Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

TÜRKÇE METİN TABANLI AÇIK ARŞİVLERDE KULLANILAN DİZİNLEME YÖNTEMİNİN DEĞERLENDİRİLMESİ

Çağdaş ÇAPKIN

Yüksek Lisans Tezi

Ankara, 2011

TÜRKÇE METİN TABANLI AÇIK ARŞİVLERDE KULLANILAN DİZİNLEME YÖNTEMİNİN DEĞERLENDİRİLMESİ

Çağdaş ÇAPKIN

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü

Bilgi ve Belge Yönetimi Anabilim Dalı

Yüksek Lisans Tezi

Ankara, 2011

KABUL VE ONAY

Çağdaş Çapkın tarafından hazırlanan "Türkçe Metin Tabanlı Açık Arşivlerde Kullanılan Dizinleme Yönteminin Değerlendirilmesi" başlıklı bu çalışma, 19.01.2011 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Prof. Dr. Yaşar A. Tonta (Başkan)

Prof. Dr. S. Serap Kurbanoğlu

Prof. Dr. Bülent Yılmaz

Doç. Dr. Nazan Özenç Uçak (Danışman)

Yrd. Doç. Dr. Mehmet Toplu

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. İrfan Çakın

Enstitü Müdürü

BİLDİRİM

Hazırladığım tezin/raporun tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin/raporumun kağıt ve elektronik kopyalarının Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü arşivlerinde aşağıda belirttiğim koşullarda saklanmasına izin verdiğimi onaylarım:

☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.

☐ Tezim/Raporum sadece Hacettepe Üniversitesi yerleşkelerinden erişime açılabilir.

■ Tezimin/Raporumun 1 yıl süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

19.01.2011

Çağdaş ÇAPKIN

Hatice ve Ahmet

ÇAPKIN'a...

TEŞEKKÜR

Danışmanlığımı üstlenip, sabırla çalışmamı destekleyen Doç. Dr. Nazan Özenç Uçak'a ne kadar teşekkür etsem azdır. Kendisi tezimle birlikte birçok farklı konuda beni desteklemiş, teknik ve bilimsel gelişimime büyük katkılar sağlamıştır.

Bu süreçte maddi ve manevi desteklerini esirgemeyen anneme, babama ve ağabeyime çok teşekkür ederim. OCR konusundaki yardımlarından ötürü Tolga ağabeyime ayrıca teşekkür ederim.

Türk Kütüphaneciliği açık arşivinin oluşturulmasında büyük desteğini gördüğüm Ali Fuat Kartal'a ve sponsorumuz EBSCO HOST'a; tez çalışmama yaptıkları katkılardan ötürü Güleda Düzyol ve Ece Dündar'a teşekkür ederim.

Ayrıca, bu süreçte manevi desteklerini esirgemeyen Dr. M. Tayfun Gülle ve Yrd. Doç. Dr. Mehmet Toplu'ya da teşekkürü borç bilirim.

ÖZET

ÇAPKIN, Çağdaş. Türkçe metin tabanlı açık arşivlerde kullanılan dizinleme yönteminin değerlendirilmesi, Yüksek Lisans Tezi. Ankara, 2011.

Bu araştırmanın amacı; açık arşivler için tasarlanabilecek bilgi erişim sistemlerinin performanslarının ve açık arşivlerde bilginin organizasyonu ve erişimini sağlayan standartların/protokollerin değerlendirilmesidir.

Bu amaçla, “Türk Kütüphaneciliği” dergisinde yer alan 2215 adet metin tabanlı doküman ile bir açık arşiv oluşturulmuş, oluşturulan açık arşivde bilgi erişim performanslarını değerlendirmek üzere Boole ve Vektör Uzayı modellerine dayalı üç farklı bilgi erişim sistemi tasarlanmıştır. Tasarlanan bilgi erişim sistemleri; sadece insana dayalı üretilmiş üstveri ile dizinlemenin yapıldığı “üstveri bilgi erişim sistemi” (ÜBES), sadece makineye dayalı (otomatik) dizinlemenin yapıldığı “tam-metin bilgi erişim sistemi” (TBES) ve hem insana hem de makineye dayalı dizinlemenin yapıldığı “karma bilgi erişim sistemi”dir (KBES).

Araştırmada betimleme yöntemi kullanılarak var olan durum betimlenmiş ve elde edilen bulgular literatüre dayalı olarak değerlendirmiştir. Araştırmada, bilgi erişim sistemlerinin performanslarını değerlendirmek amacıyla “anma-duyarlılık” ve “normalize sıralama” ölçümleri yapılmıştır.

Araştırmada aşağıdaki sonuçlara ulaşılmıştır.

Açık arşivler için tasarlanan bilgi erişim sistemlerinden KBES'in sergilediği duyarlılık performansının ÜBES'e ve TBES'e göre istatistiksel açıdan anlamlı bir fark yarattığı saptanmıştır. Ayrıca, her bir bilgi erişim sisteminde anma ve duyarlılık arasında güçlü bir negatif ilişki saptanmış, anma arttıkça duyarlılık düşmüştür.

ÜBES'in ve KBES'in sergiledikleri normalize sıralama performansının TBES'e göre istatistiksel açıdan anlamlı bir fark yarattığı saptanmıştır. Normalize

sıralama performansında ÜBES ile KBES arasında ise istatistiksel açıdan anlamlı bir fark saptanmamıştır.

ÜBES, hem en az ilgili hem de en az ilgisiz dokümana erişilen; TBES, en fazla ilgisiz ve ikinci sırada en fazla ilgili dokümana erişilen; KBES ise, en fazla ilgili ve ikinci sırada ilgisiz dokümana erişilen bilgi erişim sistemi olmuştur.

Açık arşivlerde OAI-PMH ile OAI-ORE protokollerinin birlikte kullanımının sadece OAI-PMH protokolü kullanımına göre açık arşivlerin amacına daha uygun olduğu sonucuna varılmıştır.

Anahtar Sözcükler

Açık erişim, açık arşivler, üstveri, bilgi erişim, otomatik dizinleme, Apache Lucene.

ABSTRACT

ÇAPKIN, Çağdaş. Evaluation of indexing method used in Turkish text-based open archives, Master's Thesis. Ankara, 2011.

The purpose of this research is to evaluate performance of information retrieval systems designed for open archives, and standards/protocols enabling retrieving and organizing information in open archives.

In this regard, an open archive was developed with 2215 text-based documents from "Turkish Librarianship" journal and three different information retrieval systems based on Boolean and Vector Space models were designed in order to evaluate information retrieval performances in the open archive developed. The designed information retrieval systems are: "metadata information retrieval system" (ÜBES) involving indexing with metadata created based only on human, "full-text information retrieval system" (TBES) involving (automatic) indexing based on only machine, and "mixed information retrieval system" (KBES) involving indexing based both on human and machine.

Descriptive research method is used to describe the current situation and findings are evaluated based on literature. In order to evaluate performances of information retrieval systems, "precision and recall" and "normalized recall" measurements are made.

The following results are found at the end of the research:

It is determined that the precision performance of KBES information retrieval system designed for open archives creates statistically significant difference in comparison to ÜBES and TBES. In each information retrieval system, a strong negative correlation is identified between recall and precision, where precision decreases as recall increases.

It is determined that the "normalized recall" performance of ÜBES and KBES create statistically significant difference in comparison to TBES. In "normalized

recall” performance, no statistically significant difference is identified between ÜBES and KBES.

ÜBES is the information retrieval system through which minimum number of relevant and nonrelevant documents; TBES, through which maximum number of nonrelevant and second most relevant documents, and KBES, through which maximum number of relevant and second most nonrelevant documents are retrieved.

It is concluded that using OAI-PMH and OAI-ORE protocols together rather than using only OAI-PMH protocol fits the purpose of open archives.

Key Words

Open access, open archives, metadata, information retrieval, automatic indexing, Apache Lucene.

İÇİNDEKİLER

KABUL VE ONAY i
BİLDİRİM ii
TEŞEKKÜR iv
ÖZET v
ABSTRACT vii
İÇİNDEKİLER ix
KISALTMALAR DİZİNİ xi
TABLOLAR DİZİNİ xiv
ŞEKİLLER DİZİNİ xv
1 . BÖLÜM 1
1.1 . KONUNUN ÖNEMİ 1
1.2 . ARAŞTIRMANIN AMACI VE HİPOTEZİ 3
1.3 . ARAŞTIMANIN KAPSAMI VE YÖNTEMİ 4
1.4 . ARAŞTIRMANIN DÜZENİ 5
1.5 . KAYNAKLAR 6
2 . BÖLÜM: BİLGİ ERİŞİM 7
2.1 . BİLGİ 7
2.2 . BİLGİ ERİŞİM 9
2.3 . BİLGİ ERİŞİM MODELLERİ 17
2.3.1 . Boole Bilgi Erişim Modeli 17
2.3.2 . Vektör Uzayı (Vector Space) Bilgi Erişim Modeli 22
2.3.3 . Genişletilmiş (Extended) Boole Modeli 28
2.4 . PERFORMANS DEĞERLENDİRME 31
3 . BÖLÜM: APACHE LUCENE 34
3.1 . GİRİŞ 34
3.2 . ANALİZCİLER 39
3.3 . DOKÜMAN VE ALAN 42
3.4 . DİZİN 46
3.5 . ARAMA BİLEŞENLERİ 49
3.6 . BENZERLİK 54

