltmaktadır (Ingwersen, 1996). Bilişsel örtüşmeler, farklı yorumlamalar, zaman, disiplin vb. değişkenlerden kaynaklanan belirsizlikleri azaltmaya yardımcı olmaktadır. Atıflar ise bilişsel örtüşmeler için teorik bir temel sağlaması açısından önemlidir. Bu doğrultuda Larsen (2002), atıf ilişkilerini kullanarak belirsizliğin azaldığı daha yüksek bilgi erişim performansı elde etmiştir.

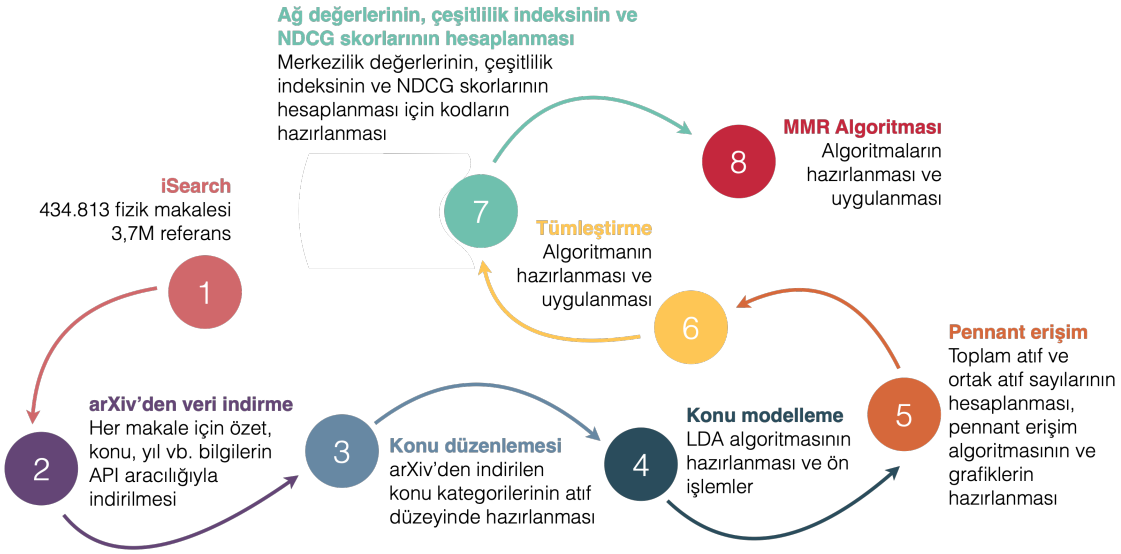
*Mantıksal belirsizlik ilkesi*ne (DBD modeli[10]) göre bir sorgu sonucu erişilecek belgeler sıralanırken her belge için belirsizliğin sayısal bir değerle ilişkilendirilmesi gerekmektedir. Bu değerleri elde etmek için *belirsizlik teorisi* kullanılabilir (van Rijsbergen ve Lalmas, 1996). Teoriye göre belgenin bilgi içeriği çeşitli durumlara göre yapılandırılır. Her bir belge için mevcut durumda bilgi içeriğinin önemini yansıtan bir ağırlık verilir. Belirsizlik derecesinin azaltılması kullanıcıya arama yapılan konu ile ilgili bağlamları ilk sıralarda gösteren bir sıralama ile mümkündür. Bizim çalışmamızda ise hem ortak atıflar aracılığı ile hem de terim sıklıkları ile kurulan bağlantılar bir arada kullanılarak belirsizlik derecesi azaltılmaya çalışmıştır. 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.

[10] Mantıksal belirsizlik ilkesi Dretske (1983), Barwise (1989, 1993) ve Devlin'in (1991) bilgi akışının doğasını ve böyle bir akışa yol açan mekanizmaları tanımladıkları teorilerini temel almaktadır. Bu yüzden belirsizlik ilkesi yazarların isimlerinin baş harflerinden oluşan "DBD modeli" olarak adlandırılmıştır (van Rijsbergen ve Lalmas, 1996).

3. BÖLÜM: YÖNTEM

3.1. GİRİŞ

Bu araştırmada atıf yapılan kaynaklara gömülü konusal ilişkiler ve terim sıklıkları bütünleşik bir şekilde analiz edilerek ilgi sıralamalarına yansıtılmıştır. Bu bağlamda önce iSearch derlemine eklemeler yapılmış ve bu derlem esas alınarak daha önceden tanımlanmış olan 65 sorgu için LDA konu modelleme ve pennant erişim algoritmaları işletilmiştir. Daha sonra algoritmaların çıktıları tümleştirilerek yeni bir sıralama önerilmiştir. Sıralamaların kalitesi için İndirimli Birikimli Kazanç (Discounted Cumulative Gain: DCG) ve Normalleştirilmiş İndirimli Birikimli Kazanç (Normalized DCG: NDCG) skorları ile kapsama ve yenilik oranları incelenmiştir. Bunun dışında LDA ve Pennant Erişim sıralamaları ile Tümleşik sıralamalara MMR (Maximal Marginal Relevance: Maksimum Marjinal İlgi) algoritması uygulanmış ve tüm sıralamalar karşılaştırılarak performans değerlendirmesi yapılmıştır. Karşılaştırma sırasında ağ değerleri ilgi, Shannon çeşitlilik indeksi ve sıralamadaki tekil konu sayıları ise çeşitlilik değeri olarak kullanılmıştır. Bu sayede kelime sıklıklarına dayalı olarak işletilen konu modelleme algoritması ile ortak atıf değerlerini dikkate alan pennant erişim algoritmasının benzer ve farklı yönleri otaya çıkarılmıştır. Bu süreçte izlenen yol genel hatlarıyla Şekil 4'teki gibidir. İlk iki adım verilerin toplanması, üçüncü adım düzenlenmesi ile ilgilidir. Dört, beş ve altıncı adımlarda algoritmalar işletilmekte ve önerilen sıralama oluşturulmaktadır. Son iki adım ise performans değerlendirme ile ilgilidir.[11]



Şekil 4. Derlemin analize uygun hale getirilmesi aşamasındaki adımlar

[11] LDA ve MMR algoritmaları ile ağ değerlerinin hesaplanması için Python, Pennant erişim algoritması için MS Access, şekillerin oluşturulması için ise R, Python ve MS Excel kullanılmıştır. Tezde kullanılan kodlara erişmek için bkz. https://github.com/mugeakbulut/LDA-Pennant_Retrieval/

Toplam 434.813 fizik makalesi ve 3,7 milyondan fazla dâhili referans bulunan iSearch derlemi (1) bu araştırma için iyi bir derlem olsa da bu derlemde çalışmaların yayın yılı bilgileri ile özetleri bulunmamaktadır. Araştırma özelinde hem yayın yılı bilgileri hem de özetler gerekli olduğu için arXiv API[12] aracılığıyla bu çalışmalara dair tüm üst veriler çekilmiştir (2). Ardından konu kategorileri atıf düzeyinde hazırlanmış (3); olasılıksal konu modellemesine hazırlık için ön işlemler (veri temizleme, konu sayısının belirlenmesi vb. gibi) gerçekleştirilmiş ve belgelerin özetleri üzerinde LDA olasılıksal konu modelleme algoritması[13] çalıştırılarak ilgi sıralaması elde edilmiştir (4). Bir sonraki aşamada sorgu ile ilgili olan konu ya da konulardaki belgelere atıf yapan belgelerin referans bilgileri de hesaplamaya dâhil edilerek pennant erişim algoritması uygulanmış (5) ve bu iki sıralama birleştirilerek tümleştirilmiş sıralama elde edilmiştir (6). Sıralamaların değerlendirilmesi için ise DCG ve NDCG skorları, Shannon çeşitlilik indeksi değerleri ve makalelerin merkezilik dereceleri hesaplanmıştır (7). Çalışma kapsamında önerilen tümleştirme algoritması kavramsal olarak MMR algoritmasına çok benzemektedir. Bu yüzden MMR algoritması sağlama amacıyla kullanılmıştır (8).

3.2. iSearch DERLEMİ

Bilgi erişim alanında performans değerlendirme çalışmalarında kullanılan derlemlerde standart olarak belgeler, senaryolar ve ilgi değerlendirmeleri olmak üzere üç bölüm bulunmaktadır (Carevic ve Schaer, 2014). Bu araştırma kapsamında kullanılan iSearch derleminde de bu bölümler yer almaktadır. Lykke ve arkadaşları (2010) tarafından geliştirilen iSearch test derlemi, 1986-2009 yılları arasında arXiv.org'a eklenen 434.813 fizik makalesi, 3.768.409 dâhili referans, 65 sorgu[14] ve her bir sorgu için kullanıcılar tarafından oluşturulmuş ortalama 200 ilgi değerlendirmesi (relevance judgements) içermektedir.[15]

iSearch'te konu (topic) olarak adlandırılan senaryolarda her senaryo için senaryo numarası, kullanıcı numarası, kullanıcının bilgi ihtiyacı, görev tanımı, kullanıcının arkaplan bilgisinin

---

[12] Bkz. https://arxiv.org/help/api

[13] Kaynak kodlar için bkz. https://colab.research.google.com/drive/1uVkPmrP2wngIl9djMegIXYdTnoEkIn6o?usp=sharing#scrollTo=p_RDL28VLXxh

[14] 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.

[15] iSearch aynı zamanda Danimarka Ulusal Kütüphanesinden alınan fizik alanındaki monografik kayıtları da içermektedir. Fakat söz konusu kaynaklara erişim ve eksik bilgileri tamamlama ihtimali olmadığından basılı kütüphane kaynakları (BK) ile ilgili olan kısım çalışma kapsamında değildir.

tanımlandığı paragraf ile arama terimleri ve (varsa) ideal cevap bilgileri yer almaktadır (bkz. Şekil 5). Bu çalışma kapsamında arama terimleri sorgu olarak kulllanılmıştır.

```
<topic>
  <topic_id>003</topic_id>
  <author_id>085</author_id>
  <current_information_need>
    I am looking for information on how to make an on chip flow cytometry
    using an LED as a light source and an APD (Avalanche photodiode) as a
    detector.
  </current_information_need>
  <work_task>
    This was part of a special course conducted in cooperation with DTU and
    Tyndall National Institute in Cork. The aim is to be able to count and
    identify particles in a micro fluidic chip using cytometry. This could
    be from scattering of light or emission from fluorescent particles.
  </work_task>
  <background_knowledge>
    I have already written a report on this subject. So the information is
    to get some extra knowledge on the subject.
  </background_knowledge>
  <ideal_answer>
  </ideal_answer>
  <search_terms>
    Flow cytometry, micro fluidic chip, APD, LED.
  </search_terms>
</topic>
<topic>
```

Şekil 5. iSearch senaryolarında tanımlı alanlar

Aynı zamanda bir açık arşiv olan arXiv ön baskı (preprint) arşivi 1991 yılında erişime açılmıştır.[16] iSearch derlemi her ne kadar performans değerlendirmesi yapmaya olanak sağlayacak veriler içerse de bu çalışma özelinde kullanılacak bazı değişkenler için derlemin genişletilmesi gerekmiştir. Örneğin, çeşitlilik değerlerinin hesaplanmasında arXiv'de yazarlar tarafından belirlenen konu kategorilerine ihtiyaç vardır. Benzer şekilde bazı makaleler için LDA konu modelleme algoritması için gereken özet bilgileri bulunmamaktadır.[17] Bu yüzden Şekil 6'da sarı ile vurgulanan alanların tümü arXiv'den indirilerek veri setine eklenmiştir. [18]

arXiv'in konu sınıflamasını detaylı olarak incelemekte yarar vardır. arXiv'e makale yüklenirken her makale için en az bir konu belirlenmesi gerekmektedir. Konu kategorileri girilirken önce

[16] Fakat arXiv.org'daki fizik makalelerinden oluşan iSearch derleminde 1991 yılı öncesine ait üç makale yer almaktadır. Bu makalelerden iki tanesi (Ginsparg ve Glashow, 1986; Ginsparg, 1988) arXiv'in kurucusu Paul Ginsparg'ın çalışmaları olup arşive deneme amacıyla eklendiği düşünülmektedir.

[17] Derlemde makaleler PF (tam metin PDF) ve PN (metadata) olarak sınıflanmıştır ve sadece PF olanların özet bilgileri bulunmaktadır.

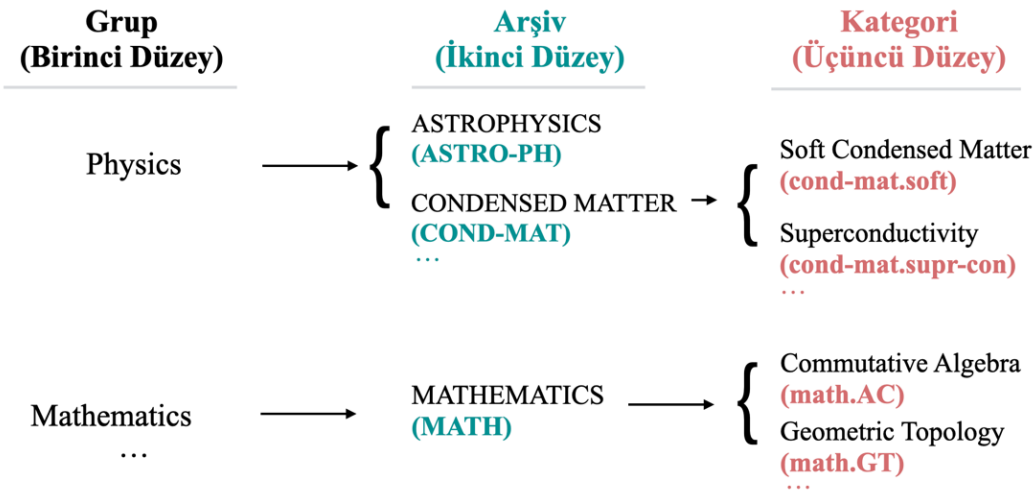
[18] Kodlar için bkz. https://mugeakbulut.com/phd/codes/iSearch_verilerini_Arxivden_indirme.py

temel kategori (primary category) seçilmektedir. Sonrasında isteğe bağlı olarak çapraz listeler (cross list) seçeneği ile alt konu başlıkları belirlenerek o makalenin başka hangi başlıklar altında listeleneceği tanımlanmaktadır.

Şekil 6. arXiv'den indirilerek derleme eklenen alanlar

arXiv'de konular genelden özele *grup* (birinci düzey), *arşiv* (ikinci düzey) ve *kategori* (üçüncü düzey) olarak sınıflandırılmakta ve sekiz adet *grup* (Computer Science, Economics, Electrical Engineering and Systems Science, Mathematics, Physics, Quantitative Biology, Quantitative Finance ve Statistics) bulunmaktadır. Bu *gruplar* en genel hatları ile çalışmanın konusunu tanımlamaktadır. Grupların altında üç ile 40 arasında *arşiv* ve son olarak *kategoriler* bulunmaktadır.[19] Bu çalışma kapsamında yapılan analizlerde üçüncü düzey temel konu *kategorileri* dikkate alınmıştır (bkz. Şekil 7).

[19] arXiv konu taksonomisi için bkz. https://arXiv.org/category_taxonomy. Görselleştirme için bkz. http://mugeakbulut.com/phd/gorsellestirme/ArXiv_konu3.png



Şekil 7. arXiv konu taksonomisi

iSearch derlemindeki makalelerin tamamına en az bir, %28'ine iki, %8'ine üç, %2'sine ise dört konu başlığı atanmıştır. Beş ve daha fazla konu başlığı atanan makalelerin oranı %1'den azdır.

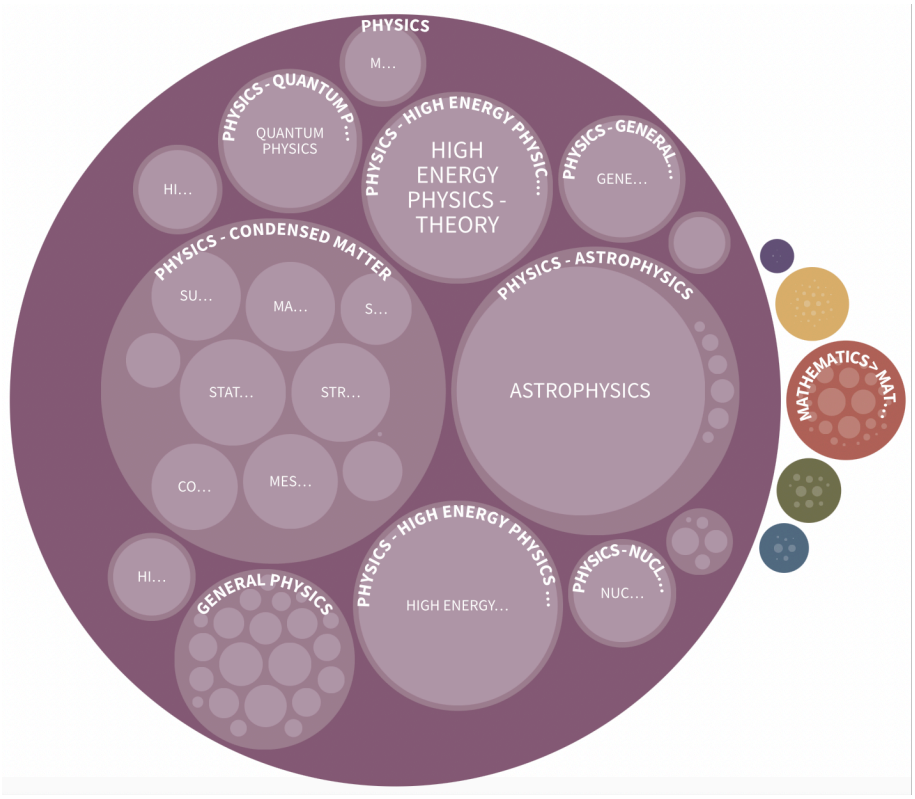
Araştırma kapsamında kullanılan iSearch derlemi fizik makalelerini içerdiği için arXiv'de tanımlanan tüm alanları kapsamamaktadır (Şekil 8).[20]

Konular grup bazında (birinci düzey) incelendiğinde yayınların %98'inin fizik grubuna, %2'sinin ise *Computer Science*, *Mathematics*, *Quantitative Biology*, *Quantitative Finance* ve *Statistics* gruplarına dâhil olduğu görülmektedir (Tablo 1).[21] Fizik grubu dışındaki makalelerin ikincil konuları fizik olarak tanımlandığı için iSearch derleminde yer almaktadır.[22]

[20] iSearch etkileşimli grafik için bkz. http://mugeakbulut.com/phd/gorsellestirme/bubble.html

[21] Sadece fizik konu grubunun etkileşimli grafiği için bkz. http://mugeakbulut.com/phd/gorsellestirme/isearch_classification.html

[22] iSearch'te, arXiv'de tanımlanmış olan Electrical Engineering and Systems Science ve Economics grupları altında sınıflanmış herhangi bir makale bulunmamaktadır.



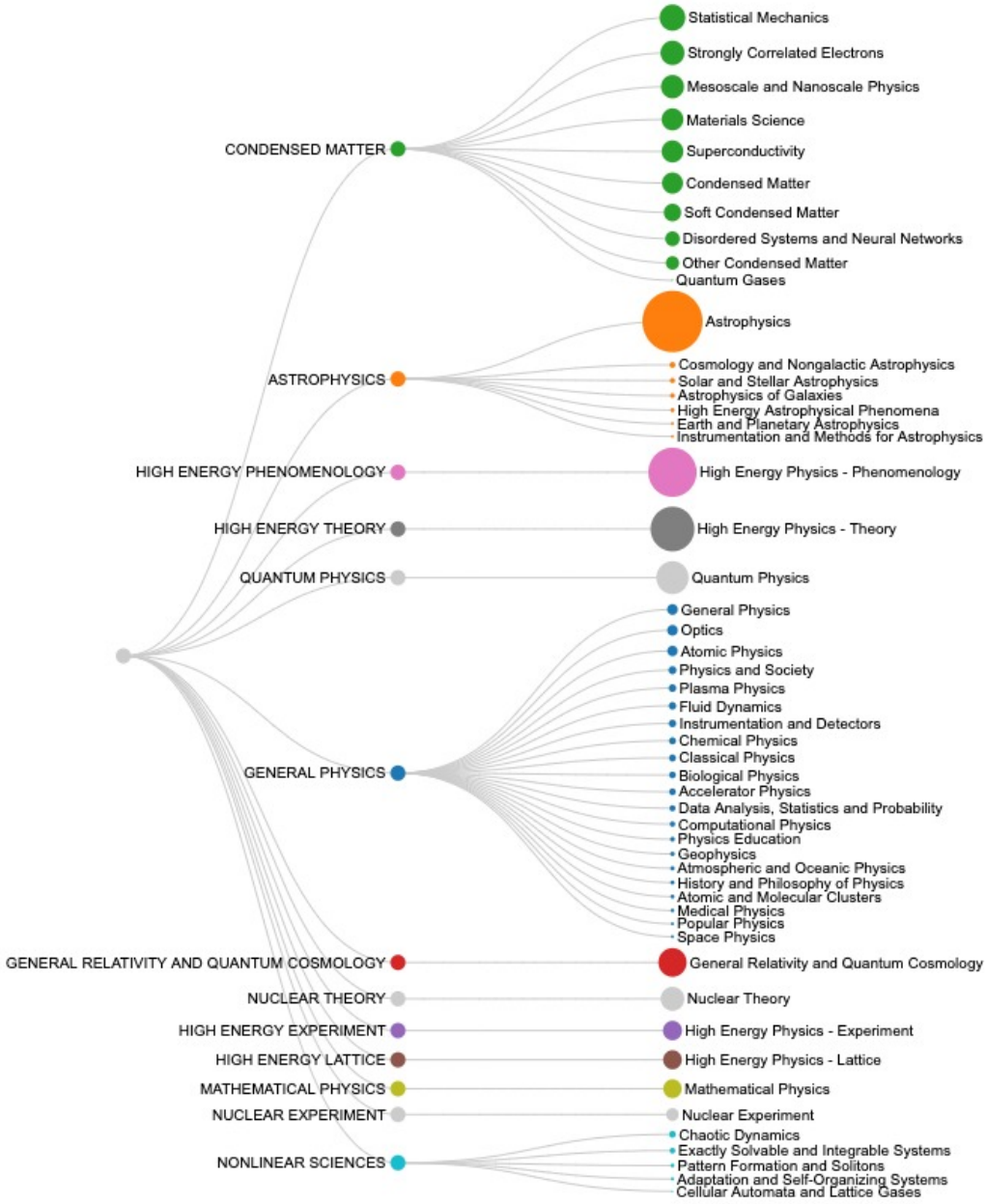
Şekil 8. iSearch derleminin konu taksonomisi

Tablo 1. iSearch derlemindeki makalelerin konulara göre dağılımı

Grup Adı	N	%
Physics	427.258	98,26
Mathematics	6081	1,40
Quantitative Biology	759	0,17
Computer Science	416	0,10
Quantitative Finance	286	0,07
Statistics	13	0,00
Toplam	434.813	100,00

Sadece fizik konu grubuna ait arşivler ve onların altında listelenen konu kategorileri ise Şekil 9'da görselleştirilmiştir. Makalelerin fiziğin alt konularına dağılımı homojen değildir (noktaların büyüklüğü her bir kategorideki makale sayılarını temsil etmektedir). Alandaki gelişmeler nedeniyle arXiv konu kategorilerinde güncellemeler yapılması da bu heterojen yapının en önemli sebeplerindendir. Örneğin, günümüzde altı temel konu kategorisi bulunan arXiv'de 2000 yılında fizik, matematik, doğrusal olmayan bilimler ve bilgisayar bilimleri olmak üzere dört temel konu

sınıfı bulunuyordu (McKiernan, 2000, s. 128). Şekil 9'da Astrophysics alt konusunun diğerlerine göre daha fazla makale içerdiği görülmektedir (tüm makalelerin %22'si). Bunun nedeni Astrophysics alanındaki gelişmeler nedeniyle alt sınıflama yapılması gereksinimi doğması ve 2004 yılında altı yeni alt kategori tanımlanmasıdır (Traag, 2021). Bu tarihe kadar olan makaleler ise doğrudan Astrophysics 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, *Machine Learning* konu sınıfı hem *Statistics* (stat.ML) hem de *Computer Science* (cs.LG) arşivi altında listelenmektedir.