4 . BÖLÜM: **AÇIK ARŞİVLER** 56
4.1 . GİRİŞ 56
4.2 . AÇIK ERİŞİM 60
4.3 . AÇIK ARŞİVLER GİRİŞİMİ VE BİRLİKTE İŞLERLİK 66
4.4 . AÇIK ARŞİV YAZILIMLARI 74
5 . BÖLÜM: BULGULAR 77
5.1 . GİRİŞ 78
5.2 . HAZIRLIK AŞAMASI 78
5.2.1 . Dermenin Elde Edilişi 78
5.2.2 . Metin Analizi 79
5.2.3 . Tasarlanan Bilgi Erişim Sistemlerinin Özellikleri 81
5.3 . TÜRK KÜTÜPHANECİLİĞİ AÇIK ARŞİVİ 82
5.4 . TEST DERMESİ 83
5.5 . TEST SORULARININ SEÇİMİ, FORMÜLASYONU, İLGİLİLİK VE NORMALİZE SIRALAMA BULGULARI 84
5.5.1 . 1. Soru: İrfan Çakın'ın Yazdığı Tüm Dokümanlar 89
5.5.2 . 2. Soru: İrfan Çakın'a Bilimsel/Hakemli Dokümanlarda Yapılan Atıflar 90
5.5.3 . 3. Soru: “Bilgi Arama Davranışı” 91
5.5.4 . 4. Soru: AACR, AACR1, AACR2 92
5.5.5 . 5. Soru: OPAC, “Çevrimiçi Katalog” 94
5.5.6 . 6. ve 7. Soru: [Engelliler, Özürlüler] ve [Engelli, Özürlü] 95
5.5.7 . 8. ve 9. Soru: [“Kullanıcılara Eğitimler”, “Okuyuculara Eğitimler”, Oryantasyonlar] ve [“Kullanıcı Eğitimi”, “Okuyucu Eğitimi”, Oryantasyon] 97
5.6 . PERFORMANS DEĞERLENDİRME SONUÇLARI 99
6 . BÖLÜM: SONUÇ VE ÖNERİLER 105
KAYNAKÇA 112

KISALTMALAR DİZİNİ

AACR : Anglo-American Cataloguing Rules

ANKOS : Anadolu Üniversiteleri Kütüphaneleri Konsorsiyumu

ANZLIS : Australia New Zealand Information Council

API : Application Programming Interface

ARL : Association of Research Libraries

BBEM : Boole Bilgi Erişim Modeli

BOAI : Budapest Open Access Initiative

CC : Creative Commans

CCSDS : Consultative Committee for Space Data Systems

CIMI : Consortium for the Computer Interchange of Museum Information

BSD : Berkeley Software Distribution

DC : Dublin Core

DOAJ : Directory of Open Access Journals

EAD : Encoded Archival Description

EDNA : Education Network Australia

GBBEM : Genişletilmiş (Extended) Boole Bilgi Erişim Modeli

GILS : Government Information Locator Service

GNU : General Public License

HTML : HyperText Markup Language

IFLA : The International Federation of Library Associations and Institutions

ISBD : International Standard Bibliographic Description

ISO : International Organization for Standardization

İVTYS : İlişkisel Veritabanı Yönetim Sistemleri

JDK : Java Development Kit

KBES : Karma Bilgi Erişim Sistemi

MARC : Machine-Readable Cataloging

NCSA : National Center for Supercomputing Applications

NISO : National Information Standards Organization

OAI : Open Archives Initiative

OAI-ORE : Open Archives Initiative Protocol - Object Exchange and Reuse

OAI-PMH : Open Archives Initiative - Protocol for Metadata Harvesting

OAIS : Reference Model for an Open Archival Information System

OCLC : Online Computer Library Center

OJS : Open Journal Systems

OPAC : Online Public Access Catalog

ORC : Optical Character Recognition

PDF : Portable Document Format

PLoS : Public Library of Science

PMC : PubMed Central

RAM : Random Access Memory

RDF : Resource Description Framework

ROAR : Registry of Open Access Repositories

RSS : Really Simple Syndication

SPARC : Scholarly Publishing and Academic Resources Coalition

SPSS : Statistical Package for the Social Sciences

SSS : Sweet Spot Similarity

TBES : Tam-metin Bilgi Erişim Sistemi

TEI : Text Encoding Initiative

TK : Türk Kütüphaneciliği

TKD : Türk Kütüphaneciler Derneği

TKDB : Türk Kütüphaneciler Derneği Bülteni

TO-KAT : Ulusal Toplu Katalog

TREC : Text REtrieval Conference

TSE : Türk Standardlar Enstitüsü

ÜBES : Üstveri Bilgi Erişim Sistemi

VRA : Visual Resource Association

VUBEM : Vektör Uzayı Bilgi Erişim Modeli

WWW : World Wide Web

XML : Extensible Markup Language

TABLOLAR DİZİNİ

Tablo 1. USMARC'ın Temel Alanları 11

Tablo 2. USMARC'ın Bazı Alt Alanları 11

Tablo 3. DC Setinde Yer Alan Elementler ve Açıklamaları 14

Tablo 4. Veri Erişim ve Bilgi Erişim Özellikleri 15

Tablo 5. Idf Parametresi Örneği 26

Tablo 6. Apache Tika'nın Metin veya Üstveri Çıkartabildiği Dosya Türleri ve Kullandığı API'lar 38

Tablo 7. Sık Kullanılan Tokenizer Türlerinin Özellikleri 41

Tablo 8. Sık Kullanılan TokenFilter Türlerinin Özellikleri 41

Tablo 9. Analizci Türlerinin Özellikleri 42

Tablo 10. Dizinlemeye Yönelik Olarak Alanın Sahip Olduğu Özellikler 45

Tablo 11. Depolamaya Yönelik Olarak Alanın Sahip Olduğu Özellikler 45

Tablo 12. Terim Vektörlerine Yönelik Olarak Alanın Sahip Olduğu Özellikler 46

Tablo 13. Lucene Dizininde Kullanılan Dosyalar, Dosya Uzantıları ve Açıklamaları 48

Tablo 14. Lucene API'ın Arama Konusundaki Temel Sınıfları 51

Tablo 15. Temel CC Koşulları 64

Tablo 16. CC Lisans Türleri ve Açıklamaları 64

Tablo 17. Alanlarda Yer Alan Karakter Uzunlukları 83

Tablo 18. Dermede En Sık Geçen 25 Terim 83

Tablo 19. Sorgularla İlgili Tüm Dokümanların Sayısı, Bilgi Erişim Sistemlerinin Erişebildikleri İlgili ve İlgisiz Doküman Sayısı 87

Tablo 20. TBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı 88

Tablo 21. ÜBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı 88

Tablo 22. KBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı 88

Tablo 23. 11 Adet Anma Basamağında Bilgi Erişim Sistemlerinin Ortalama Duyarlılık Değerleri 101

ŞEKİLLER DİZİNİ

Şekil 1. MARC'ın Yapısal Elementleri 11

Şekil 2. Bilgi Erişim Süreci 16

Şekil 3. Boole İşleçlerinin Doğruluk Tabloları 18

Şekil 4. Küme İşlemlerinin Venn Şeması (“birleşim”, “kesişim” ve “fark”) 18

Şekil 5. İkili Devrik Dizin Örneği 20

Şekil 6. İkili Devrik Dizinin Venn Şeması 21

Şekil 7. VUBEM’de Sorgu ve Doküman Vektörünün Gösterimi 24

Şekil 8. GBBEM’de AND İçin (1,1) Noktasından ve OR İçin (0,0) Noktasından Eşit Uzaklık Çizgileri 30

Şekil 9. Duyarlılık ve Anmanın Görselleştirilmesi 31

Şekil 10. Lucene ile MySQL ve PostgreSQL'in Dizinleme ve Arama Sürelerinin Karşılaştırılması 36

Şekil 11. Lucene ile MySQL ve PostgreSQL'in Bilgi Erişim Performansının Karşılaştırılması 37

Şekil 12. Lucene'in Uygulamalara Entegrasyonu 39

Şekil 13. Terimlerin Pozisyon ve Ofset Üstverileri 40

Şekil 14. Latin Harfleri ile Yazılmış Metinlerde Kullanılabilecek TokenStream Mimarisi 40

Şekil 15. Lucene'in Dizin, Doküman ve Alan İlişkisi 43

Şekil 16. 3 Segmentli Optimize Edilmemiş Lucene Dizini 47

Şekil 17. QueryParser Sınıfının İşleyişi 52

Şekil 18. Wikipedia'nın Kütüphane Terimini Vurgulama Örneği 53

Şekil 19. 1986-2008 Yılları Arasında ARL Üyesi Kütüphanelerde Harcama Trendleri 59

Şekil 20. Örnek Bir OAI-PHM Kaydı 69

Şekil 21. OAI- ORE Örneği: Kaynak Haritalama KH-1 ile Tanımlanmış Toplanma T-1'in Üç Kaynağı Toplaması 72

Şekil 22. OAI-ORE ATOM Uygulama Örneği 73

Şekil 23. “oai_rem_atom” Fili ile AOI-PHM'den OAI-ORE'ye Bağ Kurma Örneği 74

Şekil 24. Bilgi Erişim Sistemlerinin Ortalama Anma ve Duyarlılık Grafiği 100

Şekil 25. 9 Soruya Karşılık Bilgi Erişim Sistemlerinin Erişebildikleri Doküman Sayıları 102

Şekil 26. Bilgi Erişim Sistemlerinin Her Bir Sorudaki R_{norm} Değerleri 102

1. BÖLÜM

1.1. KONUNUN ÖNEMİ

Bilgi teknolojileri, 1990'lı yıllardan sonra hızlı bir gelişim geçirerek birçok alanda yaygın olarak kullanılmaya başlamıştır. Bu alanların başında yayıncılık gelmektedir. Bilgi teknolojilerinin yayıncılık alanında uygulanmasıyla ortaya çıkan elektronik yayıncılık, hem üreticilere/yayıncılara hem de tüketicilere/kullanıcılara önemli avantajlar sağladığı için ön plana çıkmıştır. Bu durum, kütüphanelerin de evrimleşme sürecine girmesine neden olmuştur. Tonta (2007), bu süreci "kütüphanelerin sanal güzergâhlara dönüşmesi" olarak ifade etmiştir. Nitekim, Amerika Birleşik Devletleri'nde faaliyet gösteren Araştırma Kütüphaneleri Derneği'nin (ARL) 2009 yılında yayınladığı rapor incelendiğinde, 1980'li yıllardan başlamak üzere kütüphane bütçelerinden en büyük payı artan bir biçimde elektronik kaynakların aldığı görülmektedir.

ARL'nin 2009 yılında yayınladığı raporda dikkat çeken bir başka nokta da dergi harcamalarının 1986 yılından 2008 yılına kadar %374 artmasıdır. Bu artışın en önemli nedeni yayıncıların "keyfi" pazarlama politikalarıdır. Keyfi pazarlama politikalarından olumsuz etkilenen kütüphaneler ve kütüphane kullanıcıları farklı çözüm yolları üretmeye çalışmıştır. Çözüm yolları arasından en etkili olanı ise "açık erişim" modeli olmuştur.

Açık erişim (open access), "bilimsel literatürün İnternet aracılığıyla finansal, yasal ve teknik bariyerler olmaksızın, erişilebilir, okunabilir, kaydedilebilir, kopyalanabilir, yazdırılabilir, taranabilir, tam metne bağlantı verilebilir, yazılıma veri olarak aktarılabilir ve her türlü yasal amaç için kullanılabilir biçimde kamuya ücretsiz açık olması" biçiminde tanımlanmaktadır (BOAI, 2002; ANKOS, 2010a).

Açık erişim modeli dünya genelinde olduğu gibi, Türkiye'de de yankı bulmuş ve bu alanda çeşitli çalışmalar yapılmıştır. 2006 yılında Anadolu Üniversiteleri Kütüphaneleri Konsorsiyumu (ANKOS) altında "Açık Erişim ve Kurumsal Arşivler Çalışma Grubu" oluşturulmuş (ANKOS, 2010b), çeşitli uygulama

projeleri geliştirilmiş (Polat, 2006; Tonta, Küçük, Al, Alır, Ertürk, Olcay ve diğerleri, 2006; Atılgan ve Bulut, 2008), doktora tezleri sunulmuş (Ertürk, 2008; Afzali, 2009), çeşitli makaleler ve bildiriler yayınlanmıştır (Kayaoğlu, 2006; Tonta, 2006; Tonta, Ünal ve Al, 2007; Ertürk ve Küçük, 2010; Karasözen, Zan ve Atılgan, 2010; Tonta, 2010a).

Ayrıca, Küçük, Al ve Olcay'ın (2008) Türkiye'deki bilimsel elektronik dergiler üzerine yaptığı bir çalışma, dergilerin % 95'inin açık erişime olanak tanıdığını ortaya koymaktadır. Aynı araştırmada elektronik yayıncılığın Türkiye için yeni olmasının sonucu olarak teknolojinin, işlemlerin ve işlerin tamamının sürece entegre edilemediğine de dikkat çekilmektedir. Mevcut durum değerlendirildiğinde, Türkiye'nin açık erişim modelinin desteklenmesi konusunda büyük bir potansiyele sahip olduğu sonucuna varılabilmektedir.

Açık erişimin bir başka önemli konusu; açık arşivlerde yer alan bilgi kaynaklarının nasıl organize edileceği, bu kaynakların nasıl eriştirileceği ve birlikte işlerliğin nasıl sağlanacağıdır. Bu konular üzerinde standartların ve protokollerin geliştirilmesi amacıyla Açık Arşivler Girişimi (OAI) kurulmuş ve OAI, 2001 yılında OAI-PMH ve 2008 yılının sonunda OAI-PMH'nin eksiklerini gidermek amacıyla OAI-ORE standartlarını/protokollerini geliştirmiştir.

Açık erişim konusu “erişim” bağlamında ele alındığında “bilgi erişim” konusu ön plana çıkmaktadır. Ancak, literatür incelendiğinde, “açık erişim” konusunun “bilgi erişim” açısından ayrıntılı olarak ele alınıp değerlendirildiğini söyleyebilmek mümkün değildir. Literatürde daha çok konunun önemi bağlamında; yayınların üretim süreci ile yayıncıların pazarlama politikaları arasındaki çarpıklıklar, açık erişimin atıf konusundaki etkisi ve yayıncılık konusunda geliştirilmiş yazılımlar gibi konular ele alınmaktadır. Bu araştırmada açık erişim konusu, OAI'nın geliştirdiği standartlar ve protokoller çerçevesinde, “bilgi erişim” bağlamında ele alınıp değerlendirilmektedir.

OAI standartları veya protokolleri ile uyumlu açık arşivler ele alındığında, farklı dizinleme teknikleriyle bilgi erişim sistemleri geliştirebilmek mümkündür. Bu dizinleme tekniklerinden biri insana dayalı, diğeri de makineye dayalıdır. Söz

konusu dizinleme teknikleri birbirleriyle karşılaştırıldığında, her iki tekniğin de birbirine karşı sağladığı çeşitli üstünlükler vardır. Makineye dayalı teknik; maliyet, zaman ve kapsam bakımından insana dayalı teknikten, insana dayalı teknik de kelime/terim yönetimi ve tanımlayıcı yapılandırma bakımından makineye dayalı teknikten üstündür (Shields, 2005). Buna ek olarak, her iki dizinleme tekniğinin avantajlı yönleri değerlendirilerek karma dizinleme yapabilmek, dolayısıyla, OAI standartları veya protokolleri ele alındığında üç farklı bilgi erişim sistemi tasarlayabilmek mümkündür. Bunlardan birincisi, sadece insana dayalı olarak üretilmiş üstverinin kullanıldığı bilgi erişim sistemi (ÜBES); ikincisi, sadece makineye dayalı olarak tam-metin ile otomatik dizinlemenin yapıldığı bilgi erişim sistemi (TBES) ve üçüncü, hem üstveriye hem de tam-metne dayalı olarak geliştirilmiş karma bilgi erişim sistemidir (KBES). Bu doğrultuda, söz konusu bilgi erişim sistemlerinin erişim performanslarının değerlendirilmesi ve en uygun bilgi erişim sisteminin saptanması gerekmektedir.

1.2. ARAŞTIRMANIN AMACI VE HİPOTEZİ

Bu araştırmanın amacı, açık arşivler için geliştirilebilecek olan ÜBES, TBES ve KBES'in performanslarını ve açık erişim standartlarını veya protokollerini değerlendirmektir. Bu amaçla; söz konusu bilgi erişim sistemleri tasarlanıp, her bir bilgi erişim sisteminin erişim performansını etkileyen faktörler ortaya konulmaya çalışılmıştır. Araştırma kapsamında erişim performansını etkileyen faktörler arasında, bilgi erişim sistemlerinin kelime/terim yönetimi, doküman uzunluğu, dizinleme tekniğine bağlı olarak dizin yapılandırılışı ve gövdeleme algoritması bulunmaktadır. Ayrıca, araştırma sonucunda elde edilen bulgular değerlendirilerek, uygulama geliştiricilerinin gelecekte tasarlayabilecekleri bilgi erişim sistemlerine yardımcı olacak önerilerin geliştirilmesine de çalışılmıştır.

Araştırmada, üç bilgi erişim sisteminin duyarlılık ve normalize erişim performansları arasında anlamlı bir fark var mıdır? Anlamlı bir fark varsa, farkı yaratan faktör veya faktörler nelerdir? sorularına yanıt aranmaktadır. Bu araştırma sorularından yola çıkarak, araştırma hipotezleri şu biçimde belirlenmiştir:

- KBES'in duyarlılık performansı ÜBES'ten ve TİBES'ten yüksektir.
- TİBES'in normalize sıralama performansı ÜBES'ten ve KBES'ten düşüktür.
- Üç bilgi erişim sisteminin **seçilen sorulara karşı eriştikleri doküman sayısı birbirinden farklıdır.**

1.3. ARAŞTIRMANIN KAPSAMI VE YÖNTEMİ

Araştırma kapsamında kullanılmak üzere ağırlıklı olarak Türkçe dokümanlardan oluşan bir açık arşiv arayışına girilmiştir. Ancak, çalışmanın başladığı dönemde tasarlanması planlanan bilgi erişim sistemlerinin performanslarının değerlendirilebilmesini sağlayabilecek sayıda Türkçe dokümana sahip herhangi bir açık arşiv bulunamamıştır. Ayrıca, ağırlıklı olarak Türkçe dokümanlara sahip ve OAI standartlarından en az biriyle uyumlu olan açık arşivlerde yer alan dokümanların metin tabanlı olmaması (imaj tabanlı olması) sebebiyle de var olan açık arşivlerden faydalanılamamıştır. Bu doğrultuda, çalışmanın kapsamını belirleyecek ve çalışmanın yapılmasını sağlayabilecek bir açık arşivin geliştirilmesi gerekmiştir. Bu gerekçelerle, *Tük Kütüphaneciliği* dergisinde 1952 yılından 2010 yılının ikinci sayına kadar yayınlanmış 2215 adet dokümandan/makaleden oluşan *Türk Kütüphaneciliği* açık arşivi oluşturulup, çalışma söz konusu açık arşivle sınırlandırılmıştır.

Araştırmada betimleme yöntemi kullanılmıştır. Kaptan (1991, s. 59), betimleme yöntemini şu şekilde tanımlamaktadır:

> ..olayların, varlıkların, kurumların, grupların ve çeşitli alanların 'ne' olduğunu betimlemeye, açıklamaya çalışır. Betimleme araştırmaları, mevcut olayların daha önceki olay ve koşullarla ilişkilerini de dikkate alarak durumlar arasındaki etkileşimi açıklamayı hedef alır.

Araştırmada veri toplamak amacıyla, performans değerlendirmesi yapılmıştır. Performans değerlendirmesinde, bilgi erişim sistemlerinin belirlenen özelliklerini

değerlendirmek amacıyla çeşitli sorular seçilip, formüle edilerek bilgi erişim sistemleri sorgulanmaktadır. Sorgu sonucu erişilen dokümanlar ilgili veya ilgisiz olarak değerlendirilmektedir. İlgililik değerlendirmesiyle elde edilen veriler, tek başına kullanılabileceği gibi, performans değerlendirme tekniklerinde kullanılarak da anlamlı verilere dönüştürülebilmektedir.

Araştırmada, bilgi erişim performansını değerlendirmek amacıyla "duyarlılık-anma" ve "normalize sıralama" teknikleri kullanılmıştır.