Şekil 9. iSearch derlemindeki makalelerin konu dağılımı

Çalışma kapsamında kullanılan pennant erişim algoritmasının performansı toplam atıf ve ortak atıf sayıları ile doğrudan ilgili olduğu için iSearch derleminin ortak atıf oranları da hesaplanmıştır. iSearch derlemi için atıf ağı yoğunluğu 0,0021'dir. Diğer bir deyişle 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 makalelere yapılan ortalama ortak atıf sayısı ise ikidir (ortanca=1).[23]

3.3. OLASILIKSAL KONU MODELLEMESİ

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, Son ve Wang, 2020). Bu araştırma kapsamında konu modellemesine hazırlık için özel karakter ve sayıların metinden atılması, büyük harflerin küçük harflere çevrilmesi,[24] metni cümle, kelime gibi birimlere ayırma (tokenization), dur kelimelerinin (stop words) atılması ve kelimelerin eklerinin kaldırılarak morfolojik köklerinin elde edilmesi (stemming) vb. gibi ön işlemler 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, bibliyografik kayıtta 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, Glänzel ve Scharnhorst, 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

---

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

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

hesaplayarak en uygun konu sayısını saptamayı amaçlamaktadır. Bazı yaklaşımlarda skorun yüksek olması, bazılarında ise düşük olması beklenmektedir (Carroll, 2018). iSearch derlemindeki yayınlar için konu sayısını belirlemek amacıyla dört ölçev temel alınmış (Arun, Suresh, Madhavan ve Murthy, 2010; Cao, Xia, Li, Zhang ve Tang, 2009; Deveaud, SanJuan ve Bellot, 2014; Griffiths ve Steyvers, 2004) ve buna göre oluşturulan kod (Nikita, 2020) iSearch derlemine uyarlanmıştır.

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)[25] 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 distributions) simetrik KL ıraksaklığı ve $L * M_2$ vektörünün dağılımı hesaplanır. Derlemdeki (C) belge sayısı d, terim (kelime) dağarcığının boyutu ise w ile temsil edilmektedir (Formül 4).

$$C_{d*w} = M1_{d*t} \times Q_{t*w} \tag{4}$$

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 (Formül 5) (Cao ve diğerleri, 2009, s. 1778).

$$korelasyon(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \quad \sqrt{\sum_{v=0}^{V} (T_{jv})^2}} \tag{5}$$

Formül 5'te $korelasyon(T_i, T_j)$ değeri küçüldükçe konular da birbirinden daha bağımsız ve ayrık olur. Konu yapısının kararlılığını ölçmek için her konu çifti arasındaki ortalama kosinüs mesafesi

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

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[26] 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 $argmax[n]$ operatörü tanımlanır (Formül 6). 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).

$$W_k = \underset{w}{argmax}[n]\, \phi_{k,w} \tag{6}$$

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 Formül 7'ye göre hesaplanır (Deveaud ve diğerleri, 2014, s. 67).

$$\widehat{K} = \underset{K}{argmax} \frac{1}{K(K-1)} \sum_{(k,k^t)\in\mathbb{T}_K} D(k\|k') \tag{7}$$

Formül 7'de 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.[27]

LDA konu modelleme algoritmasına girilecek konu sayısını belirlemek için kullanılan ve yukarıda açıklanan dört ölçev ve her ölçev için verilen formüller kullanılarak bu araştırmada konu sayısının nasıl hesaplandığı R ile yazılan kodlarda daha ayrıntılı olarak verilmektedir.[28]

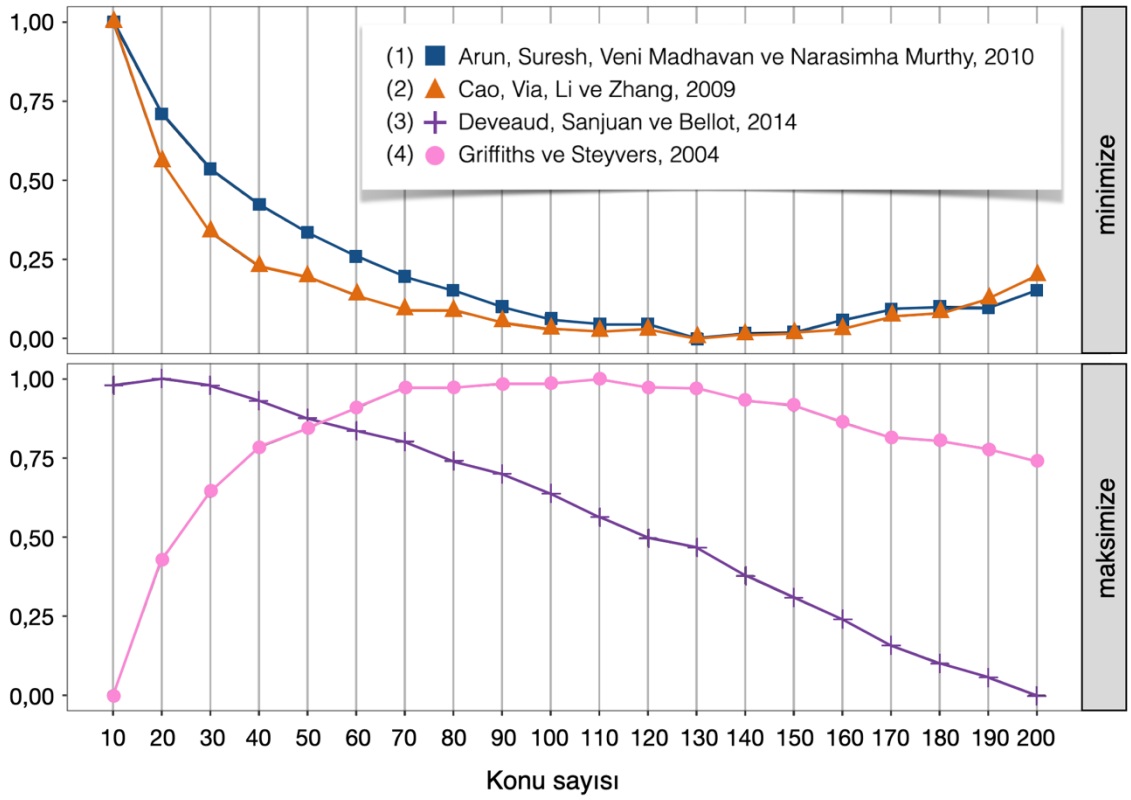
---

[26] "Maksimize etmek" terimi "mümkün olduğu kadar büyütmek, artırmak" anlamında kullanılmıştır. Bkz. Dictionary.com. https://www.dictionary.com/browse/maximize.

[27] Bu ölçevin teknik ayrıntıları için bkz. Griffiths ve Steyvers (2004).

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

LDA algoritmasına girilecek konu sayısını belirlemek için kullanılan tüm ölçevlerde skorlar 0 ile 1 aralığındadır. Dört yaklaşımın bir arada gösterildiği Şekil 10'da "maksimize" altında gösterilen ölçevler skorlarının yüksek olması beklenen, "minimize" altında gösterilenler 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 10). 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, 2019; Bonaccorsi, Melluso ve Massucci, 2022; Guillemette, Simms, Zhou ve Mills, 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.[29]



Şekil 10. iSearch derlemine en uygun konu sayısının belirlenmesi

[29] LDA algoritmasında benzerlik ölçüsü olarak iki olasılık dağılımı arasındaki mesafeyi ölçmek için kullanılan Jensen-Shannon mesafesi (Jensen-Shannon distance) tercih edilmiştir. Literatürde Information radius (yani iRad) veya ortalamaya olan toplam uzaklık (total divergence to average) olarak da geçmektedir.

LDA algoritması uygulanırken, özellik sayısı (yani sabit kelime boyutu), α ve β Dirichlet ön parametrelerine ince ayar yapılabilmektedir (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 (Wallach, Mimno ve McCallum, 2009; Zhang, Zeng, Yuan, Rao ve Yan, 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, Dascalu ve McNamara, 2017; Deerwester, Dumais, Furnas, Landauer ve Harshman, 1990, s. 394). Dolayısıyla bu araştırmada kullanılan derlem büyük ölçekli bir derlemdir. Algoritma çalıştırılırken orta düzey model için ön tanımlı parametreler kullanılmıştır.[30]

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). Daha önce de belirtildiği gibi (s. 13), Bayes yaklaşımında parametreler önsel 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.

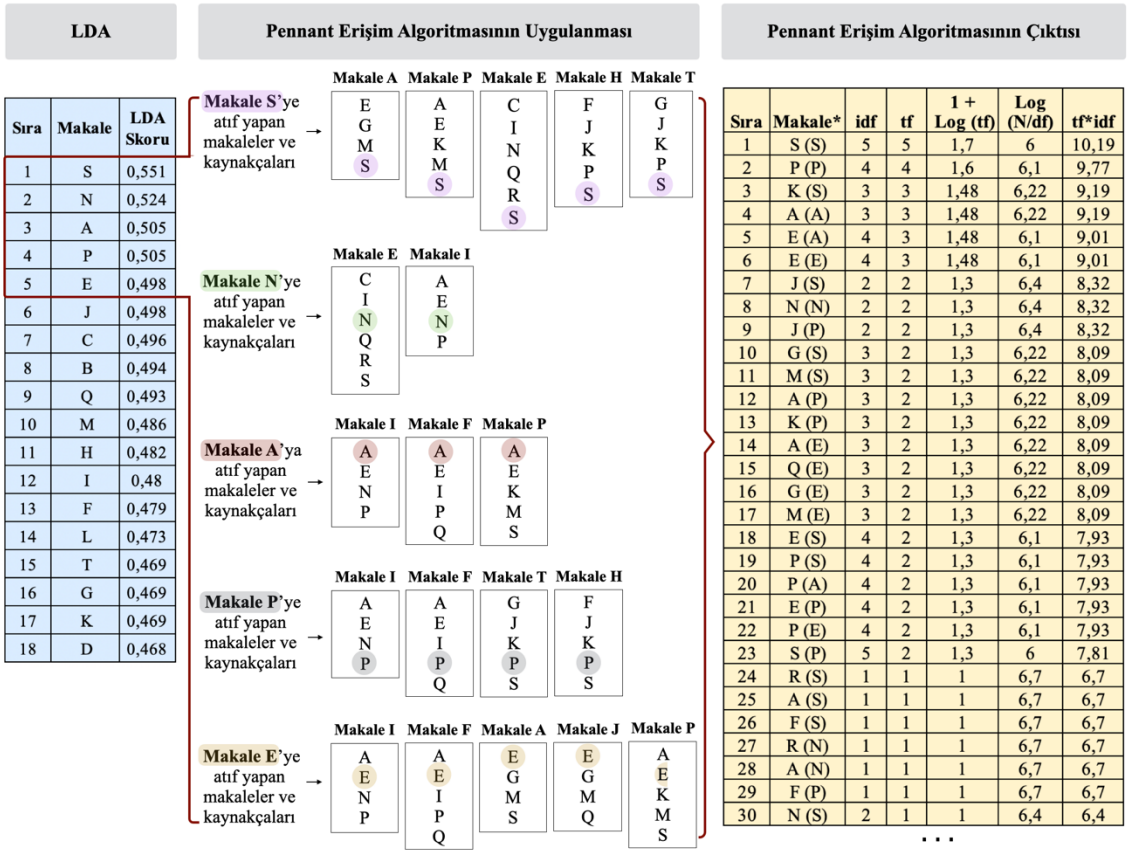
3.4. PENNANT ERİŞİM

Pennant algoritmasının işletilebilmesi için bir veya daha fazla çekirdek makaleye ihtiyaç vardır. Bunun için de LDA algoritmasının eriştiği ilk beş kaynak çekirdek kaynaklar olarak kullanılmıştır. Dolayısıyla pennant, tümleşik ve MMR algoritmalarına temel olan ve ilgi sıralaması iyileştirme işleminin yapıldığı kaynaklar için LDA algoritmasının eriştiği kaynaklar esas alınmıştır.

---

[30] 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.

Pennant erişim algoritmasının uygulama aşamalarının örnek gösterimi Şekil 11'de verilmektedir. Elimizdeki derlemde 20 makale olduğunu (Makale A-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 makalelere atıfta bulunan makalelerin kaynakçalarında yer alan her bir makale pennant sıralamasında (LDA+Pennant sıralaması) yer almaktadır. Çünkü çekirdek makalelerden en az biriyle bir ve/veya daha fazla ortak atıf almıştır. Pennant sıralamasında makaleler *tf* (ortak atıf) *ve idf* (toplam atıf) 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 sıralamaya 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.



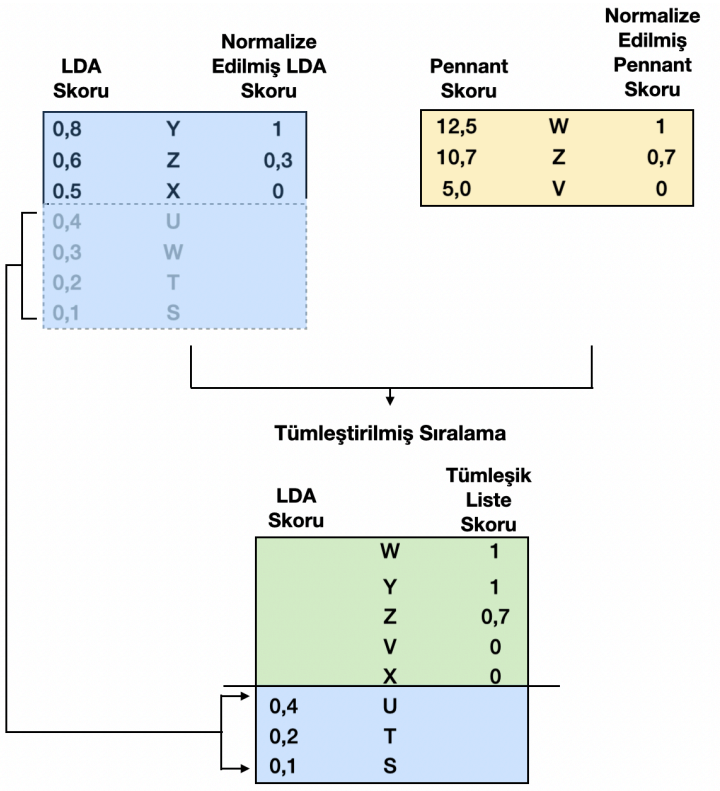
Şekil 11. Pennant erişim algoritmasının uygulama aşamaları

Not: Pennant erişim algoritmasının çıktısında makale sütununda yer alan parantez içindeki değer, ilgili makalenin hangi çekirdek makale aracılığı ile sıralamaya girdiğini göstermektedir.

iSearch derleminde 3,7 milyondan fazla atıf bulunduğu için, hesaplamanın daha hızlı yapıldığı SQL tabanlı bir platform tercih edilmiştir. Hazırlanan MS Access uygulaması[31] ile pennant erişim hesaplamaları yapılmıştır.[32]

3.5. İLGİ SIRALAMALARININ TÜMLEŞTİRİLMESİ

Ortak atıf sayısı her çekirdek makale için farklı olduğundan pennant erişim algoritmasının ürettiği ilgi sıralamalarının uzunluğu da her sorgu için farklıdır. Öte yandan LDA algoritması derlemdeki özeti olan tüm çalışmalar için bir sorgu-belge benzerlik değeri hesaplamaktadır. Bu iki sıralamanın birleştirilmesi sırasında önce pennant sıralama listesi esas alınarak iki sıralamanın uzunlukları eşitlenmiş, ardından eşitlenmiş uzunluktaki sıralamalardaki değerler normalize edilip[33] (min-max normalizasyonu) tümleştirilmiştir (bkz. Şekil 12).



Şekil 12. Sıralama tümleştirme

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

[32] İlgili değerlerden pennant diyagramı çizdirmek için ilgili kodlara http://mugeakbulut.com/phd/codes/pennant/pennant_R/pennant_not_norm.R adresinden erişilebilir.

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

Normalizasyon işlemi Formül 8'e göre yapılmıştır. Herhangi bir değerin normalize edilmiş yeni değeri hesaplanırken, o değerden en küçük değer çıkarılmakta ve en büyük değer ile en küçük değer arasındaki farka bölünmektedir.

$$x_{normalize} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

Tümleştirme aşamasında her iki sıralamada da olan ortak makaleler, normalize edilmiş değerlerinin en yüksek olduğu skorla tümleşik sıralamaya dâhil olmaktadır (örneğin Şekil 12, Makale Z için normalize edilmiş pennant skoru 0,7). Derlemde özet bilgisi olmayan makaleler varsa (Şekil 12, örneğin Makale V) LDA skoru hesaplanamamaktadır. Fakat bu makaleler için ortak atıf verisi varsa pennant erişim yöntemi ile tümleştirilmiş sıralamada bu makaleler de yer almaktadır. Tümleşik sıralama uzunluğu tamamlandığında ise LDA sıralamasından artan çalışmalar skoru en yüksek olandan başlayarak (Şekil 12'deki LDA sıralamasındaki U makalesinden itibaren) tümleşik sıralamanın arkasına eklenmektedir.

3.6. İLGİ SIRALAMALARININ KİŞİSELLEŞTİRİLMESİ

Kişiselleştirme aşamasında ise ilgi sıralaması kullanıcının ihtiyacına göre en ilgili belgeleri sıralamada önceleyecek ya da farklı konudaki çalışmaları da içerecek 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 sıralamada üst sıralarda yer alması sağlanmış ya da her ikisini de mümkün olduğu kadar artıran bir sıralama oluşturulmuştur. Kişiselleştirme aşamasında her iki sıralamada da bulunan ortak makaleler; tümleştirme öncesinde hesaplanan ve normalize edilmiş değerlerde en yüksek olan skorlarıyla değil, ağırlıklı olması istenen algoritmadaki değerleri ile sıralamaya dâhil edilmektedir (örneğin Şekil 12, Makale Z için LDA ağırlıklı bir sıralama isteniyorsa normalize edilmiş 0,3 değeri, pennant erişim ağırlıklı bir sıralama isteniyorsa 0,7 değeri kullanılmaktadır).

Daha detaylı olarak pennant erişim ağırlıklı sıralamada, *A*, *B* ya da *C* sektörlerinde yer alan makalelere öncelik verilerek, çekirdek makalelerin ardılları, akranları ve öncülleri ilk sıralara yerleştirilmektedir (bkz. Bölüm 2.3). Konusal ilgi açısından da benzer şekilde sektörlerdeki makalelere ağırlık verilerek çekirdek makaleler ile spesifik konudaki çalışmalar, ilgili konu üzerine inşa edilen ama çok spesifik olmayan çalışmalar, çekirdek makaleden türeyen çalışmalar ya da çekirdek makalenin konusuyla ilgili ama genel konudaki çalışmaların ilk sıralarda erişildiği sıralamalar oluşturulmaktadır. Bu çalışmaların etkileşimli pennant diyagramları üzerinden izlenebilmesi de mümkündür.

3.7. PERFORMANS DEĞERLENDİRME

Çalışma kapsamında performans değerlendirme işlemleri farklı detay düzeylerinde gerçekleştirilmiştir. Bu bağlamda, sıralamalar hem genel olarak karşılaştırılmış hem de makale düzeyinde ilgi ve çeşitlilik değerlerine göre yorumlanmıştır. Ek olarak makalelerin tek bir konuya indirgenmesinin yeterli olmayabileceği düşünülerek kaynakçalarındaki çalışmaların konuları da dikkate alınarak incelemeler yapılmıştır. Erişilen makalelerin sorguyla ilgileri ya da sıralamanın çeşitliliğinin değerlendirilmesi aşamasında da MMR algoritması sağlama yapmak amacıyla kullanılmıştır.

3.7.1. DCG, NDCG Değerleri ile Kapsama ve Yenilik Oranları

Sıralama kalitesinin ölçümü ile ilgili en sık kullanılan ölçevler DCG ve NDCG ölçevleridir. Bu araştırma kapsamında da bu ölçevlerden yararlanılmıştır. DCG ölçevi, ilgi sıralamasının üst sıralarında erişilen ögelerin kalitesini ölçmektedir (Cossock ve Zhang, 2008, s. 5140). Burada sorguyla çok ilgili makalelerin sıralamada daha üstlerde konumlanmasının daha kullanışlı olacağı ve bu makalelerin marjinal ilgili makalelerden daha faydalı olacağı, marjinal ilgili olanların ise alakasız makalelerden daha faydalı olacağı şeklinde iki temel varsayım vardır. Çünkü kullanıcının zaman ve çaba harcayarak bu belgeleri inceleme olasılığı daha düşüktür (Järvelin ve Kekäläinen, 2002).

Öte yandan ilgi sıralamalarının uzunluğu sorguya bağlı olarak farklılık göstermektedir. Dolayısıyla sıralama performansı tek başına DCG kullanılarak tutarlı bir şekilde ölçülemez. Bu sorunu çözmek amacıyla seçilen bir değer için her konumdaki birikimli kazancın sorgular arasında normalleştirilmesi yoluna gidilmiştir. Bu amaçla tüm ilgili belgeler ideal ilgilerine göre sıralanarak (Ideal DCG, IDCG) hesaplamaya dâhil edilir ve mümkün olan maksimum DCG değeri üretilmiş olur.

Daha ilgili belgelerin sıralamada üst sıralarda yer alması sıralama puanının daha yüksek olması anlamına gelir. Sıralama işlevi ne olursa olsun, logaritmik bir indirim benimseyen NDCG, sıralanacak makale sayısı sonsuza giderken 1'e yakınsamaktadır (Wang ve diğerleri, 2013, s. 13, Şekil 1).