1.4. ARAŞTIRMANIN DÜZENİ

Bu araştırma 6 bölümden oluşmaktadır.

Birinci bölümde; konunun önemi, araştırmanın; amacı, kapsamı, hipotezleri, yöntemi, veri toplama teknikleri, düzeni ve araştırmada kullanılan kaynaklar;

İkinci bölümde; bilgi, bilgi erişim, bilgi erişim modelleri ve performans değerlendirme teknikleri;

Üçüncü bölümde; bilgi erişim sistemlerinin tasarımında kullanılmış olan Apache Lucene;

Dördüncü bölümde; açık erişim, açık arşivler, açık arşivler girişimi ve birlikte işlerlik, açık arşiv yazılımları ve Türk Kütüphaneciliği açık arşivi;

Beşinci bölümde; bilgi erişim sistemlerinin tasarımı, üstverinin ve tam-metnin elde edilişi, test dermesi, test sorularının; seçimi, formülasyonu, ilgililiği ve bilgi erişim sistemlerinin söz konusu sorulara karşı davranışı ve performans değerlendirme sonuçları;

Altıncı bölümde ise, sonuç ve öneriler yer almaktadır.

1.5. KAYNAKLAR

Araştırmanın arka planını oluşturan bilgi ihtiyacını karşılamak üzere aşağıda yer alan dijital kütüphaneler ve bilgi erişim araçları kullanılarak ayrıntılı bir literatür taraması yapılmıştır.

- Academic Search Premier
- Blackwell Synergy
- CiteSeer
- DOAJ
- EBSCOHost Research Databases
- Emerald Library
- E-prints in Library and Information Science (E-LIS)
- Google Scholar
- IEEE Xplore
- Library, Information Science & Technology Abstracts
- OAISTER
- ProQuest Digital Dissertations
- ScienceDirect
- Science Citation Index Expanded
- Scopus
- Social Sciences Citation Index
- Springer LINK-Kluwer
- Türkiye Makaleler Bibliyografyası
- ULAKBİM Ulusal Veri Tabanları (UVT)
- Wiley Interscience

Araştırma raporunun yazımında *Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez ve Rapor Yazım Yönergesi* (2004), *Bilimsel Yayınlarda Kaynak Gösterme İlkeleri* (2006) ve *Kaynak Gösterme El Kitabı* (Kurbanoğlu, 2004) kullanılmıştır.

2 . BÖLÜM

BİLGİ ERİŞİM

2.1 . BİLGİ

Bilgi (information), insanın merak duygusunu giderme, problem çözme ve karar verme gibi çeşitli kritik süreçlerinin odak noktasında bulunan ve algılanması güç olan disiplinlerarası bir kavramdır. Birçok disiplin bu kavramın çevresinde çalışmaktadır. Bilgibilim, kütüphanecilik, bilgisayar bilimi, matematik ve felsefe bilgi üzerine önemli çalışmalar yürüten disiplinlerden birkaçıdır.

Konunun kavramsal boyutları incelendiğinde her alanı kapsayacak şekilde kabul gören ortak bir tanımın olmadığı dikkat çekmektedir. Bilgi ile ilgili disiplinlerin bilgi terimini kendilerine göre tanımladıklarına dikkat çeken Uçak (2000; 2010), günümüzde tüm disiplinleri içine alabilecek tek bir bilgi tanımı yapılmasının olanaksız ve hatta yersiz olduğunu vurgulamaktadır. Bir disiplinin kendine göre tanımladığı “information” terimi, zaman zaman diğer disiplinlerin kendilerine göre tanımladıkları “data” ve/veya “knowledge” terimlerinin anlamlarıyla örtüşebilmektedir. Özellikle Türkçede, “knowledge” ve “information” terimlerini temsil edebilecek farklı terimlerin literatüre oturtulamamış olması, her iki terimin Türkçe karşılığı olarak “bilgi” teriminin kullanılmasına yol açarak karmaşayı artırmaktadır. Uluslararası literatürde ise, bilgi (information) terimi veri (data) ve knowledge terimlerine ek olarak; haber (news), irfan (wisdom), istihbarat (intelligence) ve anlam (meaning) terimleri yerine de kullanılabilmektedir (Meadow, Boyce ve Kraft, 2007, s. 37).

Bilgi erişim (information retrieval) literatürü incelendiğinde ise, bilgi teriminin tanımı üzerinde yeterince durulmadığı; tanımların ise, zamanla değişim gösterdiği ve birbirinden farklı özellikleri ön plana çıkardığı görülmektedir. Bilginin ne olduğu konusunda kısıtlı sayıda olan ifadelerin de birbirleriyle çeliştiği ve tatmin etme bakımından çok düşük düzeyde olduğu görülmektedir.

Meadow, Boyce ve Kraft (2007, s. 38), “bilgi teorisi” kurucularından Shannon’un

entropi formülünü[1] bilgi (information) teriminin resmi/formal tanımı olarak nitelendirmektedir. Öte yandan, bilgi erişim ile ilgili diğer bir kaynak (Baeza-Yates ve Ribeiro-Neto, 1999, s. 145) Shannon'un entropy formülünü bir metinde geçen bilgi miktarının taşıdığı sembollerin dağılımıyla yakalanması olarak ele almaktadır. Rijsbergen (1979), bilgi (information) teriminin doküman (document) terimi yerine kullanıldığını vurgularken; bilgi erişim literatürüyle yakından ilişkili olan "veri erişim" (data retrieval) veya "veri tabanı yönetim sistemleri" (database management systems) litaretüründe (Rainardi, 2008, s. 24; Rob ve Coronel, 2009, s. 45; Townsend, Riz ve Schaffer, 2004, s. 337) "bilgi" veya "doküman" yerine "yapılandırılmamış veri" (unstructured data) ifadesinin sıkça kullanılmış olduğu dikkat çekmektedir. Ayrıca, "bilgi" yerine "yapılandırılmamış veri" ifadesinin kullanımına bilgi erişim literatüründe de rastlamak mümkündür (bkz: Grossman ve Frieder, 2004, s. 212). Bunlara ek olarak, bilgi erişim konusunda gerçekleştirilen önemli faaliyetlerden biri olan Metin Erişim Konferansı'nın (Text REtrieval Conference -TREC[2]), adı ve amacı göz önünde bulundurulduğunda "bilgi" yerine "metin" (text) teriminin kullanıldığını da öne sürmek mümkündür.

Kowalski ve Maybury (1998, s. 2) ise, bilginin "oluşturulmuş metin (sayılar ve tarihler dahil), imaj, ses, video ve diğer çoklu-ortam nesneleri" olabileceğini öne sürmüştür. Kowalski ve Maybury'nin bilgiye yaklaşımında dikkat çeken nokta bilginin sunuluş biçimi veya bilgi taşıyan kaynakların farklı formatta olduğu, ancak formatlar farklı olsa da iletilmek istenenin "bilgi" olduğudur. Bu yaklaşım, bilgi erişim literatüründe bilgi yerine farklı terimlerin kullanım nedenine ışık tutması bakımından önemlidir. Çünkü, bilgi erişim alanı Kowalski ve Maybury'nin belirttiği bilgi ortamlarına özgü erişim sistemleri geliştirmektedir. Bu bağlamda, bilgi yerine "doküman" veya "metin" terimlerinin kullanılmasındaki amaç hangi formattaki bilginin ele alındığıdır. Buradan yola çıkarak, veri erişim literatürünün "yapılandırılmamış veri", bilgi erişim literatürünün "doküman" veya "metin" olarak ifade ettiği olgunun "bilgi" olduğu kabul edilerek, bu çalışmada "information" teriminin Türkçe karşılığı olarak "bilgi" kullanılmıştır.

---

1 $E = -\sum_{i=1}^{\sigma} p_i \log_2 p_i$ Kaynak: (Shannon, 1959).

2 TREC konferanslarının çevrimiçi adresi: http://trec.nist.gov/overview.html

Bilgi erişim literatürü, birbiri ile içli dışlı olan ve hangisinin hangisini kapsadığı bakış açısına göre değişebilen “data”, “information” ve “knowledge” terimlerinden sadece “data” ve “information” ayrımına erişim sistemlerinin özellikleri bakımından girmektedir.

2.2 . BİLGİ ERİŞİM

Yazılı/basılı veya elektronik ortamda yer alan bilgi kaynaklarının erişilebilir olabilmesi için organize edilmesi gerekmektedir. Bilgi erişim konusu, bilginin organizasyonu bağlamında ele alındığında kütüphanecilik ve veri erişim alanlarıyla yakından ilgilidir. Kütüphanecilik ve veri erişim alanları bilginin organizasyonunu gerçekleştirmede insanı merkeze alırken, bilgi erişim alanı, insanla birlikte ağırlıklı olarak bilginin organizasyonunu gerçekleştirmede makineyi (bilgisayarları) merkeze koymaktadır.

İnsan ve makine merkezli yaklaşımların ortak hedefi bilgiyi organize etmek olsa da, her iki yaklaşımın kullandığı teknikler farklıdır. Ayrıca, her iki yaklaşımın da birbirine karşı sağladığı çeşitli üstünlükler vardır. Makineye dayalı organizasyon; maliyet, zaman ve kapsam bakımından insana dayalı bilgi organizasyonundan, insana dayalı organizasyon da kelime yönetimi ve tanımlayıcı yapılandırma bakımından makineye dayalı organizasyondan üstündür (Shields, 2005). Çalışma kapsamında her iki yaklaşımdan da faydalanılacağı için, bilgi erişim konusu insana ve makineye dayalı bilgi organizasyonu olarak ele alınmaktadır.