DCG ölçevi Formül 9'a göre hesaplanmaktadır.[34]

[34] Formüllerdeki *i* değeri sıralamadaki pozisyon, *p* kalite ölçümü yapılacak sıralamadaki toplam makale sayısı, *rel* ise ilgi değeri anlamına gelmektedir.

$$DCG = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \quad (9)$$

İdeal DCG (IDCG) ise bilinen tüm aktiflerin tahmin listesinde en iyi sıralandığı bir sıralama için hesaplanan DCG'dir (Formül 10). Her bir makale için ideal sıralamadaki ilgi değeri olarak ağ değerleri kullanılmaktadır.

$$IDCG = \sum_{i=1}^{|REL|} \frac{2^{rel_{i-1}}}{\log_2(i+1)} \quad (10)$$

NDCG değeri, DCG'nin IDCG'ye bölünmesiyle elde edilir (Formül 11) ve 0-1 aralığında bir değere sahiptir (Schuler ve diğerleri, 2022).

$$NDCG = \frac{DCG}{IDCG} \quad (11)$$

DGC, NDCG gibi yaygın olarak kullanılan değerlendirme ölçütleri her bir belgenin uygunluğunun diğer belgelerden bağımsız olarak ayrı ayrı değerlendirilebileceğini varsaymaktadır (Järvelin ve Kekäläinen, 2002). Daha detaylı analizler için ise tüm ağ dikkate alınarak derece merkeziliği değerleri hesaplanmış ve değerlendirmeler bu değerlere göre yapılmıştır.

İlgi sıralamalarını genel olarak değerlendirmek için kapsama (coverage) ve yenilik (novelty) oranları karşılaştırılmıştır. Kapsama oranı algoritmanın ilgili olduğu bilinen makalelere erişim performansını, yenilik oranı ise bir algoritma tarafından erişilen fakat ilgili olduğu daha önceden bilinmeyen makalelere erişim performansını ölçmektedir (Kaminskas ve Bridge, 2016). Her iki oranın hesaplanması için ilgili olduğu bilinen konulara ihtiyaç vardır. Çalışma kapsamında her bir sorgu için çekirdek makalelerin kaynakçalarındaki makalelerin konuları *ilgili konu* olarak kabul edilmiştir.

3.7.2. İlgi Değerlerinin Hesaplanması

Daha önce de değinildiği gibi, iSearch derleminin en önemli özelliklerinden birisi her bir sorgu için uzmanlar tarafından derecelendirilmiş ilgi değerlendirmelerini içermesidir. Söz konusu ilgi değerlendirmeleri üç ayrı fizik bölümünden öğretim görevlileri ile doktora ve yüksek lisans öğrencileri tarafından oluşturulmuştur. Fakat ilgi değerlendirmesi içeren yayınların oranı çok düşük (yaklaşık %2) olduğu için bu oran sıralamaları ilgi açısından değerlendirmek için yeterli değildir. Sorguların yarısı için ilgili olarak işaretlenen ve ilgi değeri belirlenen çalışma sayısı 20'den azdır. Bazı sorgular için (örneğin sorgu 5) ise hiç ilgi değerlendirmesi yapılmamıştır. Bu yüzden sıralamaların 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) değeri paragraf özetlemek (Marujo ve diğerleri, 2017; Ribeiro ve de Matos, 2011) ve ağ analizini geliştirmek (Giustolisi, Ridolfi ve Simone, 2020) için kullanılmaktadır. İlgi olarak merkezilik değeri 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 belirlenmiştir. Böylece iSearch derlemindeki her makale için ilgi olarak merkezilik hesaplamalarına dayanan ilgi değerleri elde edilmiştir.[35] Merkezilik değerleri NetworkX Python paketi kullanılarak hesaplanmıştır.[36] Öte yandan sıralamaların çeşitlilik oranı ise içerdikleri makalelerin arXiv temel konu kategorileri incelenerek belirlenmiştir.

3.7.3. Çeşitlilik Değerlerinin Hesaplanması

Çeşitlilik indekslerinde bir sistemin heterojenlik düzeyi çeşitliliğinin bir ölçüsü olarak ele alınmaktadır (Carpi ve diğerleri, 2019, s. 1; Jost, 2006; Rafols ve Meyer, 2010). En yaygın çeşitlilik ölçüsü olan Shannon çeşitlilik indeksi[37] de entropi kavramına dayanmakta ve genellikle tür çeşitliliğini ölçmek için kullanılmaktadır (Lande, 1996). Shannon çeşitlilik indeksi bu araştırmada da algoritmaların çıktısı olan sıralamalardaki konu çeşitliliğinin ölçümü için kullanılmıştır.

Shannon çeşitlilik indeksi Formül 12'ye göre hesaplanmaktadır.

$$\textit{Shannon Çeşitlilik İndeksi} = -\sum[(p_i) * \log(p_i)] \quad (12)$$

Birinci adımda, herhangi bir sorgu için belirli bir konunun bulunma olasılığı hesaplanır (p_i). Bunun için belirli bir konunun belirli bir sorgu için erişilen sıralamadaki geçiş sıklığı sıralama uzunluğuna (ya da kesme noktasına) bölünür. Ardından bu değerin iki tabanına göre logaritması

[35] Toplam 65 sorgudan korelasyon hesaplaması için gerekli koşulları sağlayan 57 sorgu için merkezilik derecesi ve uzmanların belirlemiş olduğu sorgu-belge benzerliği ilgi sonuçlarının korelasyonu $r = 0{,}66$'dır.

[36] Kodlar için bkz. https://tinyurl.com/pebwtkjr

[37] Shannon bu indeksi Wiener'in çalışmaları üzerine inşa etmiştir. Bu yüzden indeks "Shannon-Wiener indeksi" olarak da kullanılmaktadır. Literatürde "Shannon – Weaver indeksi" şeklinde kullanımı da mevcuttur. Fakat bu karışıklığın nedeni muhtemelen Shannon'ın Weaver ile ortak yazar olduğu bir kitap (Shannon ve Weaver, 1949) yayımlamış olmasıdır (Spellerberg ve Fedor, 2003).

hesaplanır ve ilk bulunan değer ile çarpılır. Son olarak tüm konular için elde edilen bu değerler toplanarak -1 ile çarpılır ve Shannon çeşitlilik indeksi değeri elde edilmiş olur.

LDA algoritmasında her ne kadar kelimelere dayanan konu bazlı bir analiz yapılsa da, bu konular arXiv'de entellektüel olarak belirlenen 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 kullanılarak elde edilen Shannon çeşitlilik indeksi değerleri "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.

3.7.3. Maksimum Marjinal İlgi Algoritmasının İlgi Sıralamalarına Etkisi

Çeşitlilik oranı yüksek sıralamaların 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 marjinal ilgi düzeyinin yüksek olması için hem sorguyla alakalı olması hem de önceden ilgili olarak seçilen makalelerle *minimum* benzerlik göstermesi gerekmektedir. MMR yaklaşımı bilgi erişim performansını %8 ile %17 oranında artırmaktadır (Yang, Ji ve Leong, 2007).

MMR algoritması benzer cümleleri ya da kaynakları elediği için (başka bir deyişle konunun çeşitli yönlerini içeren farklı cümle veya kaynaklara eriştiği için) metin özetleme uygulamalarında marjinal ilgi değeri daha yüksek sıralamalar oluşturmak için kullanılmaktadır. MMR algoritması kavramsal olarak bu çalışmada önerilen ve marjinal ilgili makaleleri sıralayan tümleşik sıralama algoritmasına benzemektedir (bkz. Tablo 2). Her iki algoritma da sorguyla ilgili ama marjinal kaynakların da yer aldığı bir sıralama 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 marjinal makaleleri sıralamaya 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 listeyi yeniden sıralamaktadır. Tümleştirme algoritması ise kelime sıklıklarına göre hesapladığı ilgi skorları ile atıflardan yola çıkarak hesapladığı belge-belge

benzerliklerini kullanarak her iki sıralamada da en yüksek skorları olan belgeleri üst sıralarda gösterecek şekilde bir sıralama yapmaktadır.

Tablo 2. MMR ve tümleştirme fonksiyonunun karşılaştırılması

	Tümleştirme Algoritması	MMR Algoritması
Amaç	LDA ve pennant erişimde en yüksek ilgi değerleri alanları 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.[38] MMR algoritmasının LDA, pennant erişim ve tümleştirme algoritmaları uygulanarak ayrı ayrı elde edilen ilgi sıralamalarına etkileri incelenmiştir. Etki oranlarına bakılarak hangi algoritmanın hangi özellikleri öne çıkardığı saptanmıştır.

[38] Tümleştirilmiş ilgi sıralamasına uygulanan MMR algoritması kodlarına https://colab.research.google.com/drive/1dESqgDRL6WfyCSDgHxAa1kPDoHKBFK54 adresinden erişilebilir.

4. BULGULAR VE YORUM[39]

Araştırmanın bu kısmında birinci bölümde tanımlanan araştırma sorularını cevaplamak ve hipotezleri test etmek amacıyla iSearch derlemindeki 65 sorgu için çalıştırılan algoritmaların çıktıları karşılaştırılmış ve ilgi sıralamalarına dair bulgular rapor edilmiştir. Bunun dışında önerilen algoritma için 42. sorgu özelinde yeniden sıralama yapılarak kullanıcıların ihtiyacına uygun sıralama oluşturma aşamaları incelenmiştir.

4.1. ALGORİTMALARIN ÖNE ÇIKARDIĞI ÖZELLİKLER

Çalışma kapsamında geliştirilmiş en uygun sıralama, hem sorguyla ilgili makalelerin üst sıralarda yer aldığı hem de çeşitlilik oranı yüksek sıralama olarak kavramsallaştırılmıştır. Bu bağlamda önce ilgi ve çeşitlilik oranları incelenerek algoritmaların hangi özellikleri ön plana çıkardıkları saptanmıştır. Ardından konulara göre daha detaylı analizler yapılmıştır.

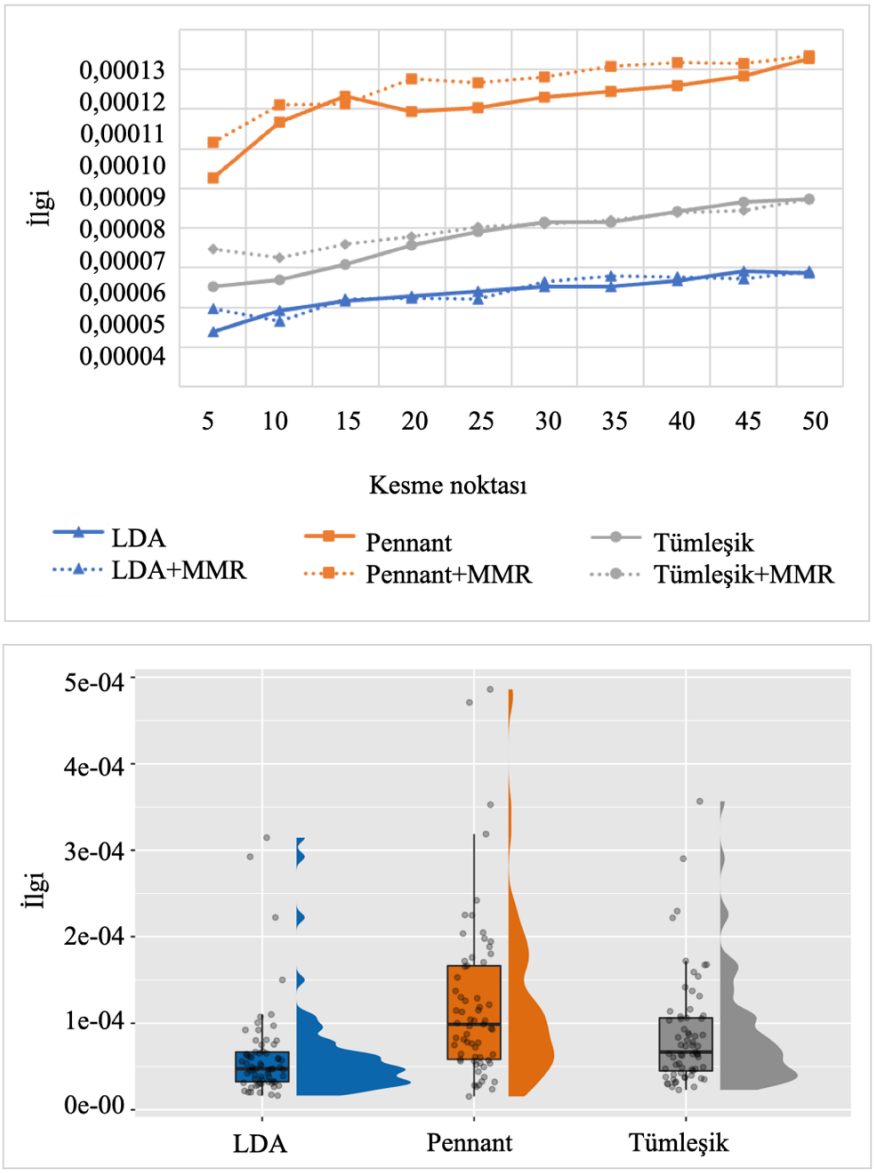
4.1.1. Algoritmaların İlgi Değerleri ve Konu Çeşitliliğine Göre Karşılaştırılması

İlk aşamada LDA ve pennant erişim sıralamaları ile tümleşik sıralamaların ilgi ve çeşitlilik değerleri incelenmiştir. Şekil 13 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 14 ise farklı algoritmaların çeşitli kesme noktalarında eriştikleri makalelerin Shannon çeşitlilik indeksi ortalamasını göstermektedir.

Şekil 13'te 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 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 13) ve çeşitliliği (Şekil 14) 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 sıralama (tümleşik) elde edilmektedir.

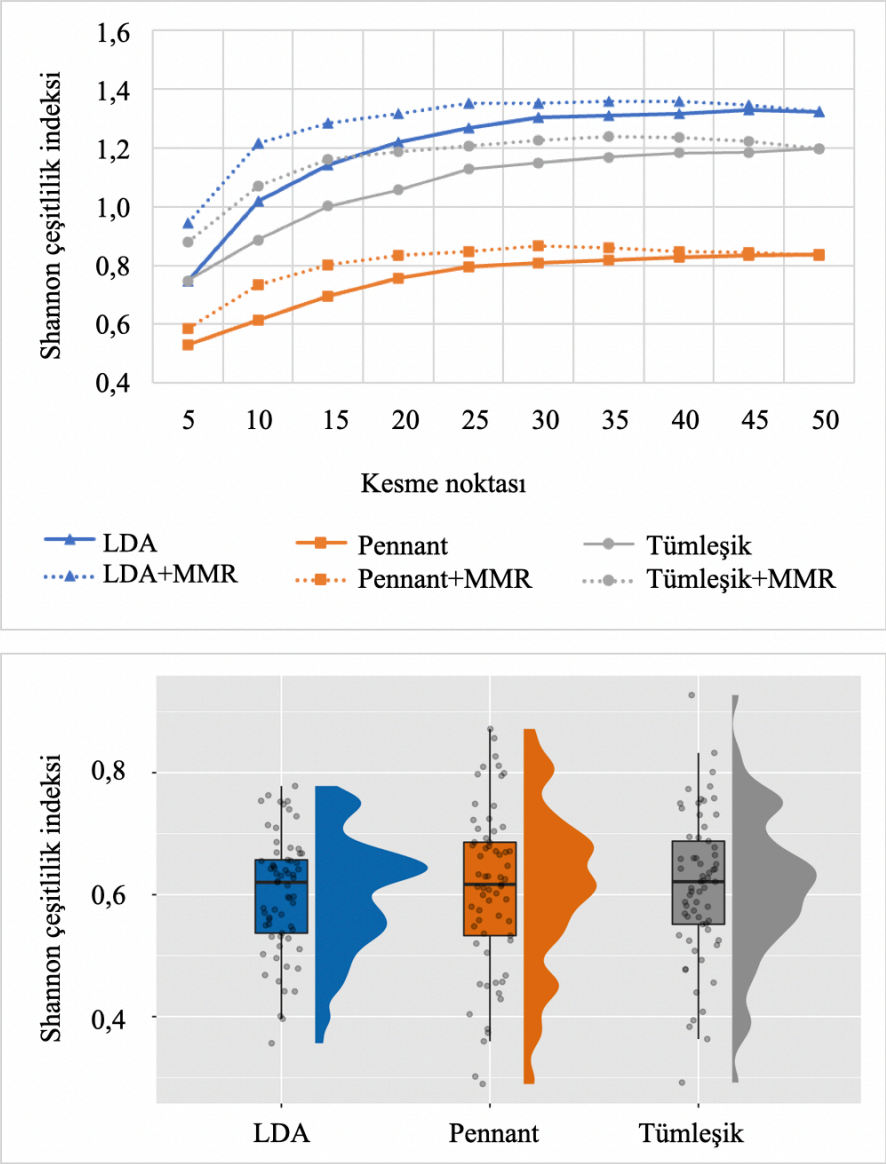
---

[39] Bu araştırmanın ilk bulguları için bkz.: Akbulut ve Tonta (basım aşamasında).



Şekil 13. Algoritmaların ilgi değerlerine göre karşılaştırılması

Çalışma kapsamında önerilen tümleşik sıralama 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 14. Algoritmaların Shannon çeşitlilik indeksine göre karşılaştırılması

Şekil 13 ve 14'teki 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 13). MMR uygulandığında kesme noktası 15 hariç tüm kesme noktalarında daha ilgili kaynaklara üst sıralarda erişilmiştir. Çünkü pennant erişim algoritması benzer konudaki makaleleri üst sıralara yerleştirirken, MMR algoritması ise benzer makaleleri, ilgi sıralamasının sonuna doğru itip farklı konuda ama gene de sorguyla ilgili olan makaleleri üst sıraya yerleştirmektedir. Pennant sıralamasındaki 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.

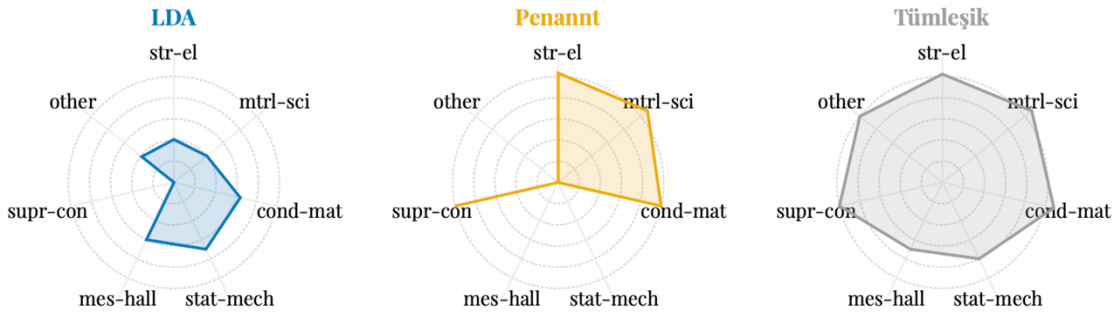
Tümleşik sıralamada ise hem kelime sıklıkları (LDA) hem ortak atıf bağlantıları (pennant erişim) kullanılarak sorguyla doğrudan ilgili makaleler -benzer olanları alt sıralara itme gibi bir endişe olmadan- üst sıralara yerleştirilmektedir. Bu sıralamaya MMR uygulandığında da üst sıralarda daha ilgili kaynaklara erişilmektedir. Fakat kesme noktası 25'ten sonra ilgi değerleri tümleşik algoritmanın eriştiği makalelerin ilgi değerleri ile neredeyse aynıdır (diğer bir deyişle MMR etkisi kaybolmaktadır). Bu durum pennant algoritmasının tümleşik sıralamaya yansıması olarak değerlendirilebilir. Pennant algoritması kesme noktası 25'e kadar genellikle benzer konudaki çalışmaları ilk sıralara yerleştirmiş, 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 sıralamaya eklendiği kesme noktası muhtemelen 25'ten daha büyük olacaktı.

Farklı sıralama algoritmalarının erişilen belgelerin hangi özelliklerini öne çıkardıklarının anlaşılabilmesi için sıralamaların bir de çeşitlilik açısından incelenmesi gerekir. LDA algoritması sorgularla ilgili çeşitli konulardan belgelere tümleşik ve pennant algoritmalarına kıyasla daha sık erişmektedir (Şekil 14). 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 algoritması çalıştırılarak erişilen ilgili belgelerin 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 (Waltman ve Van Eck, 2012).

LDA ve pennant erişim sıralamaları ile tümleşik sıralamalara MMR algoritması uygulandığında tüm kesme noktalarında daha çeşitli konulardaki kaynaklara en üst sıralarda erişildiği gözlenmektedir (Şekil 14). Fakat pennant sıralamasına 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 13'teki 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 sıralamaya üst sıralardan girebilmektedir. Bu bulgular "LDA ve pennant erişim sıralamaları ile tümleşik sıralamalarda hangi özellikler öne çıkmaktadır?" sorusunun cevabıdır (birinci araştırma sorusu). LDA algoritması çeşitli konulardaki makalelere erişim sağlarken pennant erişim algoritması ise farklı literatürlerle bağlantıları yakalayarak marjinal ama sorguyla hâlâ ilgili makalelerin tümleşik sıralamaya eklenmesini sağlamaktadır.

Önerilen sıralamanın ilgi ve çeşitlilik açısından bilgi erişim performansına katkısı 60. sorgu üzerinden gösterilmiştir (Şekil 15). Sorgu 60 için algoritmaların eriştikleri ilk 25'er makalenin konulara göre ilgi değerleri geliştirdiğimiz yöntemin sıralamayı ne kadar zenginleştirdiğine bir örnektir.

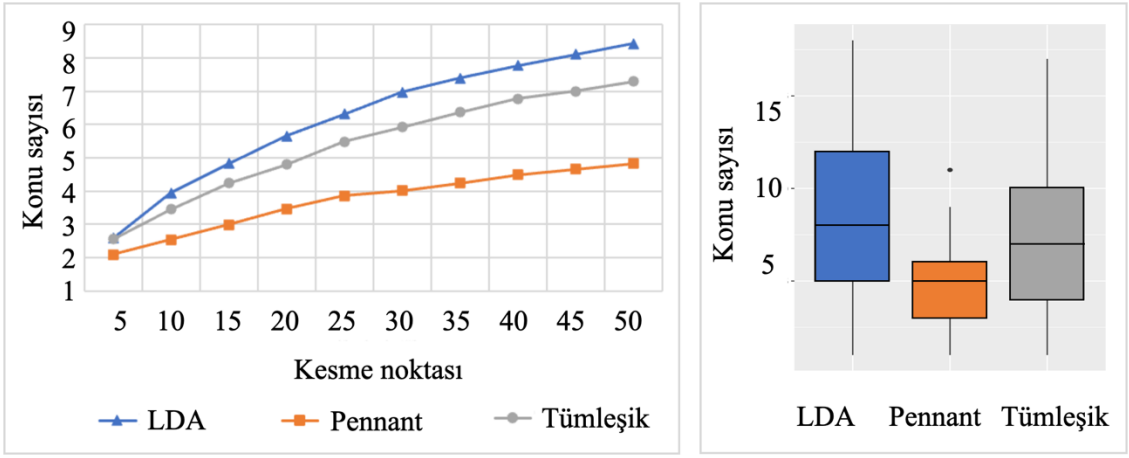


Şekil 15. 60. sorgu için algoritmaların konu çeşitliliği ve ilgi karşılaştırması

Not. "str-el": Strongly Correlated Electrons, "mtrl-sci": Materials Science, "cond-mat": Condensed Matter, "stat-mech": Statistical Mechanics, "mes-hall": Mesoscale and Nanoscale Physics, "supr-con": Superconductivity, "other": Other Condensed Matter.