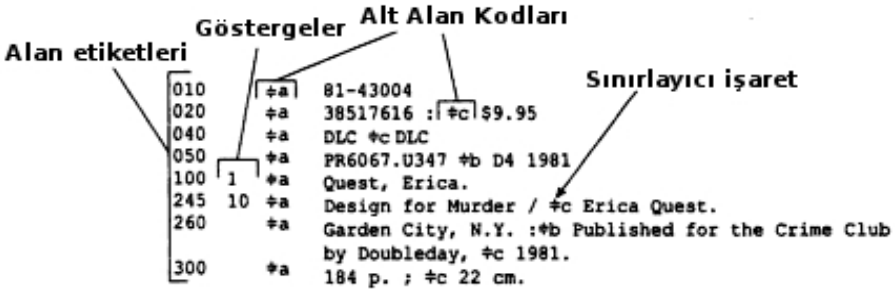
İnsana dayalı bilgi organizasyonunda, bilgi kaynaklarına erişimi sağlamak amacıyla bilgi kaynaklarının tanımlayıcı verileri (eser adı, yazar adı, konu, vd.) insan tarafından çıkartılmaktadır. Tanımlayıcı verilerin çıkartılma sürecine kütüphanecilik alanı “kataloglama” adını vermektedir. Kütüphane kataloglarının temel işlevleri; belirli bir yazarın belirli kitaplarının kütüphanede olup olmadığı, kütüphanenin belirli bir konuda hangi kitaplara sahip olduğu, var olan kitapların hangi rafların neresinde yer aldığı sorularına cevap vermektir (Koch, 1913, s. 33). Kütüphanecilik alanı bu soruları cevaplamak için çeşitli kataloglama ve sınıflama araçları geliştirerek standartlaştırmış ve bu araçlarla oluşturulan

dizinler veya kart kataloglar ilk bilgi erişim sistemleri olarak kabul edilmiştir. Kataloglardan erişim ise yazar (kişi, tüzel kuruluş ve eser adı) ile olduğu gibi denetimli dillerin (konu başlığı listeleri) kullanımıyla konu erişimi de sağlanabilmiştir (Baydur, 2010).

Bilgisayarların kütüphanelerde kullanılmasıyla, kart katalogların yerine makinenin anlayabileceği kütüphane kayıtları oluşturmak ve oluşturulan kayıtları paylaşmak/dağıtmak amacıyla da MARC (Machine-Readable Cataloging) standardı geliştirilmiştir. 1965 yılında MARC I, 1967 yılında ise MARC II geliştirilmiş ve ulusal ihtiyaçlara yönelik olarak USMARC (Z39.2), UKMARC ve CANMARC gibi çeşitli MARC formatları geliştirilmiştir (Byrne, 1998, s. 6-7, 11). MARC kayıtlarının farklı formatlarda yaratılması ise birlikte işlerlik konusunda sorun yaratmıştır. Sorunun çözümüne ilişkin 1967 yılında IFLA (International Federation of Library Associations) UNIMARC'ı (Universal MARC) geliştirmiştir. UNIMARC'ın amacı, ulusal standartlara göre oluşturulmuş MARC kayıtlarının uluslararası standartta bir kayda dönüştürülmesini veya uluslararası standartla üretilmiş olan MARC kayıtlarının ulusal formatlara dönüştürülmesini sağlamaktır (Wedgeworth, 1993, s. 541).

Kütüphane kayıtlarının paylaşımı/dağıtımı için MARC'ın standartlaştırılması süreci ise MARC II ile başlamış ve 1968 yılında Kongre Kütüphanesi (Library of Congress) yaklaşık 50.000 adet MARC formatındaki kütüphane kaydını dağıtmıştır/paylaşmıştır (Khan, 1997, s. 115). Daha sonra Internet'in ortaya çıkmasıyla da MARC kayıtlarının dağıtımında/paylaşımında Z39.50 protokolü standartlaştırılarak kullanılmaya başlamıştır.

MARC teknik olarak bir işaretleme dilidir. Şekil 1'de yapısal bir örneği bulunan MARC; alan etiketleri, göstergeler, alt alan kodları ve sınırlayıcı işaretten oluşmaktadır.



Şekil 1. MARC'ın Yapısal Elementleri (Kaynak: Byrne, 1998, s. 18)

Genel olarak MARC, oldukça ayrıntılı bir standarttır. Tablo 1'de yer alan temel alanlar ve Tablo 2'de yer alan alt alanlar aracılığı ile farklı türdeki bilgi kaynaklarının tanımlanması veya belirli özelliklerinin tanımlanması mümkündür. Ancak, MARC standardının ayrıntılı olması, öğrenim zorluğu ve maliyeti artırma gibi bir takım dezavantajları beraberinde getirmektedir.

Tablo 1. USMARC'ın Temel Alanları

Alan et.	Alan verisi
0XX	Kontrol bilgisi, kimlikleme, sınıflama numarası vd.
1XX	Temel girişler
2XX	Başlıklar ve başlık paragrafı (başlık, basım, yayın evi adı)
3XX	Fiziksel tanımlama, vd.
4XX	Seri bildirimi
5XX	Notlar
6XX	Konu erişim alanları
7XX	Konu ve seri dışındaki ek girişler
8XX	Seri ek girişleri, vd.
9XX	Yerel gerçekleştirimler için rezerv edilmiştir

Tablo 2. USMARC'ın Bazı Alt Alanları

Alt alan et.	Alan verisi
X00	Kişi adları
X10	Kurum adları
X11	Toplantı adı
X30	Tekbiçim başlıklar
X40	Bibliyografik başlıklar
X50	Konusal terimler
X51	Coğrafi isimler

1960'lı yıllardan itibaren bilgisayarların farklı disiplinlerce kullanılmaya başlamasıyla kütüphanecilik alanının “kataloglama” olarak belirttiği süreç, farklı disiplinlerce “üstveri çıkarma” (metadata extracting) olarak ifade edilmeye başlanmıştır. Üstveri (metadata) kavramı 1960'lı yıllarda coğrafi bilgi sistemleri uzmanları, veri tabanı geliştiricileri ve istatistikçiler tarafından, 1990'lı yıllardan sonra da web toplulukları tarafından yaygın olarak kullanılmıştır (Bhattacharya, 2006).

Üstveri, genel olarak “veri hakkında veri” veya “bilgi hakkında bilgi” biçiminde tanımlanmaktadır. NISO (2004, s. 1), üstveriyi “bilgi kaynağını tanımlayan, açıklayan, yerini belirleyen veya erişimini, kullanımını veya yönetimini kolaylaştıran yapılandırılmış bilgi” olarak tanımlamakta ve üç farklı türe ayırmaktadır:

1. **Tanımlayıcı üstveri:** Keşfetme ve kimlikleme gibi amaçlarla kaynağı tanımlamaktadır. Başlık, öz, yazar ve anahtar kelimeler gibi elementleri içermektedir.

2. **Yapısal üstveri:** Bileşik objelerin nasıl bir araya getirildiğini belirtmektedir. Örneğin, bölümleri biçimlendirmek için sayfaların nasıl düzenleneceği gibi.

3. **Yönetimsel üstveri:** Kaynağın nasıl yaratıldığı, dosya türünün ne olduğu, kaynağa kimin erişebileceği ve diğer teknik detaylar gibi kaynağın yönetiminde yardımcı olabilecek bilgileri sağlamaktadır.

NISO (2004, s. 1-2) ayrıca, üstverinin kaynak keşfetmede, elektronik kaynakların organizasyonunda, birlikte işlerlikte, dijital kimliklemede, arşivleme ve koruma alanlarında kullanılabileceğini belirtmektedir.

Üstveri kümeleri veya şablonları ihtiyaçlar dahilinde birçok disiplin tarafından geliştirilip standartlaştırılmıştır. En çok tanınan ve yaygın olarak kullanılan üstveri standartlarından birkaçını şu şekilde sıralayabilmek mümkündür:

- DC (Dublin Core)
- MARC
- MARCXML
- ISBD (International Standard Bibliographic Description)
- ANZLIS (Australia New Zealand Information Council)
- CIMI (Consortium for the Computer Interchange of Museum Information)
- EAD (Encoded Archival Description)
- EDNA (Education Network Australia)
- GILS (Government Information Locator Service)
- TEI (Text Encoding Initiatives)
- VRA (Visual Resource Association)

Yukarıdaki listede yer alan DC, MARC'dan sonra en yaygın kullanılan üstveri standardıdır. DC, yazarların web kaynaklarını kendi kendilerine tanımlayabilecekleri basit bir setin yaratılması amacıyla, OCLC (Online Computer Library Center) ve NCSA (National Center for Supercomputing Applications) sponsorluğunda, 1995 yılında Dublin – Ohio'da gerçekleştirilen bir atölye çalışmasında (workshop) yaratılmıştır (Weibel, Godby, Miller ve Daniel 1995; NISO, 2004, s. 3).

DC seti, Tablo 3'te yer alan 15 elementten oluşmaktadır (Küçük ve Al, 2003; DC, 2010). 15 elementten ihtiyaç duyulan herhangi biri istenildiği kadar tekrar edilebilmektedir. Ayrıca, söz konusu DC elementleri ANSI/NISA (Z39.85; 2007), Türk Standartlar Enstitüsü (TSE ISO 15836; 2007), ISO (ISO15836:2009; 2009) tarafından standart olarak yayınlanmıştır.

Tablo 3. DC Setinde Yer Alan Elementler ve Açıklamaları

No.	Terim adı	Element	Açıklama
1	Katkı sağlayıcı	Contributor	Kaynağa katkı yapmaktan sorumlu varlık
2	Kapsam	Coverage	Kaynağın uzaysal ya da zamansal özellikleri (Örneğin; coğrafi yer veya zaman periyodu)
3	Yaratan	Creater	Belgeyi meydana getiren kişi ya da tüzel kuruluşlar
4	Tarih	Date	ISO 8601 formatında kaynağın hayat döngüsüyle ilişkili bir periyod veya nokta
5	Tanımlama	Description	Kaynağın tanımlanması (Örneğin; öz, özet, içindekiler vd.)
6	Biçim	Format	Dosya biçimi, uzantısı veya boyutları
7	Kimlikleyici	Identifier	URL gibi kaynak adresi
8	Dil	Language	Kaynağın hangi dilde olduğu
9	Yayıncı	Publisher	Kaynağı elde edilebilir kılan varlık
10	İlişki	Relation	Kaynağın diğer kaynak(lar)la ilişkisi
11	Haklar	Rights	Kaynak üstünde elde tutulan hakların bilgisi
12	Kaynak	Source	Kaynağın hangi kaynaktan türediği (Örneğin; kaynak makale ise hangi dergide yayınlandığı)
13	Konu	Subject	Kaynağın konusu
14	Başlık	Title	Kaynağa verilen ad
15	Tür	Type	Kaynağın türü (Örneğin; roman, makale vd.)

İnsana dayalı bilgi organizasyonunda, çeşitli şablonlara bağlı kalınarak üretilen üstverinin bilgi kaynağını tam olarak temsil ettiği varsayılmaktadır. Erişimi sağlamak amacıyla üretilen sistemler de üstverileri şablonlara bağlı kalarak yapılandırmaya ve birbirleriyle ilişkilendirmeye izin veren veri tabanları üzerine inşaa edilmektedir. Ayrıca, kullanıcıların bilgiye erişebilmesi için bilgi ihtiyaçlarını bilgi kaynağının tanımlanmasında uyulan kurallara göre ifade etmesi gerektiği önceden belirlenmiştir. Bu doğrultuda, kullanıcı bilgi ihtiyacını önceden belirlenmiş kurallara göre ifade ettiğinde erişim sağlanabilmekte, aksi taktirde erişim sağlanamamaktadır. Yao (2004, s. 314), bu tür sistemleri bilgi erişim sistemlerinin bir parçası olan “veri erişim” sistemleri olarak tanımlamaktadır.

Bilginin organizasyonunda diğer bir yaklaşım da makineye dayalı organizasyondur. Bu yaklaşım, literatürde “bilgi depolama ve erişim” (information storage and retrieval) veya kısaca “bilgi erişim” (information

retrieval) olarak geçmektedir. Ayrıca, “bilgi depolama ve erişim” konusu, “bilgi depolama” ve “bilgi erişim” olarak iki farklı çalışma alanına da ayrılabilmektedir. Bunlara ek olarak, bilgi erişim alanı; metin, imaj, ses ve diğer çoklu-ortam bilgi kaynaklarında yer alan bilginin erişimi ile ilgilenen disiplinlerarası bir çalışma alanıdır. Çalışma kapsamında bilgi erişim ile ele alınan bilgi kaynağı türü ise metin tabanlıdır.

Bilgi erişim, kısaca “depolanmış bilgiler içerisinden ilgili (relevant) bilgilerin bulunması” biçiminde tanımlanmaktadır (Dominich, 2008, s. 2). Bilgi erişim, bilgi kaynaklarının depolanması, temsili/gösterimi, organizasyonu ve erişimi konularıyla ilgilenmektedir. Bilgi erişimin veri erişimden ayrıldığı en önemli nokta ise bilgi erişimde belirsizliklerin daha fazla olmasıdır. Ayrıca, bilgi erişim ile veri erişim arasındaki diğer farkları Tablo 4'teki biçimde toplayabilmek mümkündür.

Tablo 4. Veri Erişim ve Bilgi Erişim Özellikleri

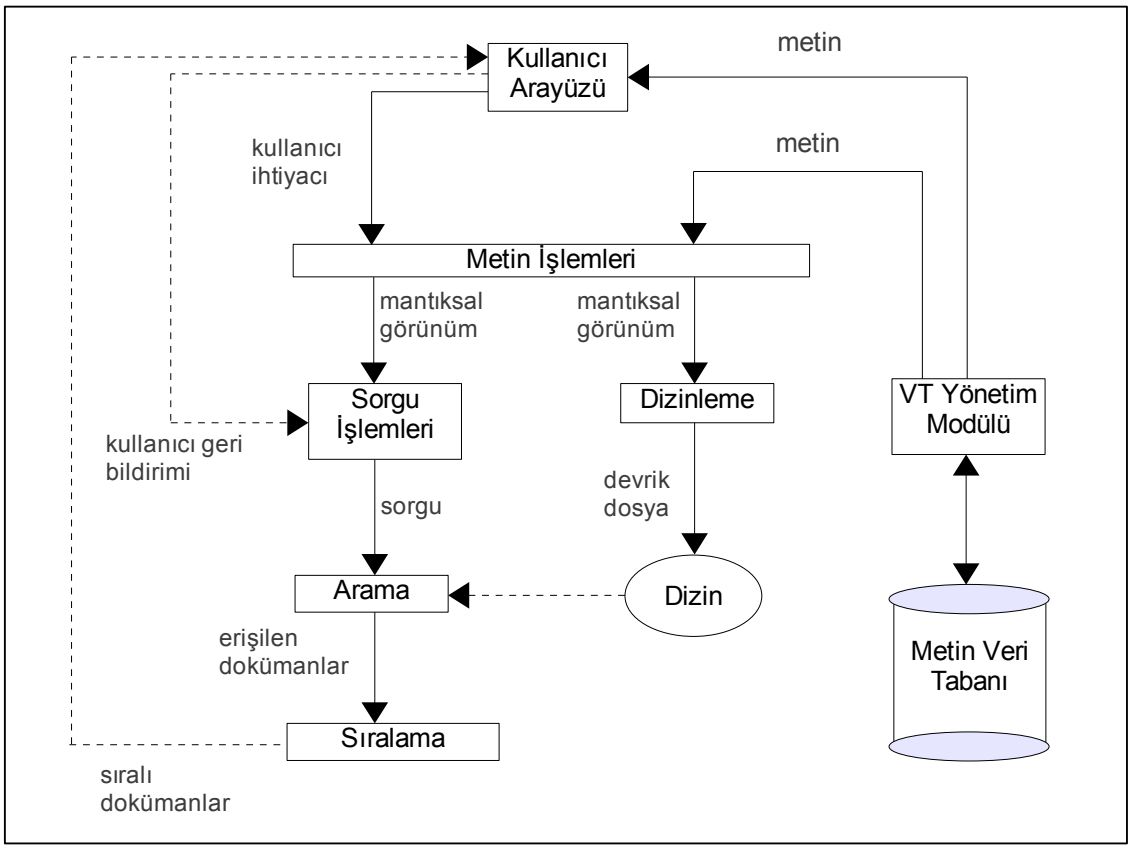
Kriter	Veri erişim	Bilgi erişim
Sorgu eşleşmesi	Tam eşleşme	Kısmi eşleşme, en iyi eşleşme
Sonuç çıkarma	Tümdengelim	Tümevarım
Model	Belirleyici	Olasılık
Sınıflama	Tekil sınıflama	Çoğul sınıflama
Sorgulama dili	Yapay	Doğal
Sorgu şartnamesi	Önceden belirlenmiş	Önceden belirlenmemiş
İstenen öğeler	Eşleştirilebilenler	İlgililer
Hata yanıtı	Hassas	Hassas değil

Kaynak: Van Rijsbergen,1979

Salton (1986), “otomatik metin erişim sistemi” olarak bilgi erişim sistemini “kullanıcı sorgularını cevaplamada doğal dil dokümanlarını aramak amacıyla geliştirilmiş sistem” olarak tanımlamaktadır. Yao (2004, s. 314) ise, herhangi bir şablona bağlı kalmadan, yarı yapılandırılmış veya yapılandırılmamış verilerin organizasyonunu sağlayan ve veri erişim sistemlerine göre belirsizliğin daha büyük rol oynadığı sistemler olarak tanımlamaktadır.

Bilgi erişim sistemlerinde erişim süreci ise Şekil 2'deki gibi işlemektedir. Kaynaklardan bilginin edinilmesiyle başlayan süreci, metin işleme ve dizinleme süreçleri takip etmektedir. Kullanıcının bilgi ihtiyacını bilgi erişim modelinin izin

verdiği biçimde sisteme aktarmasıyla başlayan süreç, sorgu terimlerinin analiz edilmesi ve arama işleminin gerçekleştirilmesiyle devam etmektedir. Arama işleminde, bilgi erişim sisteminin dayandığı modele göre sorgu-doküman benzerlikleri dizin üzerinde hesaplanıp, ilgililiğe dayalı sıralı sonuç kümesi kullanıcıya gösterilmektedir. Bazı bilgi erişim sistemlerinde erişim süreci bu aşamada tamamlanırken, bazı bilgi erişim sistemlerinde kullanıcı, kendisine döndürülmüş olan sonuç kümesi içinden ilgili bulduğu dokümanları ilgililiği daha da artırmak üzere sisteme geri bildirmektedir. Sistem, kullanıcının göndermiş olduğu ilgili dokümanların önemli terimlerinden yeni bir sorgu oluşturup, arama sonuçlarını kullanıcıya iletmektedir.



Şekil 2. Bilgi Erişim Süreci (Kaynak: Baeza-Yates ve Ribeiro-Neto, 1999, s. 10)

2.3 . BİLGİ ERİŞİM MODELLERİ

Bir bilgi erişim sistemi temel olarak belirsizliğin ele alındığı aşağıdaki üç bileşeni bünyesinde barındırmaktadır (Turtle ve Croft, 1997):

1. **Doküman temsili:** Doküman içeriklerini temsil etmek amacıyla saptanan terimler herkes tarafından kabul görmese de dokümanı temsil etmektedir. Bu alanda otomatik tekniklerin kullanımı belirsizliği daha da artırmaktadır ve hangi kavramların içeriği hangi derecede temsil edeceği karmaşıktır.
2. **Bilgi ihtiyaçlarının temsili:** Kullanıcıların bilgi ihtiyaçlarını ifade etme sürecinde de aynı temsil sorunuyla karşılaşılmaktadır. Bilgi ihtiyacı açık bir biçimde temsil edilememektedir. Bu durum, kullanıcının hangi tür dokümanların sistemde yer aldığını görmesiyle arama stratejisini değiştirmesine neden olmaktadır.
3. **Eşleşme fonksiyonu:** Eşleşme fonksiyonunda belirsizlik, bilgi ihtiyacının temsilinden ve doküman temsilinden miras alınmıştır. Bilgi ihtiyacının temsilinde ve doküman temsilinde kesin bir temsil olsa bile belirsizlik devam etmektedir, çünkü aynı kavram pek çok farklı biçimde temsil edilebilmektedir ve tek bir temsilde yer alan kavramlar birbirlerinden bağımsız da değildir.