Shannon çeşitlilik indeksinin yanı sıra konu sayısı da çeşitlilik ile ilgili bir fikir vermektedir.[40] Konu sayısı değerlerinin yansıtıldığı Şekil 16 ile Shannon çeşitlilik indeksi değerlerinin yer aldığı Şekil 14 örüntü açısından birbirine çok benzemektedir. Bu açıdan her iki değer (çeşitlilik indeksi ve konu sayısı) de çeşitlilik olarak değerlendirilebilir.

[40] Konu sayısı ve Shannon çeşitlilik indeksi ile ilgili detaylı grafikler için bkz. http://mugeakbulut.com/phd/gorsellestirme/SW_vs_konu_sayisi.png



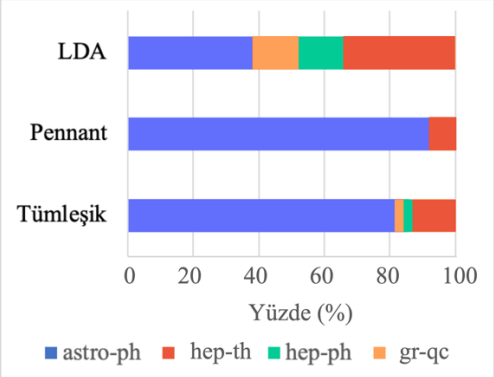
Şekil 16. Konu sayısı değerleri

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 yüksek olması, konuların ayrıntı düzeyini de artırdığı için tutarsızlık sorunlarının ortaya çıkması normaldir. Bu çalışma kapsamında kullanılan bibliyometrik veriler de büyük bir derlemden (atıf dizinleri) alınmıştır. 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).

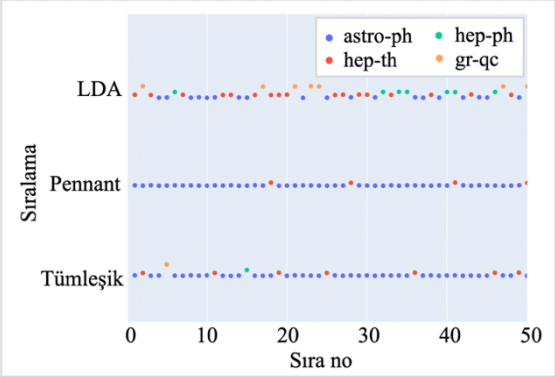
4.1.2. Algoritmaların Sıralamalara Katkılarının Konulara Göre Analizi

Hangi algoritmanın hangi konu bağlantılarını tümleşik sıralamaya eklediğini izleyebilmek için 42. sorgunun konusal açıdan detaylı bir analizi yapılmıştır. Sorgu 42 için tümleşik sıralamanın ilk 50 sırasındaki 36 makale pennant erişim algoritmasından, 14 makale ise LDA algoritmasından gelmektedir. Şekil 17'de LDA ve pennant erişim sıralamaları ile tümleşik sıralamadaki ilk 50'şer

makalenin konu dağılımı verilmektedir. Şekil 18'de ise algoritmaların ilgili konulardaki makalelere kaçıncı sırada eriştikleri bilgisi yer almaktadır.



Şekil 17. Algoritmalara göre konuların dağılımı (sorgu 42)



Şekil 18. Sıralamalara göre konuların dağılımı (sorgu 42)

Not. "astro-ph": Astrophysics, "gr-qc": General Relativity and Quantum Cosmology, "hep-ph": High Energy Physics – Phenomenology, "hep-th": High Energy Physics – Theory.

LDA sıralamasında yedi tane *High Energy Physics - Phenomenology* konulu makale olduğu görülmektedir (sıra numaraları: 6, 32, 34, 35, 40, 41 ve 46). Bunlardan altıncı sıradaki makale, tümleşik sıralamaya 15. sırada yansımıştır. LDA sıralamasında daha üst sıralarda olan makalenin tümleşik sıralamaya daha sonraki sıralarda dâhil olmasının nedeni pennant skorlarının LDA skorlarından daha yüksek olmasıdır.[41]

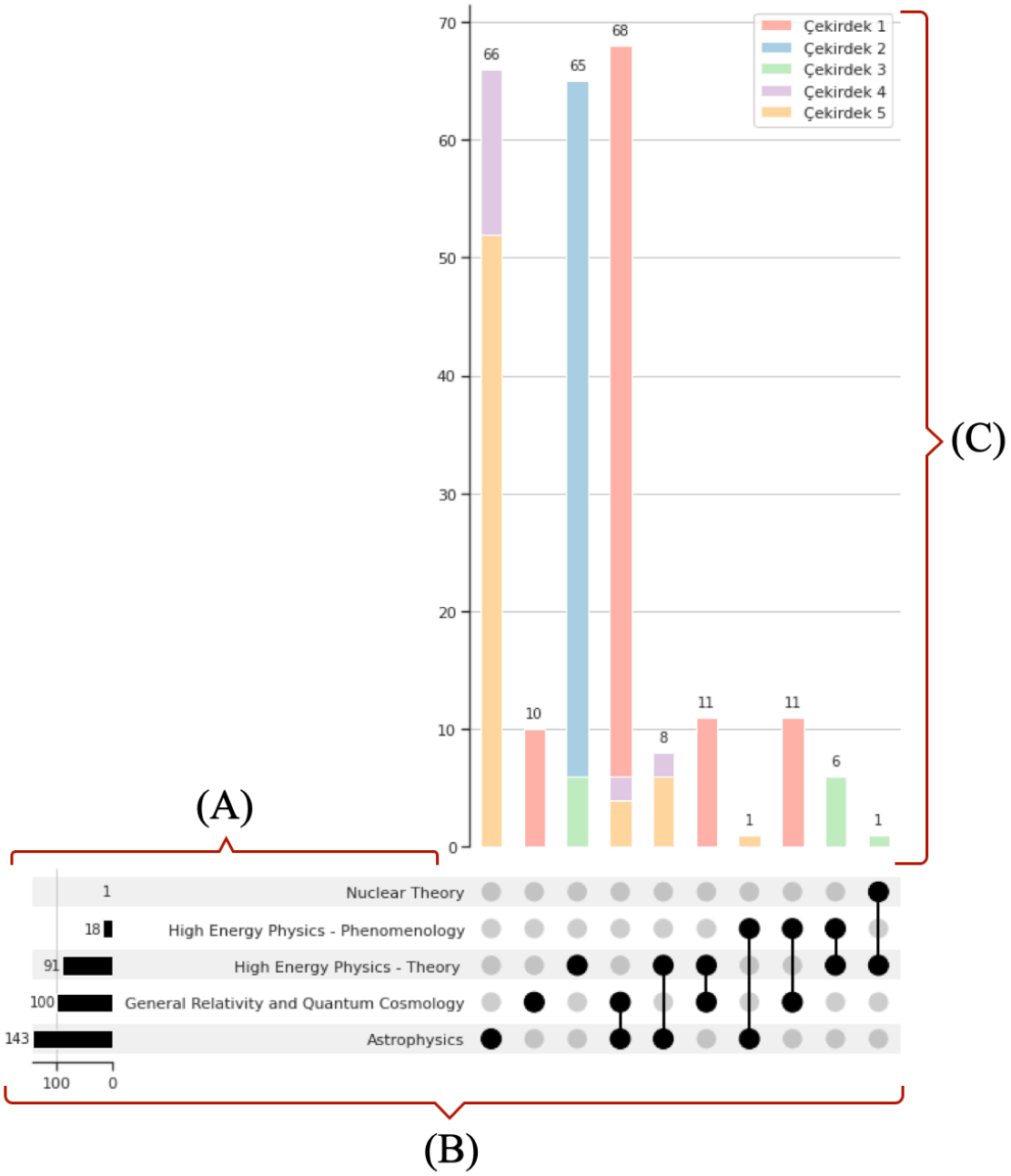
LDA sıralamasında 42. sorgu için listelenen ilk beş makale üç farklı konudadır (Tablo 3). Makalelerin tek bir konuya indirgenmesi o makalelerin içeriği hakkında sınırlı bilgi vermektedir. Öte yandan makalenin kaynakçasındaki makalelerin konuları o çalışmanın kapsamı hakkında daha çok bilgi vermektedir. Bu yüzden algoritmaların sıralamalara katkısının daha iyi gözlenebilmesi için LDA, pennant ve tümleşik ilgi sıralamalarında ilk 50'şer makale kaynakçalarıyla birlikte değerlendirilerek devrik grafikler (UpSet plots) ile detaylı olarak incelenmiştir. Devrik grafikler kaynakça düzeyinde örtüşen konuları görselleştirmek için etkili bir yaklaşım sunmaktadır (Bougioukas ve diğerleri, 2021).

[41] Bkz. http://mugeakbulut.com/phd/gorsellestirme/kesme_noktalari.pdf (sorgu 42, 2. sütun).

Tablo 3. Çekirdek makalelerin konuları

Çekirdek no (LDA sıra no)	Konu
1	High Energy Physics - Theory (hep-th)
2	General Relativity and Quantum Cosmology (gr-qc)
3	High Energy Physics - Theory (hep-th)
4	Astrophysics (astro-ph)
5	Astrophysics (astro-ph)

Sorgu 42 için Tablo 3'te konuları verilen beş çekirdek makalenin kaynakçalarında toplam 353 makale bulunmaktadır. Şekil 19'da ise çekirdek makalelerin kaynakçaları için konuların birlikte geçiş sıklıkları yer almaktadır. Daha detaylı olarak bu şekil katkıların hangi çekirdek makaleden geldiği bilgisini de içermektedir (sütun grafik bölümündeki renklendirme). Çekirdek makalelerin ve kaynakçalarının konuları genel olarak örtüşmektedir. Kaynakçaları dâhil edince *Nuclear Theory* ve *High Energy Physics- Phenomenology* konuları da eklenmiştir. Bu grafikteki konular kabaca çekirdek makalede yararlanılan makalelerin konuları olarak değerlendirilebilir.

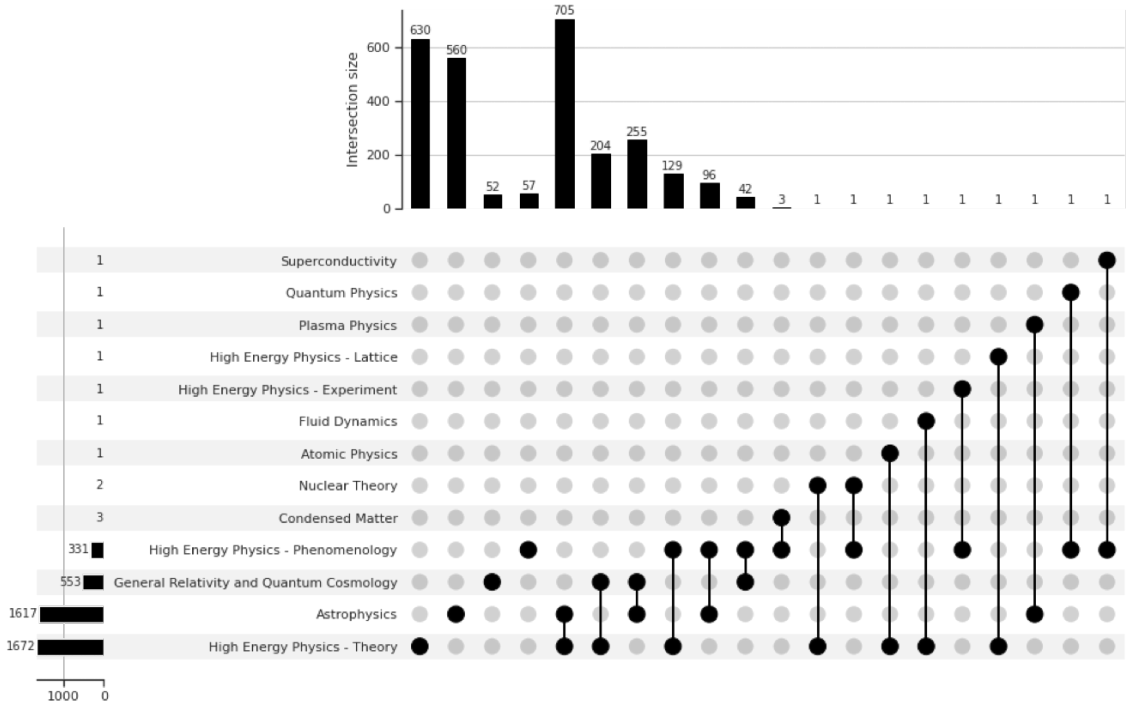


Şekil 19. Çekirdek makalelerin kaynakçaları için devrik grafik

Not. Devrik grafiklerde kabaca üç bölüm bulunmaktadır: **(A)** Yatay sütun grafiklerinin yer aldığı ve ilgili konu sayısını veren kısım, **(B)** konuların birlikte geçiş sıklıklarının görselleştirildiği çakışma bölümü ve **(C)** dikey sütunların olduğu ve çakışan konuların sıklıklarının yer aldığı bölüm. Örneğin sorgu 42 için çekirdek makalelerin kaynakçalarında yer alan 353 makaleden (143+100+91+18+1) 143'ünün konusu *Astrophysics*'tir. Bu makalelerden sekizi *High Energy Physics-Theory* konusuyla birlikte geçmiştir. Söz konusu sekiz bağlantının (*High Energy Physics-Theory* ve *Astrophysics* bağlantısı) ikisi *Çekirdek 4*, altısı ise *Çekirdek 5* aracılığı ile kurulmuştur.

LDA algoritması çalıştırıldığında 42. sorgu için erişilen 50 makale ve bu makalelerin kaynakçaları incelenerek LDA'nın hangi konularda katkı sağladığı Şekil 20'de gözlenebilir. Sıralamadaki ilk

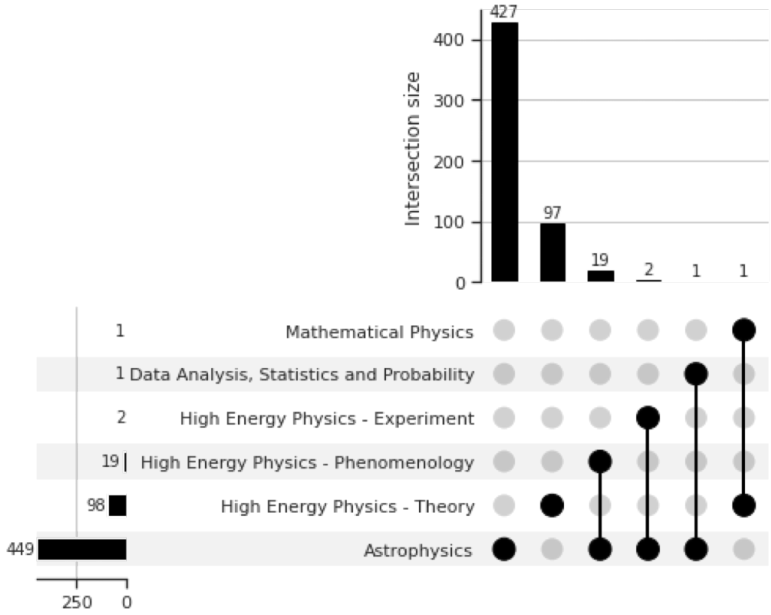
50 makale dört farklı konudan oluşmaktadır (bkz. Şekil 17 ve 18). Fakat kaynakçalar da devrik grafiğe dâhil edildiğinde sıralamaya dokuz yeni konunun daha eklendiği görülmektedir (Şekil 20). LDA algoritması ile ilk 50'de erişilen makalelerin kaynakçalarında *High Energy Physics – Theory* konusu 1672, *Astrophysics* konusu ise 1617 kez geçmiştir. Söz konusu iki konunun birlikte geçiş sıklıkları ise 705'tir. Beş çekirdek makalenin dördünün bu iki konuda olmasından dolayı çakışmaların yüksek çıkması normaldir. Benzer bir şekilde bir diğer çekirdek makale konusu olan *General Relativity and Quantum Cosmology* ise tek başına 553 kez, *High Energy Physics – Theory* konusu ile 204 kez ve *Astrophysics* ile de 255 kez birlikte geçmiştir. Grafikte çekirdek makalelerin konularının baskın olması normaldir, asıl çeşitlilik etkisi (ya da marjinal makale katkısı) diğer konular üzerinden gözlenebilir. Örneğin *Superconductivity*, *Quantum Pyhsics*, *Plasma Physics* gibi konular sorguyla marjinal ilgilidir. LDA sıralamasında yer almasa da, bu konular sıralamadaki makalelerin kaynakçalarında yer aldıkları için dolaylı olarak o konularla ilgili oldukları varsayılmaktadır. Sorguyla konuların bağlantısı LDA algoritması ile sağlanmıştır.



Şekil 20. LDA sıralaması için devrik grafik

Ortak atıf ortalaması dokuz olan 42. sorguda LDA sıralamasındaki ilk beş makaleden yola çıkarak oluşturulan pennant sıralamasında ise ilk 50'de sadece *Astrophysics* ve *High Energy Physics - Theory* konularındaki makalelere erişilmiştir (Şekil 21). Kaynakçalar da eklenerek pennant erişim algoritması için devrik grafik oluşturulduğunda LDA'ya oranla daha tutarlı bir konu dağılımı gözlenmektedir (Şekil 21). Çekirdek makalelerde üç temel konu ön plandadır (*Astrophysics*, *High*

Energy Physics – Theory ve *High Energy Physics – Phenomenology*). Bu konuların dışında sadece üç yeni konu ile bağlantı kurulmuştur (*High Energy Physics – Experiment; Data Analysis*, *Statistics and Probability* ve *Mathematical Physics*).



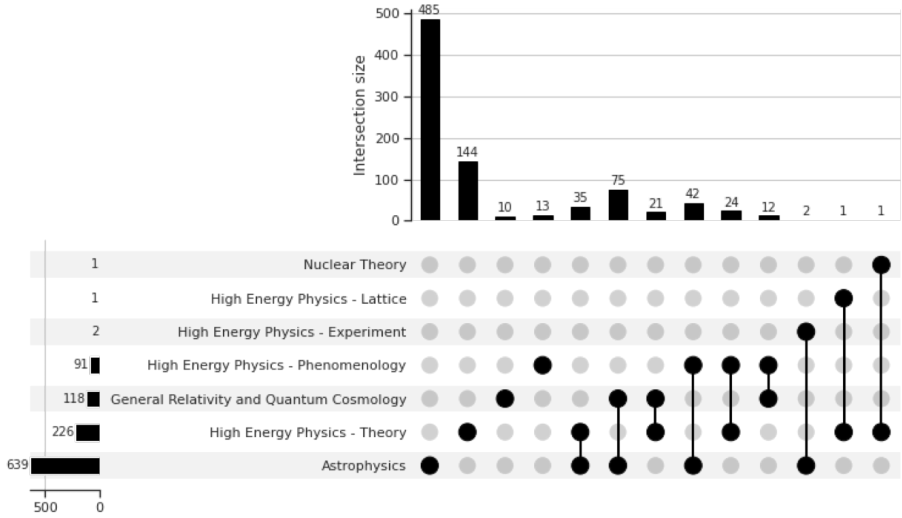
Şekil 21. Pennant sıralaması için devrik grafik

Devrik grafiklerin yapısı gereği herhangi bir konudaki bir çekirdek makalenin kaynakçasında çok sayıda makale varsa o konunun grafikteki geçiş sıklığı da yüksek olmaktadır. Fakat daha az sayıda birlikte geçiş sıklığı gözlenen konular algoritmanın eriştiği yeni bağlamları yakalamak açısından önemlidir. Bu örnekte de çekirdek makalelerin kaynakçalarından gelen üç yeni konu sorguyla marjinal olarak ilgilidir. Pennant erişim algoritmasının eriştiği makaleler ve kaynakçaları konu açısından tutarlıdır.

Tümleşik ve LDA sıralamalarındaki ilk 50'şer makalenin konuları aynı[42] olmasına karşın makalelerin kaynakçalarında referans verilen kaynakların konularıyla birlikte değerlendirildiğinde LDA sıralamasının konu çeşitliliği açısından daha zengin olduğu gözlenmektedir (Şekil 22). Öte yandan, *High Energy Physics – Phenomenology* konusuyla bağlantı da tümleşik sıralamaya pennant algoritması vasıtasıyla eklenmiştir. Grafikte yer alan *Nuclear Theory* ve *High Energy Physics – Lattice* konularının kaynakçalardan geldiği

[42] Tümleşik ve LDA sıralamalarında yer alan ilk 50'şer makalenin konuları Astrophysics, General Relativity and Quantum Cosmology, High Energy Physics – Phenomenology ve High Energy Physics – Theory'dir. Bkz. Şekil 17.

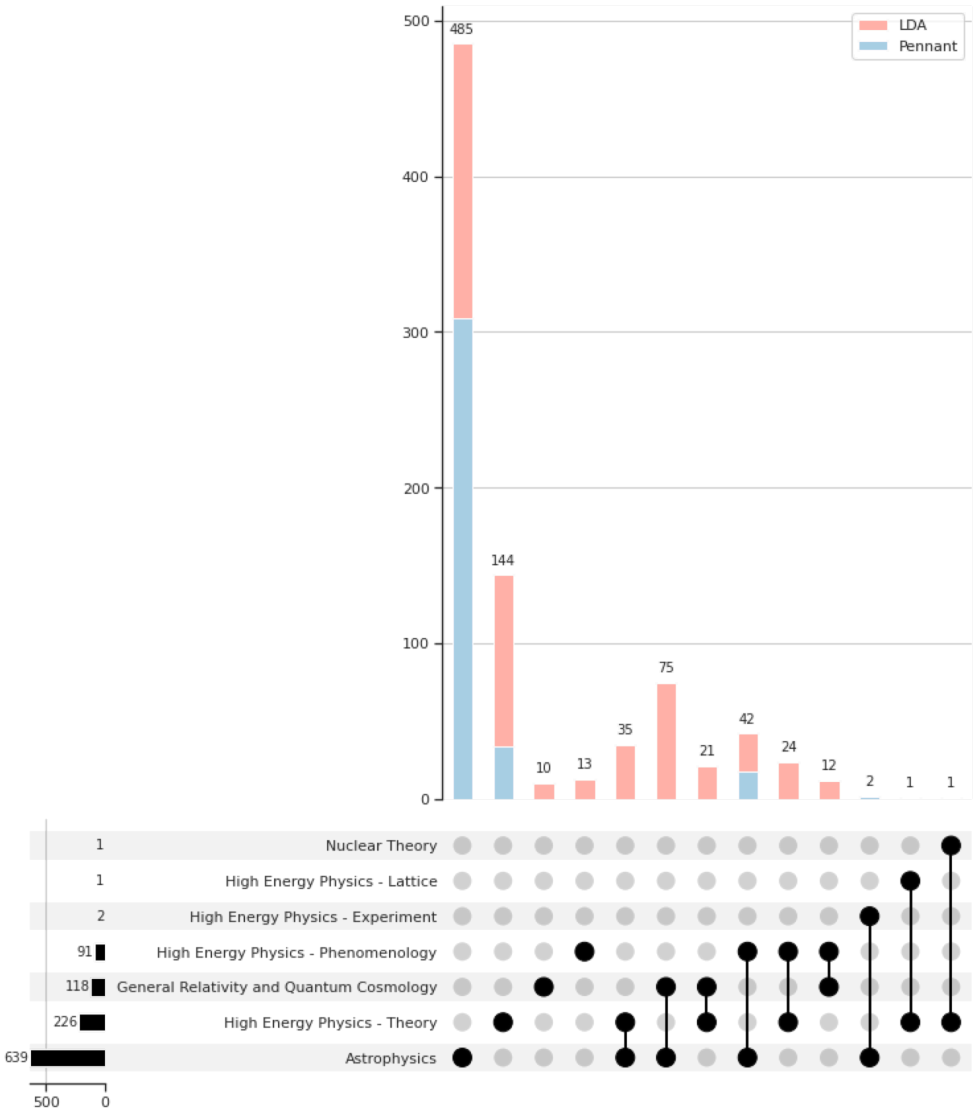
anlaşılmaktadır. Sayıca az olan bu konu grubu *High Energy Physics – Theory* konusu ile bağlantılı olarak sıralamaya girmiştir.



Şekil 22. Tümleşik sıralama için devrik grafik

Tümleşik sıralamaya daha detaylı olarak hangi çekirdek makalelerden katkı sağlandığı ayrıca incelenmiştir. Tümleşik sıralamadaki ilk 50 kaynak için pennant erişimden gelen 36 kaynağın referans listesinde toplam 363 makale bulunmaktadır. LDA algoritmasından gelen 14 makalenin kaynakçasında ise 502 makale yer almaktadır. Devrik grafikte LDA baskın bir görüntü oluşmasının nedeni kısmen bununla ilgilidir. Fakat daha önce de belirtildiği gibi 42. sorgu aslında pennant baskın bir sıralamadır. Diğer sorgularda da olduğu gibi bu sorgu için de pennant erişim algoritması tutarlı konulara erişirken, LDA algoritması marjinal konulara erişim sağlamış ve sıralamaya yeni konu bağlantıları eklemiştir.

Tümleşik sıralamaya hangi algoritmanın hangi konu bağlantılarını eklediği bilgisi Şekil 23 aracılığıyla incelenebilir. Örneğin, *Astrophysics* ile *High Energy Physics – Experiment* bağlantısı pennant erişim algoritması aracılığıyla, *Astrophysics* ile *High Energy Physics – Theory* bağlantısı ise LDA algoritması aracılığıyla sağlanmıştır.



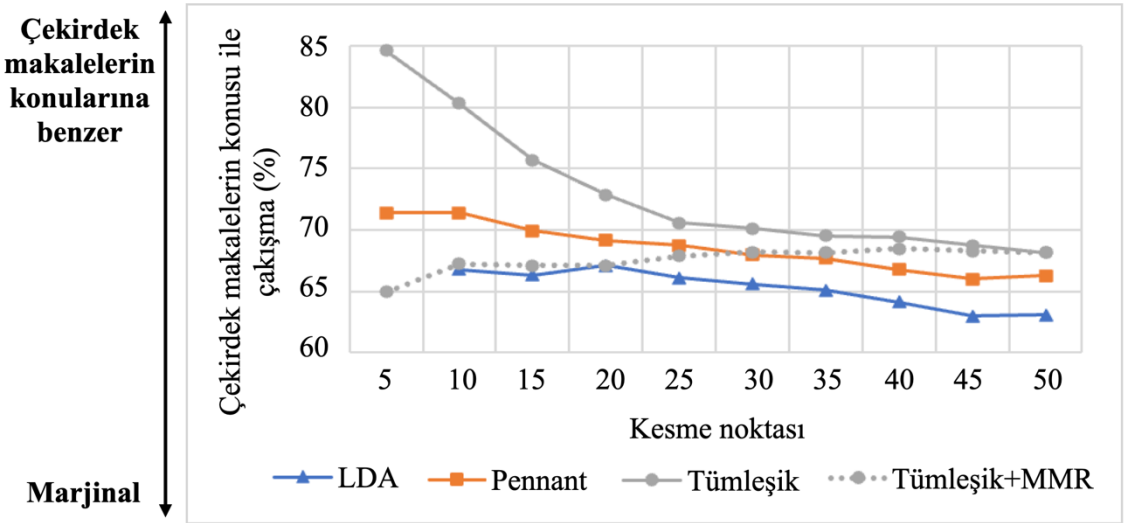
Şekil 23. Pennant ve LDA algoritmalarının tümleşik sıralamaya katkısı

Genel bir yorum yapmak gerekirse, tümleşik sıralama için oluşturulan devrik grafik (Şekil 22), çekirdek makaleler için oluşturulan devrik grafikteki (Şekil 19) konuların tümünü kapsamaktadır. Diğer bir deyişle çekirdek makalelerde yararlanılan konu bağlantıları tümleşik sıralamada da yer almaktadır. İki algoritmanın birleştirilmesiyle konunun farklı bağlamlarını yakalayan kapsamlı bir sıralama oluşturulmuştur.

4.1.2. Çekirdek Makalelerin Konuları ile Çakışma

Çekirdek makalelerin konuları ile çeşitli algoritmalar tarafından erişilen makalelerin çakışma oranları incelenmiştir. Tüm sorgular için ilgili kesme noktalarında ortalama çakışma oranları izlenerek algoritmaların hangi kesme noktasında nasıl davrandıkları kestirilebilir (bkz. Şekil 24).

Çekirdek makaleler LDA algoritması ile erişilen ilk beş makale olduğu için LDA algoritmasında kesme noktası beş için herhangi bir değer yansıtılmamıştır.



Şekil 24. Çekirdek makalelerin konuları ile çakışma oranları

Konu çeşitliliğinde sorgunun interdisiplinerlik derecesi de önemli bir etkendir. Örneğin, spesifik bir sorgu ile LDA'nın eriştiği ilk beş makalenin konuları aynı olursa o zaman çakışma olasılığı da artmaktadır.

Toplam 65 sorgu için çekirdek makalelerin tekil konu sayısı ortalaması ve ortancası üçtür. Bu sorgulardan 15'i için çekirdek makalenin hepsi de aynı konudadır.[43] Dolayısıyla bu sorgular için çakışma olasılığı diğer 50 sorguya göre daha fazladır. Çekirdek makaleler için tekil konu sayıları ile ortak atıf sayısı ortalamaları arasındaki korelasyon çok düşük ve negatiftir (r = -0,08). Yani ortak atıf ortalaması arttıkça pennant erişim algoritmasının daha tutarlı ve az sayıda konuya erişeceği söylenebilir. Ortak atıf sayısı fazla olan sorgular için pennant skorları da yüksek olacağından pennant baskın sıralama oluşturulması da daha olasıdır.

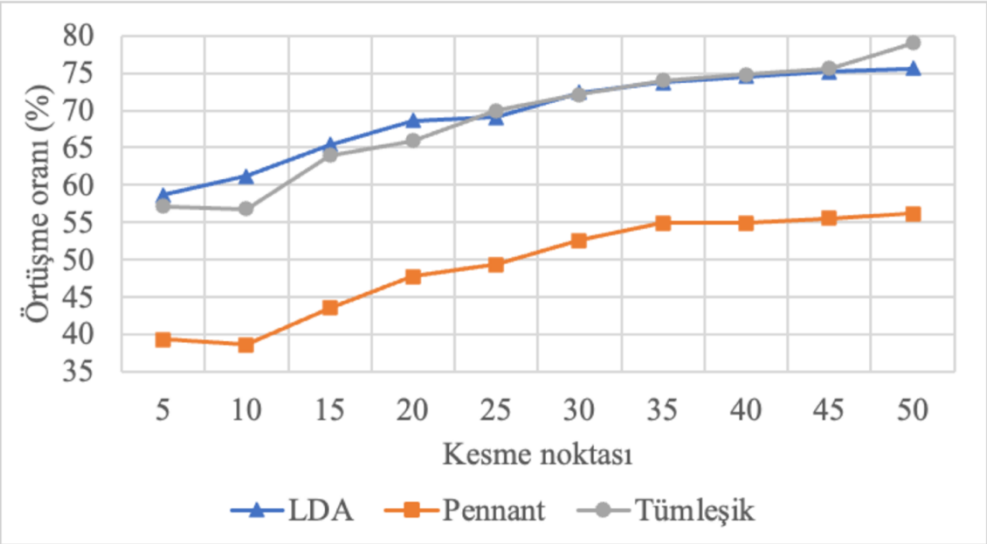
4.1.3. Çekirdek Makalelerin Kaynakçalarındaki Makalelerin Konuları ile Erişilen Makalelerin Konularının Çakışması

Çekirdek makalelerin kaynakçalarındaki konu dağılımı kabaca yazarların o araştırmaları yaparken yararlandıkları kaynakların konularını yansıtmaktadır. Pennant erişim algoritmasının

[43] Geri kalan çekirdek makalelerdeki tekil konuların dağılımı 17 sorgu için iki; 19 sorgu için üç, sekiz sorgu için dört ve altı sorgu için beştir.

atıflara ve ortak atıflara dayanarak eriştiği çalışmaların konu başlıkları ise çekirdek makalelerin literatüre etkisini yansıtmaktadır.

LDA, pennant erişim ve tümleşik ilgi sıralamalarındaki ilk 50'şer makalenin konuları ile LDA'nın eriştiği ve pennant erişimde çekirdek olarak kullanılan ilk beş makalenin kaynakçalarındaki konuların örtüşme oranı Şekil 25'te izlenebilir.[44] Pennant daha ilgili ve tutarlı bir şekilde benzer konudaki makaleleri sıraladığı için çekirdek makalelerin kaynakçaları ile örtüşme oranı daha düşüktür. LDA sıralaması ve tümleşik sıralama ise birbirine çok yakın değerlere sahiptir.



Şekil 25. Çekirdek makalelerin kaynakçaları ve farklı algoritmalar tarafından erişilen makalelerin konularının örtüşme oranları

Tüm algoritmalar için kesme noktası arttıkça konu çeşitliliğinin ve dolayısıyla çakışmanın artması beklenen bir durumdur. LDA için kesme noktası 5'te çakışma oranı %58'dir. Diğer bir deyişle 65 sorgu için LDA sıralamasındaki ilk beş kaynağın üçü çekirdek makalenin kaynakçasında yer alan konularla örtüşmektedir. Çekirdek makalelerin farklı konularda olma olasılığı da düşünüldüğünde %58 çakışma oranına bakılarak algoritmanın tutarlı bir yapıda olduğu söylenebilir.

4.2. İLGİ SIRALAMALARININ GENEL DEĞERLENDİRMESİ

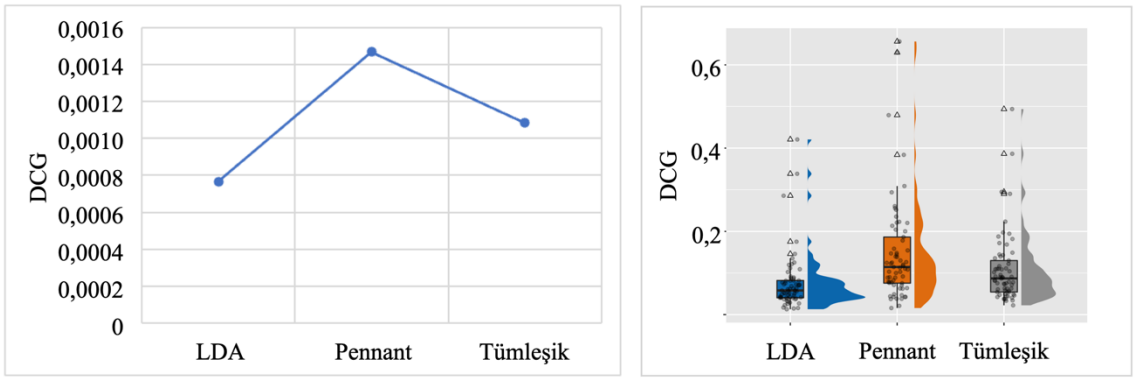
Sıralamaların hangi özellikleri ön plana çıkardıkları belirlendikten sonra bu alt bölümde sıralamalara dair genel değerlendirmelere yer verilmiştir.

[44] iSearch derleminde makalelerin kaynakçalarındaki tüm kaynaklar yer almamaktadır. Kaynakçadaki makalelerin sadece arXiv'de yer alanları listelenmektedir.

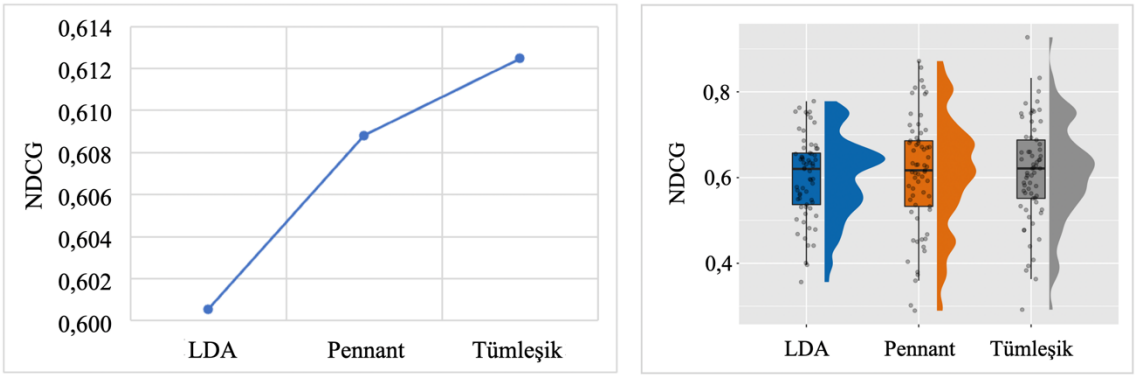
4.2.1. DCG ve NDCG Değerleri

Sıralamaların kalitesini karşılaştırmak için iki ölçevden faydalanılmıştır. Daha önce de belirtildiği gibi DCG ölçevi sıralamanın en üstüne yakın ögelerin kalitesini ölçmektedir. NDCG ise bu ölçevin ideal sıralama da hesaba katılarak DCG ölçevinin normalleştirilmiş halidir.[45]

Şekil 26'da iSearch'teki tüm sorgular için ortalama DCG değerleri gösterilmektedir. Herhangi bir algoritma ile daha ilgili belgelere (ağ değeri daha yüksek makaleler) üst sıralarda erişildiğinde DCG skorları da yükselmektedir. Pennant erişim algoritması en ilgili makalelere üst sıralarda eriştiği için ortalama DCG değeri diğer algoritmalara göre daha yüksektir. İdeal sıralama dikkate alınarak yapılan NDCG değerlerine göre ise en başarılı sıralama tümleşik sıralamadır (Şekil 27). NDCG değerlerinin DCG değerlerinden yüksek olması önerilen algoritmanın ilgi açısından ideal sıralamaya daha yakın olduğunu göstermektedir.



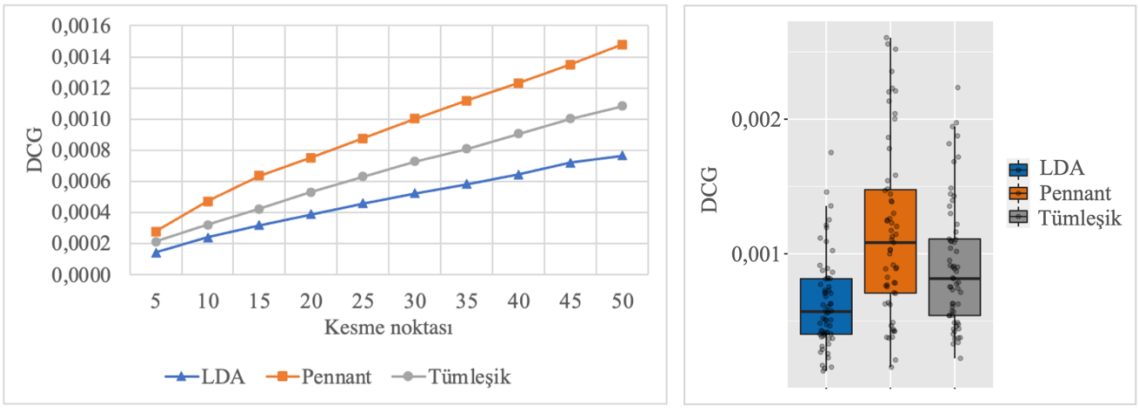
Şekil 26. Algoritmaların ortalama DCG değerleri



Şekil 27. Algoritmaların ortalama NDCG değerleri

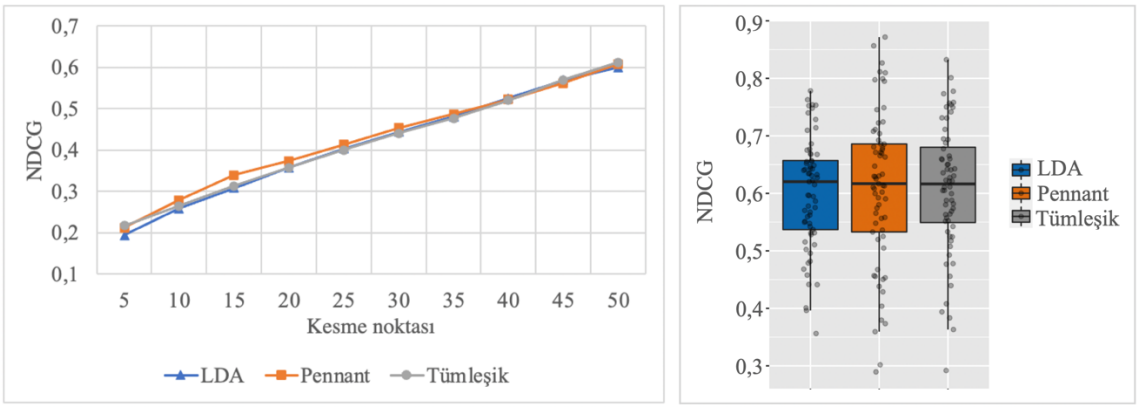
[45] DCG değerleri ideal ilgi sıralamasına bölünerek 0-1 aralığında normalize edilmiş değerler elde edilir. Her makale için ilgi değeri (derece merkeziliği) bilindiği için erişilen makaleler en ilgiliden en ilgisize doğru sıralanarak ideal ilgi sıralaması elde edilmektedir.

DCG ve NDCG değerleri detaylı olarak çeşitli kesme noktalarında da incelenmiştir.[46] DCG değerleri tüm algoritmalar için kesme noktası arttıkça yükselmektedir (bkz. Şekil 28).



Şekil 28. Çeşitli kesme noktalarında algoritmaların ortalama DCG değerleri

NDCG değerlerinde ise kesme noktası 15 ve 20'de pennant algoritmasının NDCG skorunun nispeten daha yüksek olduğu göze çarpmaktadır (bkz. Şekil 29). Söz konusu kesme noktalarında pennant erişim algoritması marjinal kaynakları sıralamaya eklemeye başladığı için ilgi değerleri düşmüştür. Marjinal makaleler özellikle kesme noktası 45'te gözlenebilmektedir.



Şekil 29. Çeşitli kesme noktalarında algoritmaların ortalama NDCG değerleri

4.2.2. Kapsama ve Yenilik Oranları

Araştırma kapsamında konu düzeyinde kapsama ve yenilik oranları hesaplanmıştır. Kapsama oranı algoritmanın ilgili olduğu bilinen makalelere, yenilik oranı ise ilgili olduğu bilinmeyen makalelere erişim açısından ne kadar başarılı olduğunu belirtmektedir (Kaminskas ve Bridge, 2016) Bu bağlamda çekirdek makalelerin kaynakçalarındaki makalelerin konuları *ilgili konu*

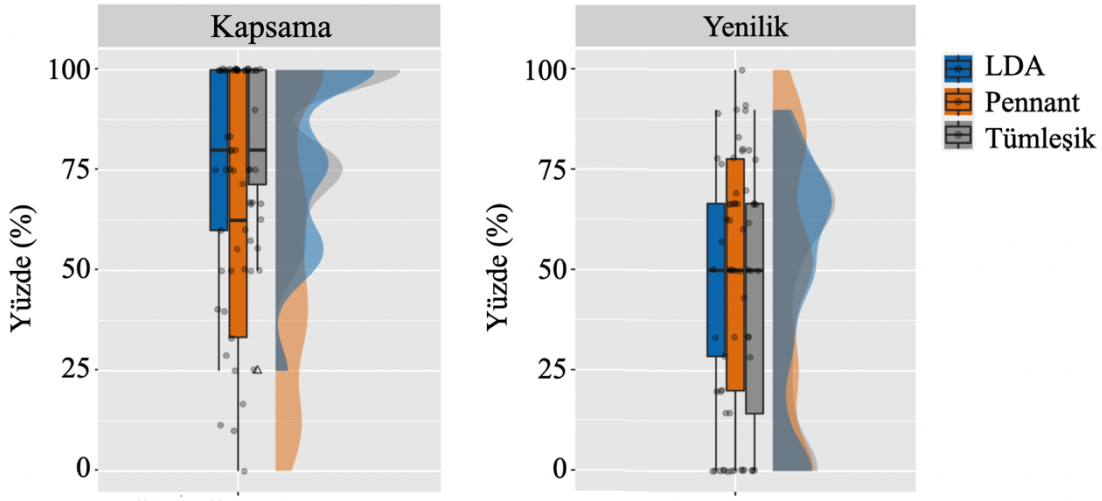
[46] Dağılımın daha iyi gözlenebilmesi için Şekil 28 ve Şekil 29'da uç değerler (outliers) kutu grafiklerden çıkarılmıştır. Uç değerlerin dâhil olduğu grafikler için bkz. Şekil 26 ve Şekil 27.

olarak kabul edilmiştir. Buradan hareketle tüm algoritmalar için erişilen ilk 50'şer makalenin konuları ile çekirdek makalenin kaynakçalarındaki makalelerin konuları karşılaştırılmıştır.

Örneğin sorgu 15'in çekirdek makalelerinin kaynakçalarında altı tekil konu bulunmaktadır. Algoritmalar tarafından erişilen ve konusu altı makalenin konularıyla aynı olan makaleler ilgili olduğu bilinen makaleler olarak işaretlenmektedir. LDA sıralaması ve tümleşik sıralamada ilk 50'de yer alan makalelerin konuları çekirdek makalelerin kaynakçalarındaki altı konunun tümünü, pennant algoritması ile erişilen ilk 50 makalede ise altı konunun beşini içermektedir. Dolayısıyla sorgu 15 için kapsama oranları LDA=%100 $((6/6) * 100)$, pennant=%83 $((5/6) * 100)$ ve tümleşik=%100 $((6/6) * 100)$ olarak hesaplanmıştır.