Yukarıda yer alan üç temel belirsizliğe yönelik getirilen çözümler de “bilgi erişim modeli” olarak adlandırılmaktadır. Çalışma kapsamında, temel bilgi erişim modellerinden Boole, Vektör Uzayı ve Genişletilmiş Boole modelleri ele alınmaktadır.

2.3.1. Boole Bilgi Erişim Modeli

Boole bilgi erişim modeli (BBEM), küme (set) teorisi ve Boole cebirine dayalı olarak geliştirilmiş basit bir bilgi erişim modelidir. BBEM, ilk klasik bilgi erişim modeli olmakla beraber, geniş çevrelerce en çok benimsenmiş modeldir (Dominich, 2001, s. 97). BBEM neredeyse tüm veri tabanı yönetim sistemi üreticilerince desteklenip, geliştirilmiştir. Bu durum, modele ulaşımı kolaylaştırdığı gibi, kullanımını da yaygınlaştırmıştır.

Modelin dayanaklarından Boole cebiri, 0 ve 1 (true ve false) değerlerine dayalı, tümdengelimli matematiksel bir sistemdir (Hyde, 2005, s. 192). "VE" (AND), "VEYA" (OR) ve "DEĞİL" (NOT) olmak üzere farklı üç temel işleç Boole mantığının tasarımında kullanılmaktadır. İşleçlerin doğruluk tablosu ise Şekil 3'teki gibidir.

VE (AND)

A	B	S
0	0	0
0	1	0
1	0	0
1	1	1

VEYA (OR)

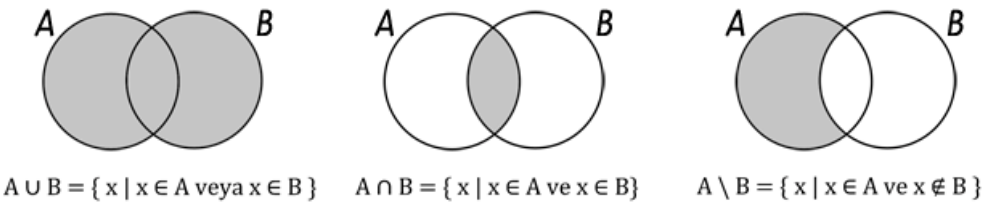
A	B	S
0	0	0
0	1	1
1	0	1
1	1	1

DEĞİL (NOT)

A	S
0	1
1	0

Şekil 3. Boole İşleçlerinin Doğruluk Tabloları (Kaynak: Gillies, 2010)

Modelin bir diğer dayanağı ise küme teorisidir. İyi tanımlanmış (yani belirgin, başka nesnelerden ayırt edilebilir) nesnelerin oluşturduğu herhangi bir topluluğa küme denir ve kümeyi oluşturan nesneler arasında belirgin ortak bir özellik olabildiği gibi olmayabilir de (Özer, 1998). Kümeler arasında *N* tane kümeye, yeni bir küme karşılık getirme biçiminde birçok işlem tanımlanabilmektedir. BBEM için bu işlemlerden en önemlileri Şekil 4'teki "birleşim $(\cup$ veya $\vee)$", "kesişim $(\cap$ veya $\wedge)$" " ve "fark $(\setminus)$"dır.



Şekil 4. Küme İşlemlerinin Venn şeması ("birleşim", "kesişim" ve "fark")

BBEM ise şöyle tanımlanmaktadır:

> Boolean modeli için, dizin terimi ağırlık değişkenlerinin tümü ikilidir (örn., $w_{i,j} \in \{0,1\}$). Bir *q* sorgusu geleneksel bir *Boole* ifadesidir. Sorgu *q* için $\vec{q}_{dnf}$ ayrık normal form olsun. Ayrıca, $\vec{q}_{dnf}$ bağlaç bileşenlerinden

herhangi biri $\vec{q}_{cc}$ olsun. Doküman d_j'nin sorgu q'ya benzerliği şu şekilde tanımlanır*:

$$sim(d_j, q) = \begin{cases} 1, & eğer\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{k_i},\ g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0, & aksi\ takdirde \end{cases}$$

Eğer, $sim(d_j, q) = 1$ ise Boole modeli şunu öngörmektedir; d_j dokümanı sorgu q ile ilgilidir (olmayabilir de). Aksi takdirde, öngörü "doküman ilgili değildir" biçimindedir. (Baeza-Yates ve Ribeiro-Neto, 1999, s. 26-27)

Tanım açılacak olursa; Boolean modelinde, olası dizin terimleri Boole işleçleriyle (VE, VEYA, DEĞİL) birbirlerine bağlanarak sorgu oluşturulur (bir başka ifadeyle, bilgi ihtiyacı formüle edilir). Belirlenen koşullar çerçevesinde; sorgudaki terim(ler)in, dizindeki terim(ler)le çakışması durumunda ilgililik kararı verilir ve *N* kümeye, ilgili olduğu varsayılan yeni bir erişim kümesi tanımlanır. Modele göre, dizin terimleri dokümanda ya geçmektedir ya da geçmemektedir. Bu yüzden, erişim kümesinde ilgili kabul edilen tüm terimlerin ağırlığı 1'dir. Erişim kümesine giremeyen terimlerin ağırlıkları ise 0'dır.

Dokümanları temsil eden terim kümesinden faydalanarak BBEM aşağıdaki gibi örneklendirmek mümkündür:

Dokümanları temsil etmekte kullanılacak terim kümesi *T* aşağıdaki gibi olsun:

T = {t_1 = elma, t_2 = armut, t_3 = kiraz, t_4 = kitap}

Doküman kümesi *D* aşağıdaki gibi olsun:

D = {D_1 , D_2 , D_3 , D_4 , D_5 , D_6 , D_7, D_8}

D kümesindeki her bir elemanı (dokümanı) tanımlamak/temsil etmek üzere şu terimler atansın:

* Tamında, k_i dizin teriminin ağırlığını döndüren fonksiyon g_i'dir (*örn.*, $g_i(\vec{d}_j) = w_{i,j}$).

D_1 = {elma, armut } D_5= {kiraz }

D_2 = {armut, kiraz } D_6= {elma }

D_3 = {elma, armut, kiraz } D_7= {armut }

D_4 = {elma, kiraz} D_8= { kitap}

Sorgu *Q* aşağıdaki gibi olsun*:*

Q = elma ∧ armut ∧ kiraz

İlk aşamada, D_i dokümanlarının S_1 , S_2 ve S_3 erişim kümeleri şunlardır:

$S_1 = \{ D_i \mid \text{elma} \in D_i \} = \{D_1, D_3, D_4, D_6\}$

$S_2 = \{ D_i \mid \text{armut} \in D_i \} = \{ D_1, D_2, D_3, D_7\}$

$S_3 = \{ D_i \mid \text{kiraz} \in D_i \} = \{ D_2, D_3, D_4, D_5\}$

Son aşamada, *Q* sorgusundaki işleme karşılık aşağıdaki erişim kümesi tanımlanır:

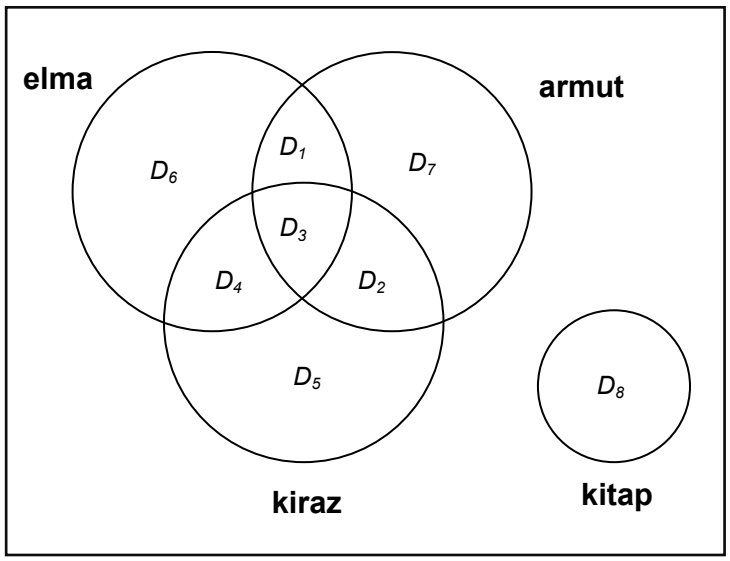
$$\{ D_i \mid D_i \in S_1 \cap S_2 \cap S_3\} = \{D_1, D_3, D_4, D_6\} \cap \{ D_1, D_2, D_3, D_7\} \cap \{ D_1, D_2, D_3, D_7\} = \{D_3\}$$

Sonuç olarak, sadece D_3 dokümanına erişilebilmektir.

Yukarıdaki örneğin devrik dizini Şekil 5'teki gibidir. Devrik dizine ait Venn şeması ise Şekil 6'daki gibidir.

T	***D_1***	***D_2***	***D_3***	***D_4***	***D_5***	***D_6***	***D_7***	***D_8***
elma	1	0	1	1	0	1	0	0
armut	1	1	1	0	0	0	1	0
kiraz	0	1	1	1	1	0	0	0
kitap	0	0	0	0	0	0	0	1

Şekil 5. İkili Devrik Dizin Örneği



Şekil 6. İkili Devrik Dizinin Venn Şeması

BBEM, kullanıcı gruplarının özellikleri göz önünde bulundurulduğunda göreli avantajlar sağlamaktadır. Göker ve Davies (2008, s. 3), modelin uzman kullanıcılarda sistem üzerinde kontrol hissi uyandırdığını, gönderilen sorguya karşılık dokümanın neden geldiğinin kolay anlaşıldığını ve sonuç kümesinin küçük veya büyük gelmesi durumunda hangi işleçlerle istenilen düzeyde sonuç kümesi elde edilebileceğinin kolay anlaşılır olduğunu vurgulamaktadır. Ayrıca, modelin kolay uygulanabilir olması ve hesaplama verimliliği de modelin avantajları arasında sayılabilmektedir (Spoerri, 1995, s.31).