Yenilik oranı ise çekirdek makalelerin kaynakçalarındaki konulardan farklı tekil konu sayısının, erişilen makalelerdeki tekil konu sayısına oranıdır. Diğer bir deyişle, yenilik oranı algoritmanın marjinal makalelere erişim oranı hakkında bilgi vermektedir. LDA ve pennant erişim sıralamaları ile tümleşik sıralamada erişilen ilk 50 makalenin tekil konu sayıları sırasıyla 7, 5 ve 6'dır. Üç sıralama için de ilk 50'de çekirdek makalelerin kaynakçalarındaki altı tekil konunun tümü yer almaktadır. LDA sıralaması ise bu altı konuya ek olarak yeni bir konu daha içermektedir. Dolayısıyla sorgu 15 için yenilik değerleri LDA=0,14 (1/7), pennant=0 (0/5) ve tümleşik=0 (0/6) olarak hesaplanmıştır.

Şekil 30 iSearch'teki tüm sorgular için hesaplanan yenilik ve kapsama oranlarını göstermektedir. Kapsama açısından LDA sıralamalarının ve tümleşik sıralamaların ortancaları aynıdır. Pennant sıralamasının kapsama oranı ortanca değeri ise daha düşüktür. Öte yandan pennant erişim ve LDA sıralamaları bütünleştirildiğinde kapsama oranı tam olan sorguların sayıca fazla olduğu gözlenmektedir. Tümleşik sıralama için %75 ve tam kapsama değerlerinde yoğunlaşma vardır (çift modlu dağılım). LDA ve pennant algoritmalarının birleşiminden oluşan sıralamanın kapsama oranı daha yüksek sıralamalar oluşturduğu söylenebilir. Bu bulgu "Çalışmaların özetlerine uygulanan LDA konu modelleme algoritmasıyla oluşturulan sıralama ile, toplam atıf ve ortak atıf verileri de dâhil edilerek hesaplanan pennant erişim çıktıları artırımlı olarak tümleştirilerek (fusion) arama yapılan konuyu daha geniş kapsamda içeren ilgi sıralaması elde edilebilir mi?" (ikinci araştırma sorusu) sorusunun cevabı niteliğindedir. Önerilen algoritma ile kapsama oranı yüksek olan sıralamalar elde edilmiştir.

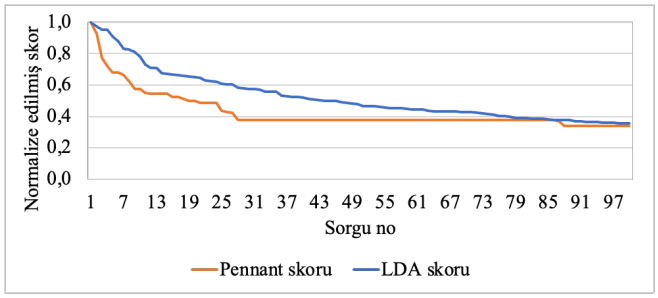


Şekil 30. Kapsama ve yenilik oranları

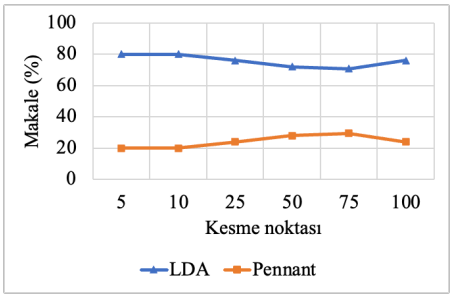
Yenilik oranları için ortanca değerleri tüm algoritmalar için aynıdır (%50). Pennant algoritması için yenilik oranı %100 olan sorgular da gözlenmektedir. Bunun nedeni seyrek ortak atıf ilişkileri ile erişilen marjinal makalelerdir. LDA sıralamaları ve tümleşik sıralamalar için en yüksek yenilik oranları %90'dır.

4.3. ÖNERİLEN ALGORİTMANIN İŞLEYİŞİ

İlgi sıralamalarının artırımlı olarak geliştirilmesinde pennant erişimin katkısı her sorgu için farklılık göstermektedir. Çünkü tümleştirme algoritmasında skorlar ön plandadır ve bu skorları da kelimeler ve atıflar belirlemektedir. Örneğin, sorgu 23'e karşılık erişilen makalelerde kelime sıklıklarının olasılıksal dağılımlarında çok baskın bir örüntü olduğu için tümleştirilmiş sıralamada "LDA baskın" bir yapı gözlenmektedir. Sıralamada 79. sıraya kadar LDA skorları pennant skorlarından açık ara daha yüksektir (Şekil 31 ve Şekil 32). Dolayısıyla tümleştirilmiş sıralamada ilk 100 makalenin 76'sı LDA algoritması aracılığıyla sıralamaya girmiştir (Şekil 33).

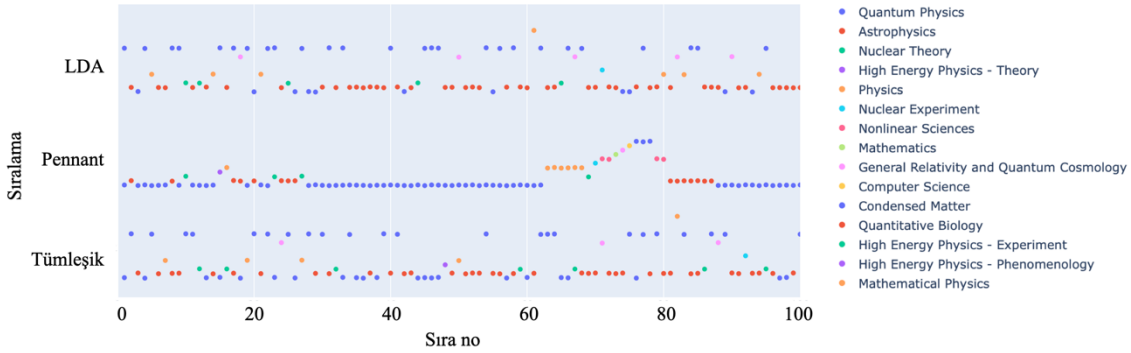


Şekil 31. Algoritmaların ilgi skorları (sorgu 23)



Şekil 32. Algoritmaların çeşitli kesme noktalarında tümleşik sıralamaya katkıları (sorgu 23)

Pennant erişim algoritması 30. sıra ile 64. sıra arasında konusu sadece *Quantum Physics* olan çalışmalara erişmiştir (Şekil 33). Sıralamada 50. sıradan sonra çoğunlukla *Quantum Physics* ve *Astrophysics* ile ilgili çalışmalar baskındır.



Şekil 33. LDA baskın tümleşik sıralama (sorgu 23)

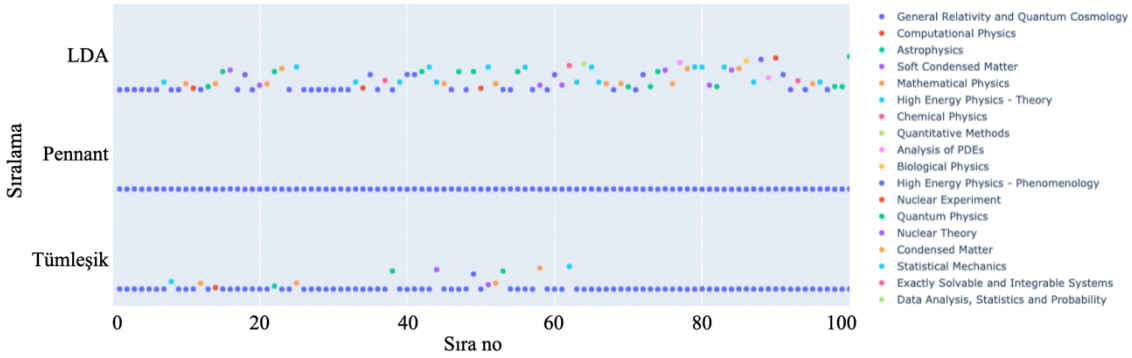
Pennant erişim algoritmasında genellikle tutarlı bir şekilde çekirdek makaleler ile aynı konuda olan makalelere ilk sıralarda erişilmektedir. Daha sonraki sıralarda ise daha az ortak atıfa sahip olan marjinal makalelere erişilmektedir. Marjinal makaleler aynı zamanda farklı disiplinlerle bağlantı kuran makalelerdir. Bu bulgu üçüncü araştırma sorusu olan "Kelime dağılımına dayalı olasılıksal konu modelleme algoritması LDA kullanılarak oluşturulan sıralamaya pennant erişim yönteminin katkıları nelerdir?" sorusunun cevabı niteliğindedir. Fakat nadir de olsa ilk sıralarda çekirdek makalelerin konularıyla örtüşen çalışmalara erişilemeyebilir ve marjinal makaleler ilk sıralarda yer alabilir. Buradaki en önemli etken atıf ağının seyrek ve dolayısıyla ortak atıf sayısı ortalamasının düşük (2) olmasıdır. Ortak atıf sayısı ortalaması azaldıkça marjinal makalelere ilk sıralarda erişilme oranı da artmaktadır.[47]

Pennant sıralamasında ilk sıralarda erişilen makalelerin çeşitliliğini etkileyen bir diğer önemli etken de çekirdek makalelerin konularındaki çeşitliliktir. Örneğin, sorgu 23 için çekirdek makalelerin konuları sırasıyla: *Soft Condensed Matter*, *Astrophysics*, *Quantum Physics*, *Statistical Mechanics* ve *Atomic Physics*'tir. Fakat oluşturulan pennant sıralamasında ilk 100'de *Soft Condensed Matter* ve *Atomic Physics* konulu herhangi bir çalışma yer almamaktadır.

[47] Ortak atıf sayısı ortalaması bilgisi yorumlanırken toplam atıf değerlerinin de dikkate alınması gerekmektedir. Örneğin, ortak atıf sayıları 40, 40 ve 40 olan makaleler için de, 118, 1 ve 1 olan makaleler için de ortak atıf sayısı ortalaması 40'tır. Fakat buradaki önemli unsur toplam atıf değerleridir. İlk üç makalenin toplam atıf değerleri ortak atıf değerlerine çok yakınsa ve ikinci setteki 118 ortak atıf alan makalenin toplam atıf sayısı çok yüksekse o zaman ilk setteki makaleler için pennant skorları ikinci setteki ortak atıfı 118 olan makaleden çok daha yüksek çıkacaktır.

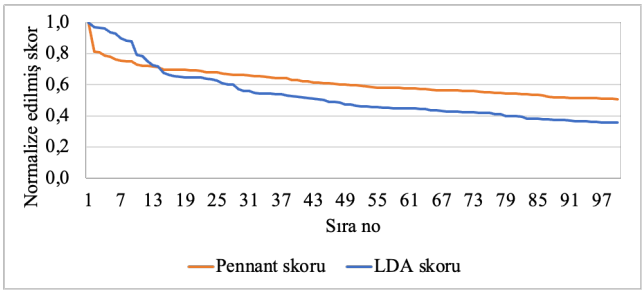
Çekirdek makalelerin farklı konularda olması pennant sıralamasında (ve dolayısıyla tümleşik sıralamada) çeşitliliğin yüksek oluşunda önemli bir etkendir.

Toplam atıf sayısı az ama ortak atıf sayısının yüksek olduğu durumlarda ise tümleşik sıralamada "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 34, pennant baskın tümleşik sıralamaya bir örnektir (sorgu 66). LDA algoritmasının eriştiği ilk 100 makale çok farklı konulardadır (20 farklı konu). Öte yandan ortak atıf ortalaması 15 olan ve çekirdek makalelerinin tümü *General Relativity and Quantum Cosmology* konusunda olan bu sorgu için pennant erişim algoritması uygulandığında ilk 100'de erişilen makalelerin tümünün çekirdek makalenin konusuyla aynı konuda olduğu görülmektedir.

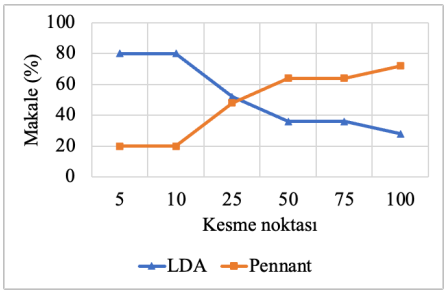


Şekil 34. Pennant baskın tümleşik sıralama (sorgu 66)

Şekil 35'te LDA ve pennant erişim algoritmalarının skorları, Şekil 36'da ise tümleşik sıralamaya LDA ve pennant erişim algoritmalarının katkısı gösterilmektedir (66. sorgu için). İlk 12 makale için kelime sıklıklarına göre olan sonuçlar (LDA) daha ön planda olmasına karşın 13. sıradan sonra ortak atıf sayılarının (pennant) ön plana çıktığı izlenmektedir. İlgi skorlarında 13. sıradaki çakışmanın yansıması tümleşik sıralamaya katkı grafiğinde de gözlenebilmektedir. Kesme noktası 25'ten sonra pennant ağırlıklı bir sıralama göze çarpmaktadır.



Şekil 35. Algoritmaların ilgi skorları (sorgu 66)



Şekil 36. Algoritmaların çeşitli kesme noktalarında tümleşik sıralamaya katkıları (sorgu 66)

Ortak atıf verilerine göre işletilen pennant erişim çıktısının kaç makaleye erişeceği kestirilememekte, çok uzun ya da çok kısa sıralamalar olabilmektedir.[48] Pennant sıralamasının kısa olduğu durumlarda tümleşik sıralamada bir noktadan sonra sadece LDA algoritması tarafından erişilen kaynaklar yer almaktadır.[49]

4.3.1. Ortak Atıf Sayılarının Algoritmaların İşleyişine Etkisi

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

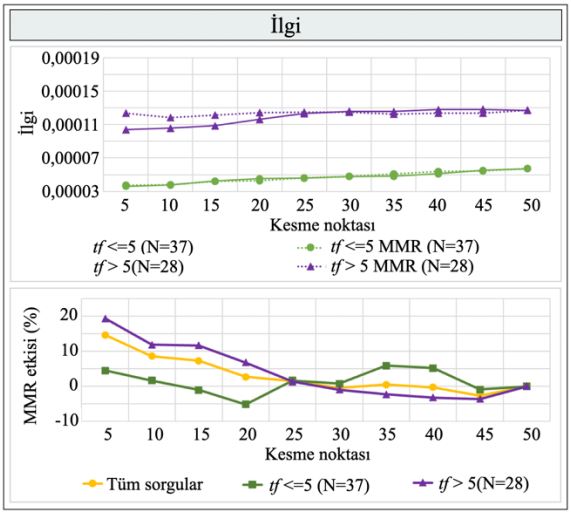
Genelde ortak atıf sayısı arttıkça ilgi oranları da artmaktadır. Örneğin, tümleşik sıralamalarda ortak atıf sayısı beşten büyük olan sorguların ilgi değeri, beş ve daha küçük olanların ilgi değerinin ortalama 2,2 katıdır (Şekil 37). MMR algoritmasının tümleşik sıralamaya etkisi kesme noktası 25'e kadar yüksek iken 25'ten sonra azalmaktadır. Etkinin neredeyse yok olduğu kesme noktası 25'te tümleşik sıralamaya marjinal olarak tanımlanabilecek ortak atıf aracılığıyla farklı disiplinlerle bağlantısı keşfedilen makaleler eklenmektedir. Ortak atıf sayısı beşten büyük

[48] Toplam 65 sorgu için liste uzunlukları maksimum 4243, minimum 22'dir (ortalama 532, ortanca 222).

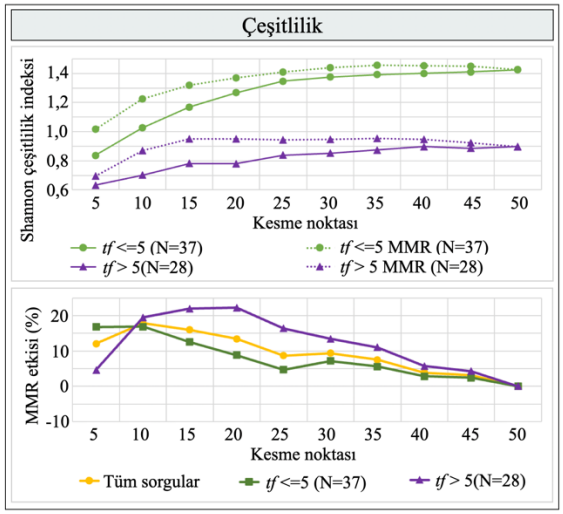
[49] Örneğin, 65. sorguda pennant algoritması çalıştırıldığında sadece 22 makaleye erişilmektedir (dolayısıyla LDA çıktısı da 22'ye sabitlenmiştir, bkz. https://mugeakbulut.com/phd/gorsellestirme/liste_konu_grafik_100.pdf, sorgu 65). Tümleşik listede ise 39. sıradaki makaleden sonra sadece LDA listesinden gelen kaynaklar yer almaktadır.

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 sıralamalarda ortak atıf sayısı beş ve beşten küçük olan sorguların eriştiği makalelerin Shannon çeşitlilik indeksi ortalaması beşten büyük olan sorgularınkinin 1,6 katıdır[50] (Şekil 38). Ortak atıf sayısı beşten büyük olan 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 sıralamalara 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 daha az 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 37. Tümleşik sıralama için ilgi değerleri ve MMR etkisi (*tf* eşiği <=5)



Şekil 38. Tümleşik sıralama için çeşitlilik değerleri ve MMR etkisi (*tf* eşiği >5)

Not: "tf": ortak atıf sayısı ortalaması.

[50] Atıf sayısı beş ve daha küçük olan sorguların eriştiği makalelerdeki tekil konu sayıları ise beşten büyük olan sorgularınkinin 1,4 katıdır.

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.

Tablo 4'te ortak atıf sayısı ortalaması beş ve beşten daha küçük olan sorgular, beşten daha büyük olan sorgular ve tüm sorgular için tümleşik sıralamanın ilk 50 sırasına LDA ve pennant algoritmalarının katkısı, Şekil 39'da ise Tablo 4'teki istatistiklere dayanan yağmur bulutu grafiği (raincloud plot) verilmektedir.

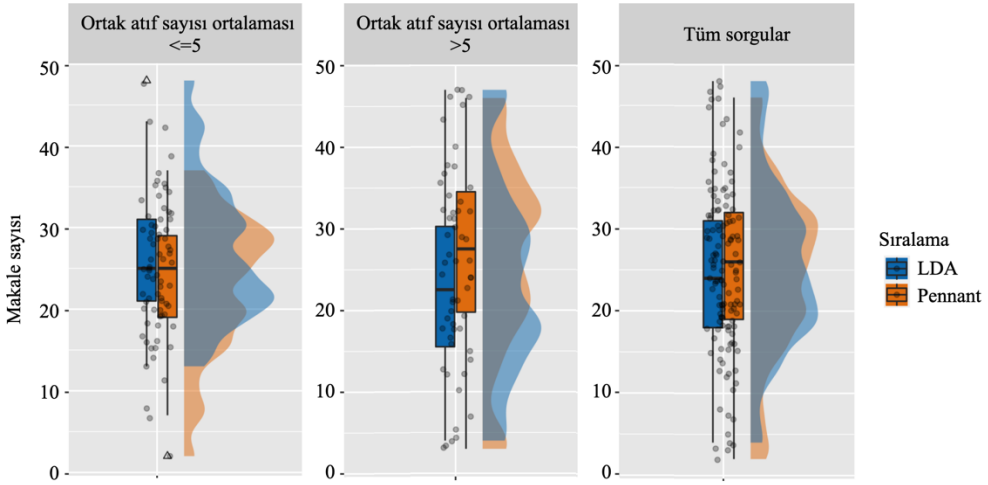
Çeyreklik değerler dikkate alındığında en ilgili makalelerin yer aldığı varsayılan ilk %25'lik çeyrekte pennant erişimin katkısı tüm sorgular için ve ortak atıf ortalaması beş ve beşten daha küçük olan sorgular için 19 makaledir. Ortak atıf sayısı beşten büyük olduğunda ise pennant erişim algoritmasının katkısı ortalama bir makale daha fazladır. Ortak atıf sayısının beş ve beşten daha küçük olduğu durumlarda LDA'nın ve pennant erişimin tümleşik sıralamadaki ilk 50 makaleye katkıları sırasıyla ortalama 27 ve 24 makaledir.

Tablo 4. Algoritmaların tümleşik sıralamaya katkılarının ortak atıflara göre karşılaştırılması

	LDA'nın katkısı			Pennant erişimin katkısı		
Tanımlayıcı istatistikler	*tf* <=5	*tf* >5	Tüm sorgular	t*f* <=5	*tf* >5	Tüm sorgular
Minimum	13,0	4,0	4,0	2,0	3,0	2,0
Yüzde 25	21,0	15,5	18,0	19,0	19,8	19,0
Ortanca	25,0	22,5	24,0	25,0	27,5	26,0
Ortalama	26,7	23,5	25,3	23,3	26,5	24,7
Yüzde 75	25,0	22,5	24,0	25,0	27,5	26,0
Maksimum	48,0	47,0	48,0	37,0	46,0	46,0

Not: "tf": ortak atıf sayısı ortalaması. Tablodaki değerler sıralamalardaki ilk 50'şer makale içindir.

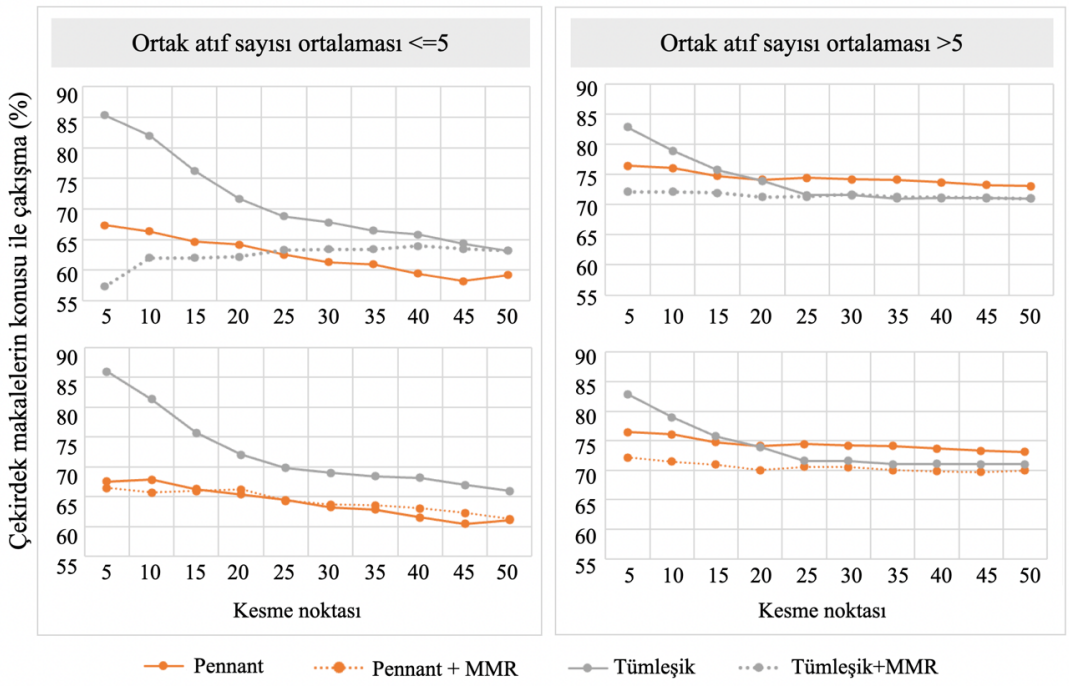
Her ne kadar sorgudan sorguya katkı oranları değişse de genel örüntüye bakıldığında ortak atıf sayısı ortalaması arttığında pennant erişim algoritmasının etkisinin de arttığı gözlenmektedir (Şekil 39). Ortak atıf sayısı beş ve beşten daha küçük olan sorgular ortak atıf sayısı beşten büyük olanlarla karşılaştırıldığında pennant algoritması aracılığıyla ilk çeyreklikte (%25'lik dilim ya da ilk 15 sıra) kabaca fazladan 1, ikinci ve üçüncü çeyreklikte ise 3'er kaynak eklenmektedir.



Şekil 39. Algoritmaların tümleşik sıralamaya katkılarının ortak atıflara göre karşılaştırılması

Şekil 40'ta tümleşik sıralama ve pennant sıralamasının ortak atıf sayısı ortalaması beş veya beşten az ya da beşten daha yüksek olmasına göre nasıl davrandıkları yer almaktadır. LDA algoritması ortak atıf sayısından bağımsız olduğu için grafiklere eklenmemiştir.

Önerilen tümleşik algoritmada LDA'nın eriştiği ilk beş çekirdek makaleden yola çıkılmaktadır. LDA'nın ilk sıralarda eriştiği makalelerin sorguyla ilgili olduğu varsayılmaktadır. Şekil 40, LDA algoritmasının da ne kadar isabetli sonuçlar getirdiğinin izlenebilmesi açısından önemlidir. Terim sıklıklarına dayanan algoritmada çeşitlilik oranının yüksek olması beklenirken tüm sorgular için ortalama %65-%68 aralığında bir çakışma söz konusudur.



Şekil 40. Ortak atıf sayılarının çakışma oranlarına etkisi

Ortak atıf ortalaması beşten büyük olduğunda (Şekil 40, sağ taraftaki grafikler) pennant erişim algoritması ilk sıralarda çekirdek makalelerin konuları ile uyumlu makalelere erişmektedir. Bu tutarlılık kararlı bir şekilde kesme noktası 50'den sonra da devam etmektedir. Ortak atıf ortalaması düşük olan sorgular için ise erişilen makaleler çekirdek makalenin konuları ile daha az uyuşmaktadır. Bu durum ortak atıf sayısı düşük olan sorgular için marjinal makalelerin (çekirdek makalenin konusuna benzemeyen) ilk sıralarda listeye girebildiği şeklinde yorumlanabilir. (Şekil 40, sol alttaki grafik). Bu bulgular "Pennant erişimin artırımlı olarak geliştirilen ilgi sıralamasına katkısı ortak atıf ve toplam atıf sayıları ile doğrudan ilgili midir? Ortak atıf sayısı az olan derlemlerde de bu yönteme başvurulabilir mi?" sorusunun (dördüncü araştırma sorusu) cevabıdır.

Tümleşik ilgi sıralaması hem pennant hem de LDA algoritması ile erişilen en ilgili makaleleri (her ne kadar farklı yöntemlerle belirleniyor olsalar da) ilk sıralara yerleştirdiği için çekirdek makalelerin konuları ile uyum oranı yüksektir. Tümleşik sıralamaya MMR uygulandığında benzer konudaki makaleler ötelendiği için marjinal ilgili makaleler ilk sıralara yerleşmektedir. Şekil 40'taki grafiklerde tümleşik sıralama için oluşan makasın nedeni budur. Ortak atıf sayısı beşten büyük olduğunda potansiyel olarak konu sayısı da arttığı için dalgalanma gözlenmesi normaldir.

4.4. İLGİ SIRALAMALARININ KİŞİSELLEŞTİRİLMESİ

4.4.1. İlgi ve Çeşitliliğe Göre Ağırlıklandırma

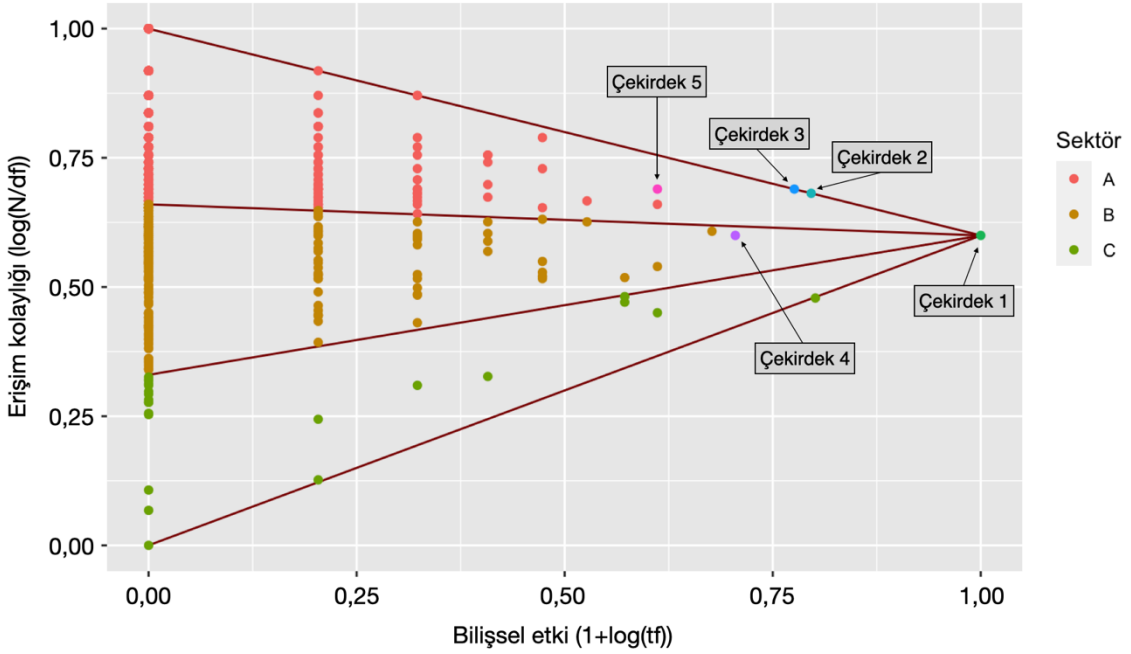
LDA ve pennant erişim algoritmaları işleyiş biçimleri farklı olduğu için farklı özelliklerde sıralamalar oluşturmaktadır. İki algoritmanın çıktıları tümleştirildiğinde ise ilgi ve çeşitliliğin artırımlı olarak geliştirildiği sıralamalar elde edilmektedir. Kullanıcıya sunulan bu sıralama kullanıcının isteğine göre yeniden sıralanabilir. Çeşitlilik oranının yüksek olduğu bir sıralama için LDA'nın eriştiği makalelere, ilgi oranının yüksek olduğu bir sıralama için ise pennant erişim algoritmasının eriştiği makalelere ağırlık verilir ve yeniden sıralama yapılır.

4.4.2. Pennant Diyagramları

İlgi ve çeşitliliğe göre ağırlıklandırma dışında, pennant erişimden elde edilen sektör bilgileri de gerektiğinde ağırlıklandırılarak ilgi sıralamalarına yansıtılmakta ve bu sayede araştırmacıların istediği özellikte sıralamalar oluşturulmaktadır. Pennant diyagramı gösterimi ile sıralamada gözlenmesi mümkün olmayan ilişkiler (örneğin, yazarlar ve makaleler arasındaki konusal ilişkiler, makalelerin etkisi vb. gibi) kolaylıkla gözlenebilmektedir. Diyagramlar bu yönüyle araştırmacıların literatürü izlemesini kolaylaştırmaktadır.

Sorgu 42 için oluşturulan örnek pennant diyagramları Şekil 41'teki gibidir. Pennant diyagramları üç bölüme ayrılıp yorumlanmaktadır (bkz. Bölüm 2.3 ve Şekil 3). *A* sektöründe yer alan makalelerin çekirdek makalenin ardılları, *B* sektöründekilerin akranları, *C* sektöründe olanların ise öncülleri olması beklenmektedir. Bu araştırma özelinde beş çekirdek makale bulunduğu için çekirdek makaleler aynı sorgu için potansiyel olarak farklı konularda olabilmektedir. Bu yüzden sektör içeriklerini yorumlarken özellikle çizgilere yakın olan makalelere dikkat etmek gerekmektedir. Fakat yine de çekirdek makaleler diyagramda büyük çoğunlukla *x* ekseninin en sağında ve *y* ekseninin ortalarında yer aldığından (toplam atıf ve ortak atıf değerleri en yüksek olan makaleler) sektör içerikleri pennant diyagramlarında tanımlanan içeriklerle uyuşmaktadır.[51]

Örnek olarak sorgu 42 için pennant erişim algoritmasıyla erişilen makaleler ayrıntılı olarak incelenmiştir. Bu sorgu için çekirdek makalelerin hepsi astrofizik konusundadır ve yayın yılları ortancası 2002'dir (sırasıyla 2002, 2004, 2002 ve 1998).



Şekil 41. Sorgu 42 için pennant diyagramı[52]

Çekirdek makalelerin özelliklerine göre (konusu, yayın yılı vs.) sektörleri yorumlamak bazen yanıltıcı olabilir. Yine de izlenen yöntem kelime torbası (bag of words) yöntemine benzer bir

[51] Aslında diyagramdaki her bir makalenin hangi çekirdek makale aracılığı ile diyagrama eklendiği bilgisi de elimizde bulunmaktadır. Dolayısıyla seçilen bir çekirdek makale için de kolaylıkla daha detaylı pennant diyagramları çizdirilebilir.

[52] Tüm sorgular için oluşturulan pennant diyagramları için bkz. http://mugeakbulut.com/phd/gorsellestirme/pennant_diagrams/

yaklaşım içerdiği için sektörlerde yer alan çalışmalar daha önceden tanımlanan özellikleri taşımaktadır.

İlgili sektörlerde yer alan makalelerin konularının sektörlere göre dağılımı ve sektörler için ortanca yıl bilgileri Tablo 5'te verilmiştir. İlgili sektörlerdeki makaleler için beklenen özellikler aşağıda yer almaktadır.

Çekirdek makaleler ile spesifik olarak aynı konuda olan ardıl çalışmalar *A sektöründe* konumlanmaktadır. Bu sorgu için *A* sektöründe yer alan 263 makalenin 236'sının (%89,7) konusu doğrudan çekirdek makalelerin konusu olan *Astrophysics* ile ilgilidir. Bu makalelerin yayın yılı ortancası 2004'tür. Çekirdek makalelerin ortancasının 2002 olması *A* sektöründeki makalelerin ardıl çalışmalar olduğu bulgusunu desteklemektedir.

Tablo 5. Sorgu 42 için makalelerin sektörlere dağılımı

Konu	A Sektörü	B Sektörü	C Sektörü	Çekirdek
*Astrophysics	236	201	11	5
High Energy Physics - Theory	22	47	9	-
Algebraic Geometry	1	-	-	-
High Energy Physics - Phenomenology	3	13	2	-
General Relativity and Quantum Cosmology	1	1	-	-
Classical Physics	-	1	-	-
Condensed Matter	-	2	-	-
Atomic Physics	-	1	-	-
Data Analysis, Statistics and Probability	-	-	1	-
Toplam	263	266	23	5
Ortanca (yıl)	2004	2002	1998	2002

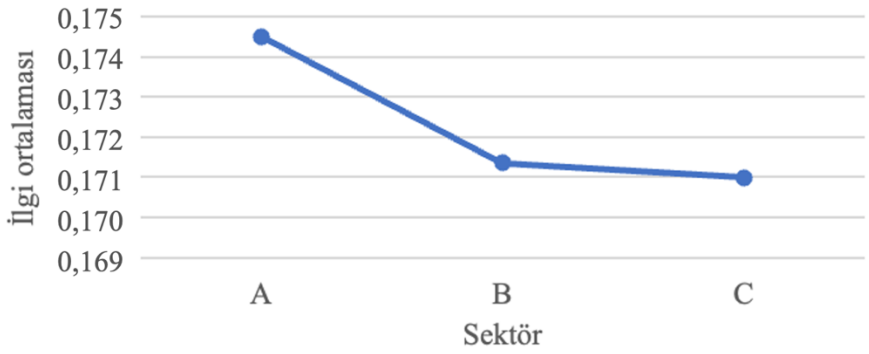
*çekirdek makalelerin konusu

B sektöründe yer alan makaleler çekirdek makalelerin konusu ile spesifik olarak doğrudan ilgili olmayan, daha genel konuları olan ve genellikle çekirdek makale üzerine inşa edilen çalışmalardan oluşmaktadır (derivative). Diğer bir deyişle, *B* sektöründeki makaleler terim eşleşmesi içermekten ziyade daha derin ilgi özelliği taşırlar. Bu sorgu için *B* sektöründe yer alan 266 makalenin %75,5'i (201) konusu çekirdek makalenin konusuyla aynıdır. Bu oranın *A* sektöründekinden (%89,7) düşük olmasının sebebi çekirdek makale ile doğrudan ilgili olmayan çalışmaların da bu sektörde yer almasıdır. Benzer olarak *B* sektöründe yer alan makaleler sayıca *A* sektöründeki makalelere yakın olmasına karşın *B* sektörü tekil konu açısından daha farklı

konularda makaleler de içermektedir (*A* sektöründe yer alan 263 makale beş farklı konudadır, *B* sektöründe yer alan 266 makale ise 7 farklı konudadır). *B* sektöründeki makalelerin de çekirdek makalelerin de yayın yılı ortancası 2002'dir. Bu bulgu *B* sektöründe akran yazarların çalışmalarının yer aldığı varsayımı ile uyuşmaktadır.

C sektöründe ise genelde rehber niteliğinde olan öncül (prior) makaleler yer almaktadır. Çekirdek makaleler genellikle bu sektördeki makalelerin üzerine inşa edilmektedir. Makalelerin yıllarının ortancalarına bakıldığında (1998) en eski makalelerin bu sektörde yer aldığı görülmektedir.

A, *B* ve *C* sektörlerinde yer alan kaynakların ilgi değerlerini karşılaştırmak için 42. sorgu özelinde bir hesaplama yapılmış, erişilen 552 makalenin çekirdek makalelere uzaklıkları hesaplanmıştır. Her bir makale için elde edilen beş değerin ortalaması alınarak ilgi değeri belirlenmiştir. Beklendiği gibi *A* sektöründeki makalelerin çekirdek makalelerle en ilgili makaleler olduğu, *B* sektöründekilerin derin ilgi özelliği taşıdığı ve *C* sektöründekilerin çok daha genel konulardaki çalışmalar olduğu saptanmıştır (Şekil 42).



Şekil 42. Sektörlere göre ortalama ilgi değerleri

Tablo 6'da sorgu 42 için konusal ilgi (topical relevance) açısından çekirdek makalelerin başlıklarında geçen terimlerin pennant erişim algoritması tarafından erişilen makalelerin başlıklarındaki terimlerle eşleşme oranları verilmektedir. *A* sektöründe yer alan 263 makalenin %34'ünün başlığındaki terimlerden en az biri çekirdek makalelerin başlıklarında da yer almaktadır. *A* sektöründeki makaleler spesifik olarak çekirdek makale ile ilgili olduğu için bu oranın yüksek çıkması normaldir. *B* sektöründe yer alan makaleler ise daha derin ilgi özelliği taşıdıkları için terim çakışma oranları daha düşüktür. *C* sektöründeki çoğu çalışmanın çekirdek makale ile ilgisi ancak bu alanda uzman kişiler tarafından fark edilebilmektedir (White, 2015). Dolayısıyla buradaki makaleler için çakışma oranı çok daha düşüktür.

Tablo 6. Makale başlıklarındaki terimlerin çakışma oranları

Sektör	Toplam makale sayısı	Başlığında çekirdek makaledeki terimlerin geçtiği makaleler	
		Sayı	%
A	263	88	33,5
B	266	40	15,0
C	23	2	8,7

Bu bölümdeki bulgular araştırmanın beşinci araştırma sorusu olan "Arama sırasında araştırmacının önceliğine göre belirli özelliklerin (örneğin arama yapılan konu üzerine inşa edilen makaleler) ön planda çıktığı ilgi sıralamaları anlık olarak oluşturulabilir mi?" sorusunun da cevabı niteliğindedir. Seyrek bir atıf ağında bile kullanıcının ihtiyacına göre ilgili ya da çeşitli makalelerin olduğu sıralamalar oluşturulabilmekte, pennant erişim algoritmasındaki sektör bilgileri daha detaylı olarak sıralamalara yansıtılabilmektedir.

4.4.3. Tüm Sorgular İçin Pennant Diyagramı Sonuçları

Pennant diyagramları tüm sorgular için oluşturulup incelendiğinde çekirdek olarak erişilen makalelerin yıl ortalamasının 2002 olduğu belirlenmiştir. *A*, *B* ve *C* sektörlerinde yer alan makalelerin yıl ortancaları sırasıyla 2003, 2001 ve 1999'dur. *A*, *B* ve *C* sektöründe erişilen makalelerin sırasıyla %78, %75 ve %69'unun konuları çekirdek makale konuları ile eşleşmektedir. Diğer bir deyişle *C* sektöründe yer alan makaleler çekirdek makalelerin konularına göre daha çeşitli konulardadır. Toplu bulgular da *A* sektöründe ardıl, *B* sektöründe akran ve *C* sektöründe öncül çalışmalar olması gerektiği; çekirdek makalelerin *C* sektöründe yer alan makalelerin üzerine inşa edilmiş çalışmalar olduğu ve *A* sektöründeki çalışmaların çekirdek çalışmayla ilgisinin net olduğu yönündeki diğer çalışmaların bulgularını desteklemektedir (Akbulut 2016, Akbulut ve diğerleri 2020, Tonta ve Özkan Çelik, 2013; White, 2007a, 2007b, 2009, 2010, 2011, 2015; White ve Mayr, 2013).

Genel olarak 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 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. Bu bulgu altıncı araştırma sorusu olan "LDA konu modelleme algoritması pennant erişim ile desteklenerek, LDA modelinin kaçırdığı temel kaynaklara ve ilgili terimin kullanıldığı diğer alanlardaki çalışmalara erişim sağlanabilir mi?" sorusunun cevabıdır. Toplam 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 algoritması, bu çalışmanın 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).

Bu bölümde sunulan bulgular araştırmada önerilen atıflara ve ortak atıflara dayanan pennant erişim ile desteklenen tümleşik bilgi erişim modelinin incelenen hemen hemen bütün ölçütlere göre (ilgi, konu çeşitliliği, yenilik, kapsama, vd.) ilgi sıralamalarını artırımlı olarak geliştirdiğini göstermektedir. "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 konuyu tüm yönleri ile ele alan artırımlı olarak geliştirilmiş ilgi sıralamaları oluşturulabilir" şeklinde formüle edilen araştırmamızın temel hipotezi (H1) desteklenmektedir. Böylece daha önce sadece örnek olaylara dayanan pennant erişimin bilgi erişim sistemlerinin performansını çeşitli açılardan artırdığı yönündeki bulgular (Akbulut, 2016; Akbulut ve diğerleri, 2020; Larsen, 2008; Schneider ve diğerleri, 2007; Tonta ve Özkan Çelik, 2013; White, 2007a, 2007b, 2009, 2010, 2015) daha büyük bir derlem üzerinde ve daha çok sayıda sorguya dayanarak doğrulanmış olmaktadır.

"İlgi sıralamaları kullanıcının ihtiyacına göre; öncül çalışmalar, spesifik makaleler, makalelerin alana etkisi gibi özellikleri öncelenecek şekilde yeniden sıralanabilir" şeklinde oluşturulan ve ilk kez test edilen araştırmamızın ikinci hipotezi (H2) de desteklenmektedir. Başka bir deyişle, LDA, pennant erişim ve tümleşik algoritmalar yardımıyla oluşturulan ilgi sıralamalarının kullanıcıların isterlerine göre ilgi ya da çeşitlilik oranları daha yüksek olacak şekilde kişiselleştirilebilmektedir. Pennant erişimdeki sektörel bilgiler erişim sonuçları öncül, akran ya da ardıl çalışmalardan oluşacak şekilde ağırlıklandırılarak yeni sıralamalar elde edilebilmektedir.

5. BÖLÜM: SONUÇ VE ÖNERİLER

Araştırma sonuçlarına göre, kelimeler arası ilişkilere odaklanan LDA konu modelleme algoritması ile oluşturulan ilgi sıralamaları ilk sıralarda çoğunlukla marjinal ilgili makaleleri içermektedir. 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 makaleler yer almaktadır. Pennant erişim ile sonradan eklenen marjinal ilgili makaleler ise 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. Çalışma kapsamında elde edilen bulgular hipotezleri destekler niteliktedir. LDA algoritması pennant erişim algoritması ile artırımlı olarak geliştirildiğinde 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.**

LDA algoritmasında farklı sınıflarda benzer olasılıklara sahip sözcükler fazla bilgi içermezler. Sözcüklerin ayırt edici ve kullanışlı olabilmeleri için bir (ya da az sayıda) sınıf sözcük için olasılıklar yüksekken geri kalan sözcükler için düşük olması gerekir. Oysa pennant erişim algoritmasında toplam atıf sayısı ve ortak atıf sayısı birbirine yakın ve ortak atıf sayısı nispeten yüksek olan makaleler için ayırt edicilik de daha yüksektir. Önerilen yöntemde tümleşik sıralamada her iki algoritma için de bu özellikteki makaleler üst sıralarda yer almıştır. Tümleştirme aşamasında her iki algoritmadan elde edilen skorlar normalize edilerek kullanıldığı için bir sorguya karşılık erişilen makalelerde kelime sıklıklarının olasılıksal dağılımlarında baskın bir örüntü varsa tümleştirilmiş sıralamada "LDA baskın", ortak atıf yoğunluğu yüksek bir örüntü varsa "pennant baskın" bir yapı gözlenmiştir. Bu açıdan bakıldığında, **önerilen yöntem kelime sıklığı ve atıf bilgilerini orijinal yapıyı koruyacak şekilde bütünleştirmektedir.** Algoritmalardaki ayırt edicilik özellikleri saklı kaldığı için tümleşik sıralamada herhangi bir ön yargı söz konusu değildir. Önerilen yöntemde sistematik bir birleştirme yapısı olmadığı için bu yöntem benzer amacı olan diğer algoritmalardan (örneğin, MMR) ayrılmaktadır.

Konu modellemesi makalelerin sadece özet ve başlık bölümlerine uygulandığı için süreç hızlandırılmış, seyrek bir atıf ağında bile istenen özellikte sıralamalar elde edilebilmiştir. Pennant erişim algoritması ortak atıf sayısı yüksek olduğunda daha başarılı sonuçlar vermesine karşın, ortak atıf oranı düşük olan iSearch derleminde bile amacına ulaşmıştır. LDA algoritması pennant erişim algoritmasıyla artırımlı olarak geliştirilerek alternatif ve kişiselleştirilebilir özellikte ilgi sıralamaları oluşturulmuştur. Öte yandan, görselleştirme için pennant erişim diyagramlarından yararlanılması kullanıcıların literatürü izlemelerini kolaylaştırmaktadır (Akbulut ve diğerleri,

2020). Bu sonuçlar önerilen yöntemin atıf veri tabanlarına entegre edilerek literatür taramaları için daha uygun ilgi sıralamaları elde edilebileceğini göstermektedir.

5.1. ÇALIŞMANIN SINIRLILIKLARI VE GELECEKTE YAPILMASI ÖNERİLEN ÇALIŞMALAR

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 mümkün olduğunca artırılmış 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.

Çalışma kapsamında sadece ilgi sıralamalarının artırımlı olarak geliştirilmesi üzerine odaklanılmıştır. Bunun ötesinde önerilen yöntem için farklı senaryolarla geçerlik ve uygunluk değerlendirmeleri yapılmasında fayda vardır (Herlocker, Konstan, Terveen ve Riedl, 2004). 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 dâhil 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, Beeferman ve Huang, 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.

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.

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

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, Antons ve Salge, 2021). Pennant diyagramları hem araştırmacıların literatürü izlemelerini kolaylaştırılabilir hem de bir çalışmanın belli bir alanda ya da farklı alanlardaki çalışmaları nasıl etkilediği belirlenebilir.

Önerilen ilgi sıralaması yöntemini durağan bir derlem (iSearch) üzerinde test etmemize karşın, bu yöntemin zamanla dinamik yapıdaki atıf dizinleri ve web aramaları için 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. Öte yandan özeti olmayan ya da çok kısa özeti olan makaleler için sorgu ya da özet genişletme teknikleri kullanılabilir (Jeong, Baek, Park ve Park, 2021). Kısa belgelerin genellikle tek bir konu hakkında olduğu hipotezinden yola çıkan yaklaşımlar izlenebilir (Efron, Organisciak ve Fenlon, 2012). Kısa belgeler, terim sıklığı bilgisi açısından çok az verim sağladığından, belge genişletme özellikle konu aşamasında modellemede yardımcı olabilir. Önerilen algoritmadaki çekirdek makale sayısı (5) ön analizler sonucu belirlenmiştir. Fakat özellikle güncel kaynakların da yer aldığı dinamik atıf dizinleri söz konusu olduğunda ortak atıf çıkmadığı durumlarla karşılaşılabilir. Bu durumda çekirdek kaynak sayısı artırılarak LDA algoritması artırımlı olarak desteklenebilir.

Benzer şekilde dinamik yapıdaki web aramaları için de atıf verileri yerine web sayfaları arasındaki bağlantılar kullanılarak aynı yöntemle ilgi sıralamalarında artırımlı geliştirme yapılabilir. Bu yöntem belirli platformlarda *öneri sistemi* olarak da kullanılabilir.

Bilindiği gibi, ilgi sıralamalarını geliştirmek için yapay ve derin öğrenme teknolojilerinden de yararlanılmaktadır (Fiorini ve diğerleri, 2018; Lages ve Carvalho, 2020; Liu, 2009, s. 226; Zong ve Huang, 2014, s. 155). Bu bağlamda önerilen algoritma için veriler (eğitim verileri) çözümlenip öngörü yapmaya yarayacak bilgiye dönüştürülerek öğrenen sistemler oluşturulabilir. Yapay öğrenme algoritmalarından birisi olan yapay sinir ağları (artificial neural networks – ANN) da pennant erişim algoritması ile desteklenerek ilgi sıralamalarının geliştirilmesinde kullanılabilir.

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* (Yüksek lisans tezi, Hacettepe Üniversitesi). http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/3529

Akbulut, M. ve Tonta, Y. (basım aşamasında). İlgi sıralamalarının artırımlı olarak geliştirilmesi: Pennant erişimle desteklenen yeni bir yöntem önerisi. *Türk Kütüphaneciliği***,** *36*(2).

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

Barwise, J. (1989). *The situation in logic. CSLI Lecture Notes 17*. Stanford University Center for the Study of Language and Information.

Barwise, J. (1993). Constraints, channels and the flow of information. P. Aczel, D. Israel, Y. Katagari ve S. Peters (Yay. haz.). *Situation theory and its applications-Vol. III* içinde (s. 3-27). Stanford University Center for the Study of Language and Information.

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

Berberich, K., Vazirgiannis, M. ve Weikum, G. (2005). Time-Aware authority ranking. *Internet Mathematics, 2*(3), 301-332. https://doi.org/10.1080/15427951.2005.10129110

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

Bornmann, L. ve Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222. https://doi.org/10.1002/asi.23329.

Boualili, L., Moreno, J. ve Boughanem, M. (2022). *Highlighting exact matching via marking strategies for ad hoc document ranking with pretrained contextualized language models.* Research Square. https://doi.org/10.21203/rs.3.rs-550456/v1

Bougioukas, K. I., Vounzoulaki, E., Mantsiou, C. D., Savvides, E. D., Karakosta, C., Diakonidis, T., Tsapas A. ve Haidich, A. B. (2021). Methods for depicting overlap in overviews of systematic reviews: An introduction to static tabular and graphical displays. *Journal of Clinical Epidemiology*, *132*, 34-45. https://doi.org/10.1016/j.jclinepi.2020.12.004

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

Börner, K., Chen, C. ve Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179-255. https://cns.iu.edu/docs/publications/2003-borner-arist.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

Carpi, L. C., Schieber, T. A., Pardalos, P. M., Marfany, G., Masoller, C., Díaz-Guilera, A. ve Ravetti, M. G. (2019). Assessing diversity in multiplex networks. *Scientific Reports, 9*(4511), 1-12. https://doi.org/10.1038/s41598-019-38869-0

Carpineto, C., D'Amico, M. ve Romano, G. (2012). Evaluating subtopic retrieval methods: Clustering versus diversification of search results. *Information Processing & Management*, *48*(2), 358-373. https://doi.org/10.1016/j.ipm.2011.08.004

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

Chan, S. H. (2021). *Introduction to probability for data science*. Michigan Publishing. https://probability4datascience.com/index.html

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

Clark, B. (2013). *Relevance theory.* Cambridge University Press. https://doi.org/10.1017/CBO9781139034104

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S. ve MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 659-666). https://dl.acm.org/doi/pdf/10.1145/1390334.1390446

Comins, J. A. ve Leydesdorff, L. (2016a). *Identification of long-term concept- symbols among citations: Can documents be clustered in terms of common intellectual histories?* arXiv. http://arxiv.org/abs/1601.00288

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

Cormack, G. V. ve Grossman, M. R. (2017). Navigating imprecision in relevance assessments on the road to total recall: Roger and me. *SIGIR 2017: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 5-14). ACM. https://doi.org/10.1145/3077136.3080812

Cossock, D. ve Zhang, T. (2008). Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, *54*(11), 5140-5154. https://doi.org/10.1109/TIT.2008.929939

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://asistdl.onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9

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

Devlin, K. (1991). *Logic and information*. Cambridge University Press.

Dretske, F. (1983). Précis of knowledge and the flow of information. *Behavioral and Brain Sciences, 6*(1), 55-63. https://doi.org/10.1017/S0140525X00014631

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. ve Zemel, R. (2012) Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* içinde (s. 214-226). ACM. https://doi.org/10.1145/2090236.2090255

Efron, M., Organisciak, P. ve Fenlon, K. (2012). Improving retrieval of short texts through document expansion. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 911-920). ACM. https://doi.org/10.1145/2348283.2348405

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

El-Arini, K., Veda, G., Shahaf, D. ve Guestrin, C. (2009). Turning down the noise in the blogosphere. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* içinde (s. 289-298). ACM. https://dl.acm.org/doi/10.1145/1557019.1557056

Fenton, J. E., Roy, D., Hughes, J. P. ve Jones, A. S. (2002). A century of citation classics in otolaryngology-head and neck surgery journals. *The Journal of Laryngology and Otology*, 116, 494-498. https://pubmed.ncbi.nlm.nih.gov/12238666/

Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., Osipov, M., Kholodov, M., Ismagilov, R., Mohan, S., Ostell, J. ve Lu, Z., (2018). Best Match: New relevance search for PubMed. *PLOS Biology*, 16, e2005343. https://doi.org/10.1371/journal.pbio.2005343

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

Garfield, E. (2001). From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography: A citationist's tribute to Belver C. Griffith. https://garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf

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

Ginsparg, P. (1988). *Applied conformal field theory*. arXiv. https://arxiv.org/abs/hep-th/9108028

Ginsparg, P. (2016). Preprint déjà vu. *The EMBO Journal*, *35*(24), 2620-2625. https://doi.org/10.15252/embj.201695531

Ginsparg, P. ve Glashow, S. (1986). *Desperately seeking superstrings.* arXiv. https://arxiv.org/abs/physics/9403001

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

Grant, R. M. (1996a). Toward a knowledge-based theory of the firm. *Strategic Management Journal*, 17, 109–22. https://doi.org/10.1002/smj.4250171110

Grant, R. M. (1996b). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, 7, 375–87. https://doi.org/10.1287/orsc.7.4.375

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

Harter, S. P., Nisonger, T. E. ve Weng, A. (1993). Semantic relationships between cited and citing articles in library and information science journals. *Journal of the American Society for Information Science*, *44*(9), 543-552. https://doi.org/10.1002/(SICI)1097-4571(199310)44:9<543::AID-ASI4>3.0.CO;2-F

Haunschild, R. ve Marx, W. (2019). *Discovering seminal works with marker papers.* arXiv. https://arxiv.org/abs/1901.07352

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

Heibi, I., Peroni, S. ve Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, 121, 1213–1228. https://doi.org/10.1007/s11192-019-03217-6

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

Hornik, K. ve Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30. https://epub.wu.ac.at/3987/1/topicmodels.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

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 3-50. https://doi.org/10.1108/eb026960

Järvelin, K. ve Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422-446. https://dl.acm.org/doi/pdf/10.1145/582415.582418

Jeong, S., Baek, J., Park, C. ve Park, J. C. (2021). *Unsupervised document expansion for information retrieval with stochastic text generation*. arXiv. https://arxiv.org/abs/2105.00666

Jiang, Z., Liu, X. ve Gao, L. (2015). Chronological citation recommendation with information-need shifting. *Proceedings of the 24th ACM International on Conference on Information*

and Knowledge Management içinde (s. 1291-1300). ACM. https://doi.org/10.1145/2806416.2806567

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

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363-375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kaminskas, M. ve Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, *7*(1), 1-42. https://doi.org/10.1145/2926720

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. (2012). *Recommendation on academic networks using direction aware citation analysis*. arXiv. https://arxiv.org/pdf/1205.1143.pdf

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

Lages, J. ve Carvalho, J.P. (2020). Relevance ranking for web search. *29th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2020, Glasgow, UK, 19-24 July 2020* içinde (s. 1-8). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9177802

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, 5-13. https://www.jstor.org/stable/3545743

Larsen, B. (2002). Exploiting citation overlaps for Information Retrieval: Generating a boomerang effect from the network of scientific papers. *Scientometrics,* 54, 155–178. https://doi.org/10.1023/A:1016011326300

Larsen, B. (2008). *Informetrics and IR*. Nordic Research School in Information Studies (NORSLIS), Umea, Sweden. http://itlab.dbit.dk/*blar/files/Norslis_Umea-june2008_BL2.ppt

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

Leonhardt, J., Rudra, K., Khosla, M., Anand, A. [Abhijit]. ve Anand, A. [Avishek]. (2021). *Fast Forward Indexes for Efficient Document Ranking.* arXiv. https://arxiv.org/pdf/2110.06051.pdf

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, T. Y. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, *3*(3), 225-331. https://www.nowpublishers.com/article/Details/INR-016

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

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

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

McKiernan, G. (2000), arXiv.org: The Los Alamos National Laboratory e-print server. *International Journal on Grey Literature, 1*(3), 127-138. https://doi.org/10.1108/14666180010345564

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_StatisticalNLP.pdf

Manning, C., Raghavan, P. ve Schütze, H. (2008). *An introduction to information retrieval*. Cambridge University Press. http://nlp.stanford.edu/IR-book/pdf/irbookprint.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

Marx, W., Bornmann, L., Barth, A. ve Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751-764. https://doi.org/10.1002/asi.23089

Maslov, S. ve Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, *28*(44), 11103-11105. https://doi.org/10.1523/JNEUROSCI.0002-08.2008

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

Montazeralghaem A., Rahimi R. ve Allan J. (2020). Relevance Ranking Based on Query-Aware Context Analysis. M. J. Jose ve diğerleri (Yay. Haz.). *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol. 12035* içinde (s. 446-460). Springer, Cham. https://doi.org/10.1007/978-3-030-45439-5_30

Mowshowitz, A. ve Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing & Management, 38*(1), 141–156. https://doi.org/10.1016/S0306-4573(01)00020-6

Nabiyev, Y. (2013). *Teoriden uygulamaya algoritmalar.* Seçkin.

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

Nuray, R. ve Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management*, *42*(3), 595-614. https://doi.org/10.1016/j.ipm.2005.03.023

Oral, B. ve Eryiğit, G. (2022). Fusion of visual representations for multimodal information extraction from unstructured transactional documents. *International Journal on Document Analysis and Recognition.* https://doi.org/10.1007/s10032-022-00399-3

Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J. ve Cheng, X. (2017). Deeprank: A new deep architecture for relevance ranking in information retrieval. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* içinde (s. 257-266). ACM. https://dl.acm.org/doi/pdf/10.1145/3132847.3132914?casa_token=08zPqim61f0AAAAA:WVv4VRH8S9BVQ9l6lh3bA7TIhqKDpa1K4lICsYrmdc_n0EjGGPuduXLZUC4qnHLzQY7YbMXc43HG-Q

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

Pinski, G. ve Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, *12*(5), 297-312. https://doi.org/10.1016/0306-4573(76)90048-0

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

Portenoy, J. (2021). *Harnessing Scholarly Literature as Data to Curate, Explore, and Evaluate Scientific Research.* (Doktora Tezi, Washington Üniversitesi). https://digital.lib.washington.edu/researchworks/handle/1773/47601

Portenoy, J. ve West, J. D. (2020). Constructing and evaluating automated literature review systems. *Scientometrics*, *125*(3), 3233-3251. https://doi.org/10.1007/s11192-020-03490-w

Radlinski, F., Bennett, P. N., Carterette, B. ve Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. *ACM SIGIR Forum, 43*(2), 46-52. https://dl.acm.org/doi/pdf/10.1145/1670564.1670572?casa_token=2F_1Z3DD17wAAAAA:jLdJy-jPTdYZdVTMEXedaQI-dw_Xr4Idy254DIb2H5o65eZeAXPe1NAvcAoj6LkANcMe2t7X8YKdLQ

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

Rafols, I. ve Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82,* 263–287. https://doi.org/10.1007/s11192-009-0041-y

Rao, R. (2004). From IR to Search, and Beyond: Searching has come a long way since the 60s, but have we only just begun? *Queue*, *2*(3), 66-73. https://dl.acm.org/doi/pdf/10.1145/1005062.1005070

Ren, P., Chen, Z., Ma, J., Zhang, Z., Si, L. ve Wang, S. (2017). Detecting temporal patterns of user queries. *Journal of the Association for Information Science and Technology*, *68*(1), 113-128. https://doi.org/10.1002/asi.23578

Ren, P., Chen, Z., Ma, J., Wang, S., Zhang, Z., Ren, Z. ve Ma, T. (2018). User session level diverse reranking of search results. *Neurocomputing*, *274*, 66-79. https://doi.org/10.1016/j.neucom.2016.05.087

Ren, P., Chen, Z., Song, X., Li, B., Yang, H. ve Ma, J. (2013). Understanding temporal intent of user query based on time-based query classification. *CCF International Conference on Natural Language Processing and Chinese Computing* içinde (s. 334-345). Springer. https://doi.org/10.1007/978-3-642-41644-6_31

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

Rousseau, R., Zhang, L. ve Hu, X. (2019). Knowledge integration: Its meaning and measurement. W. Glänzel, H. F. Moed, U. Schmoch ve M. Thelwall (Yay. haz.). *Springer handbook of science and technology indicators* içinde (s. 69-94). Springer. https://doi.org/10.1007/978-3-030-02511-3_3

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

Say, C. (2018). *50 soruda yapay zekâ*. Bilim ve Gelecek Kitaplığı.

Schneider, J. W., Larsen, B. ve Ingwersen, P. (2007). Pennant diagrams, what is it, what are the possibilities and are they useful? The Nordic Workshop on Bibliometrics and Research Policy, Copenhagen, September 13-14, 2007. https://pdfs.semanticscholar.org/b674/7068496b8b72a5b017281b2dce75844b1e3d.pdf

Schuler, J., Falls, Z., Mangione, W., Hudson, M. L., Bruggemann, L. ve Samudrala, R. (2022). Evaluating the performance of drug-repurposing technologies. *Drug Discovery Today, 27*(1), 49-64. https://doi.org/10.1016/j.drudis.2021.08.002

SCImago JCR. (2007). SJR — *SCImago Journal & Country Rank.* http://www.scimagojr.com

Shannon, C. E. ve Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana. https://pure.mpg.de/rest/items/item_2383164/component/file_2383163/content

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269. https://doi.org/10.1002/asi.4630240406

Sparck Jones, K., Walker, S. ve Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing & Management*, *36*(6), 779–808. https://doi.org/10.1016/s0306-4573(00)00015-7

Spellerberg, I. F. ve Fedor, P. J. (2003). A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. *Global Ecology and Biogeography*, *12*(3), 177-179. https://doi.org/10.1046/j.1466-822X.2003.00015.x

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

Strohman, T., Croft, W. B. ve Jensen, D. (2007). Recommending citations for academic papers. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 705-706). ACM. https://doi.org/10.1145/1277741.1277868

Sugimoto, C. ve Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

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. (1992). *An Analysis of Search Failures in Online Library Catalogs*. (Doktora Tezi, Kaliforniya Üniversitesi, Berkeley). http://yunus.hacettepe.edu.tr/~tonta/Yayinlar/tonta-phd-dissertation-1992.pdf

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

Tonta, Y. ve Özkan Çelik, A. E. (2013). Cahit Arf: Exploring his scientific influence using social network analysis, author co-citation maps and single publication h index. *Journal of Scientometric Research*, *2*, 37-51. https://www.jscires.org/article/38

Traag, V. A. (2021). Inferring the causal effect of journals on citations. *Quantitative Science Studies*, *2*(2), 496–504. https://doi.org/10.1162/qss_a_00128

van Rijsbergen, C. J. ve Lalmas, M. (1996). Information calculus for information retrieval. *Journal of the American Society for Information Science*, *47*(5), 385-398. https://doi.org/10.1002/(SICI)1097-4571(199605)47:5<385::AID-ASI6>3.0.CO;2-S

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

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management, 36*(5), 697–716. https://doi.org/10.1016/S0306-4573(00)00010-8

Vorontsov, K. ve Potapenko, A. (2014). Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *International Conference on Analysis of Images, Social Networks and Texts* içinde (s. 29-46). Springer. http://www.machinelearning.ru/wiki/images/1/1f/voron14aist.pdf

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

Waltman, L. ve Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378-2392. https://www.doi.org/10.1002/asi.22748

Wang, Q., Cao, Z., Xu, J. ve Li, H. (2012). Group matrix factorization for scalable topic modeling. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (s. 375-384). https://doi.org/10.1145/2348283.2348335

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

Wang, Y., Wang, L., Li, Y., He, D. ve Liu, T. Y. (2013). A theoretical analysis of NDCG type ranking measures. *JMLR Workshop and Conference Proceedings,* vol. 2013 içinde (s. 1-30). PMLR. http://proceedings.mlr.press/v30/Wang13.pdf

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. (2017). Relevance theory and distributions of judgments in document retrieval. *Information Processing & Management*, *53*(5), 1080-1102. https://doi.org/10.1016/j.ipm.2017.02.010

White, H. D. (2018a). 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. (2018b). Pennants for Garfield: Bibliometrics and document retrieval. *Scientometrics,* 114, 757–778 (2018). https://doi.org/10.1007/s11192-017-2610-9.

White, H. D, Buzydlowski, J. ve Lin, X. (2000) Co-cited author maps as interfaces to digital libraries: Designing pathfinder networks in the humanities. *Proceedings of the*

International Conference on Information Visualization içinde (s. 25–30). Computer Society Press. https://doi.org/10.1109/IV.2000.859732

White, H. D. ve Griffith, B. C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-171. https://doi.org/10.1002/asi.4630320302

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.

Winograd, J. M. (1997). *Incremental refinement structures for approximate signal processing*. (Doktora Tezi, Boston Üniversitesi). https://www.proquest.com/dissertations-theses/incremental-refinement-structures-approximate/docview/304338225/se-2?accountid=142289

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

Zong, W. ve Huang, G-B. (2014). Learning to rank with extreme learning machine. *Neural Processing Letters, 39*(2), 155–166. https://doi.org/10.1007/s11063-013-9295-8

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

EK 1. ORİJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 13/06/2022

Tez Başlığı: Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve d) Sonuç kısımlarından oluşan toplam 76 sayfalık kısmına ilişkin, 13/06/2022 tarihinde şahsım/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda işaretlenmiş filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %2'dir.

Uygulanan filtrelemeler:

1- ☒ Kabul/Onay ve Bildirim sayfaları hariç
2- ☒ Kaynakça hariç
3- ☒ Alıntılar hariç
4- ☐ Alıntılar dâhil
5- ☒ 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

Adı Soyadı:	Müge Akbulut
Öğrenci No:	N15247015
Anabilim Dalı:	Bilgi ve Belge Yönetimi
Programı:	Bilgi ve Belge Yönetimi
Statüsü:	☒ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Prof. Dr. Yaşar Tonta

EK 2. MUAFİYET FORMU

HACETTEPE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
TEZ ÇALIŞMASI ETİK KOMİSYON MUAFİYETİ FORMU

HACETTEPE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih:13/06/2022

Tez Başlığı: Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi

Yukarıda başlığı gösterilen tez çalışmam:

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır,
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne müdahale içermemektedir.
4. Gözlemsel ve betimsel araştırma (anket, mülakat, ölçek/skala çalışmaları, dosya taramaları, veri kaynakları taraması, sistem-model geliştirme çalışmaları) niteliğinde değildir.

Hacettepe Üniversitesi Etik Kurullar ve Komisyonlarının Yönergelerini inceledim ve bunlara göre tez çalışmamın yürütülebilmesi için herhangi bir Etik Kurul/Komisyon'dan izin alınmasına gerek olmadığını; aksi durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

Adı Soyadı:	Müge Akbulut
Öğrenci No:	N15247015
Anabilim Dalı:	Bilgi ve Belge Yönetimi
Programı:	Bilgi ve Belge Yönetimi
Statüsü:	☐ Yüksek Lisans ☒ Doktora ☐ Bütünleşik Doktora

DANIŞMAN GÖRÜŞÜ VE ONAYI

Prof. Dr. Yaşar Tonta