Modelin avantajlarının yanı sıra, dikkate alınması gereken bazı temel dezavantajları da bulunmaktadır. Salton (1984) ve Cooper (1988) genel olarak modelin üç önemli dezavantajı üzerinde durmuştur. Bunlar; Boole formülasyonunun zorluğu, boş çıktı veya fazla yüklü çıktı alınması ve ağırlıklandırma eksikliğidir. Modelin temel problemleri aşağıdaki biçimde açıklanabilir:

- Kullanıcıların doğal dilde kullandıkları “VE” ve “VEYA” sözcükleri bilgi erişim sistemlerinde farklı manalara gelmektedir. Bilgisayar ve Boole cebiri hakkında bilgi sahibi olmayan kullanıcılar VE-VEYA işleçlerinin mantığını kavramada ve sorgu formülize etmede zorluk çekmektedir. Özellikle tecrübesiz kullanıcılar karmaşık sorgularda parantez kullanımı

konusunda hata yapabilmekte ve sistemi kullanabilmek için çok çaba sarf etmektedir.

- Boolean arama taleplerinde VE işleçlerinin fazla kullanılması durumunda boş çıktı ile karşılaşılabilmektedir. VEYA işleçleriyle oluşturulmuş arama taleplerinde ise çok fazla sonuçla karşılaşılabilmektedir.
- Klasik Boolean modelinin erişilen dokümanlarda ilgililik sıralaması yapma konusunda herhangi bir yaklaşımı bulunmamaktadır. Sıralama, yapılandırılmış verilerin karakteristiklerine uygun olarak, “artan-azalan”, “büyüktür-küçüktür” veya “arasında” gibi çeşitli kıstaslara göre yapılabilmektedir.

BBEM’in yukarıda sıralanan temel eksikliklerini giderebilmek üzere birçok çalışma yapılmıştır. Boole sorgularının oluşturulması konusunda insan bilgisayar etkileşimi ve bilgi görselleştirme gibi alanlar kullanıcı dostu arayüz tasarımlarıyla büyük ölçüde sorunların üstesinden gelebilmiştir. Boole sorgusu oluşturma ve boş çıktı veya aşırı yüklü çıktı sorunlarının üstesinden gelebilmek üzere, sorgu genişletmeye veya daraltmaya odaklı “akıllı Boolean” (smart Boolean) (Marcus, 1991) geliştirilmiştir. Ağırlıklandırma ve ilgililik sıralaması sorunlarının üstesinden gelmek üzere Vektör Uzayı modelinden de faydalanılarak Genişletilmiş (Extended) Boole Modeli geliştirilmiştir.

Sonuç olarak, BBEM’in yaygın kullanım alanının “bilgi erişim”den ziyade, “veri erişim” olduğu dikkat çekmektedir. Bunun sebebi ağırlıklandırmanın ikili (binary) olmasına dayanmaktadır. Ağırlıklandırılmanın ikili yapılması “bilgi erişim” modeli olarak tatmin edici olmasa da “veri erişim” için idealdir.

2.3.2. Vektör Uzayı (Vector Space) Bilgi Erişim Modeli

Vektör Uzayı bilgi erişim modeli (VUBEM), temel olarak, BBEM’in ikili ağırlıklandırma kaynaklı sıralama yeteneğinin olmayışının üstesinden gelebilmek üzere istatistiksel yönteme dayalı olarak geliştirilmiş bir bilgi erişim modelidir.

Modelin temel istatistiksel dayanağı ise Luhn tarafından ortaya konulmuştur. Luhn (1957), terimlerin dokümanlardaki geçiş sıklıklarının dokümanı temsil etmede veya doküman için önem belirlemede kullanılmasını önermiştir. Ayrıca, kullanıcıların bilgi ihtiyaçlarını ifade etmek için doküman hazırlayabileceğini, hazırlanan doküman ile dermedeki dokümanların benzerlik derecelerinin ilgiliğe dayalı sıralamalı sorgu sonucunu verebileceğini ortaya koymuştur. Luhn'un benzerlik ölçütü aşağıdaki biçimde ifade edilmektedir.

- Doküman metninden çıkarılmış veya deneyimli dizinleme yapan kişiler tarafından elle atanan terimler ve sorgudan çıkarılmış terimler, doküman içeriğini belirlemede kullanılabilir. Her iki durumda da dokümanlar *terim vektörleri* olarak gösterilebilir. Bu durum aşağıdaki gibi ifade edilebilmektedir.

  Dizin terimleriyle ilgili her bileşen: $d_k = (1 \leq k \leq m)$

  Doküman terim vektörü: $\vec{d} = (d_1, d_2, \dots, d_m)$

- Aynı doküman terimleriyle ilişkili sorgular veya sorgu dokümanları da vektör olarak ifade edilebilir. Bu durum aşağıdaki gibi ifade edilebilmektedir.

  Sorgu vektörü: $\vec{q} = (q_1, q_2, \dots, q_m)$

- Sonuç olarak, sorgu-doküman benzerliği aşağıdaki formül ile hesaplanabilmektedir.

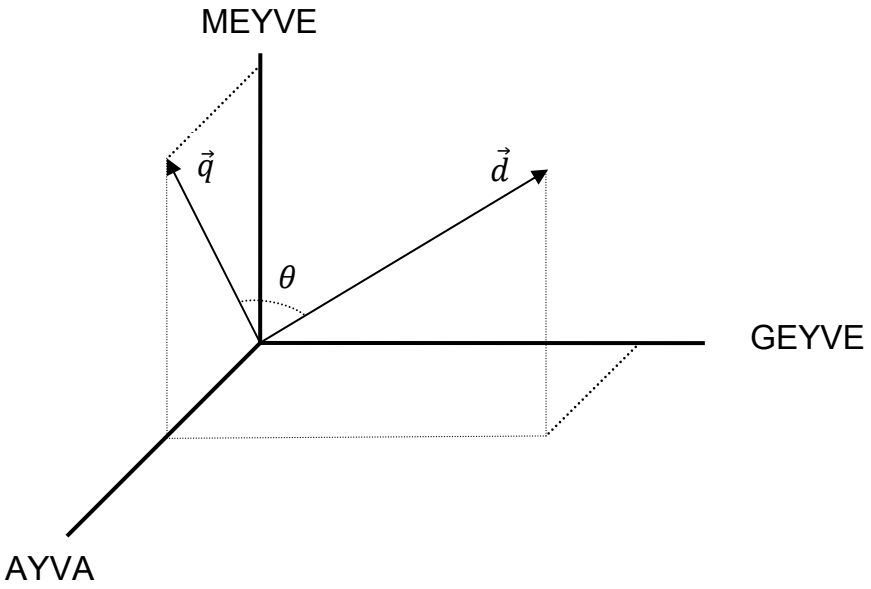
$$score = (\vec{d}, \vec{q}) = \sum_{k=1}^{m} d_k \, . \, q_k$$

- Terim sorguda ve dokümanda geçiyorsa vektör bileşenlerinin değeri 1'dir, geçmiyorsa 0'dır.

Salton, Wong ve Yang ise, (1975) Luhn'un ortaya attığı istatistiksel yaklaşımı geliştirip güçlü bir model ortaya koymuşlardır. Bu doğrultuda, VUBEM'i temel olarak aşağıdaki gibi özetleyebilmek mümkündür:

- İkili ağırlıklandırma ideal bir erişim modeli için oldukça kısıtlıdır. Bu kısıtlamayı ortadan kaldırmak üzere dizin terimlerine, sorgulara ve dokümanlara ikili olmayan ağırlıkların atanması gerekmektedir.
- VUBEM'de, dokümanlar ve sorgular *t* boyutlu vektörler olarak gösterilir.
- Ağırlıklandırma, dokümanlar ile sorgular arasındaki benzerlik derecesinin hesaplanmasında kullanılmaktadır.
- Sonuç olarak, kullanıcılara benzerliği azalan bir sıralama ile sonuç kümesi döndürülebilmektedir.

VUBEM'de hem doküman terimleri hem de sorgu terimleri ağırlıklandırılmaktadır. Dokümandaki ve tüm dermedeki terimlerin önemini terim ağırlıkları göstermektedir. Sorgudaki terimlerin önemini ise sorgudaki terimlerin ağırlıkları belirlemektedir. Ayrıca, dokümanlar; dokümandan çıkarılmış terimlerin vektörü biçiminde kavramsal olarak gösterilmektedir. Vektörlerin boyutu terim sayısı kadardır. Şekil 7, MEYVE, AYVA ve GEYVE terimlerinin doküman ve sorgu vektörlerini göstermektedir.



Şekil 7. VUBEM'de Sorgu ve Doküman Vektörünün Gösterimi

VUBEM'de terimlerin ağırlıklandırılmasının ardından sorgu vektörü ($\vec{q}_j$) ve doküman vektörleri ($\vec{d}_j$) arasındaki benzerliğin hesaplanması gerekmektedir.

Benzerliğin hesaplanmasında, iki vektör arasındaki derecenin kosinüs bağıntısı kullanılmaktadır. Çok boyutlu uzayda, vektörler dik ise açının kosinüsü 0'dır, eğer açı 0^0 ise 1'dir. Bu durumda, 90^0 ile 0^0 arasındaki benzerlik 0 ile 1 arasındaki değerlere tekabül etmektedir. Kosinüs bağıntısı ise aşağıdaki gibidir (Baeza-Yates ve Ribeiro-Neto, 1999, s. 27; Göker ve Davies, 2008, s. 6) .

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{k=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Kosinüs bağıntısının ne anlama geldiğinin anlaşılabilmesinde, bağıntının içerisinde yer alan "ağırlıklandırma" ve "normalizasyon" konuları büyük öneme sahiptir.

Doküman vektöründeki bir terimin ağırlığı pek çok farklı yöntemle belirlenebilmektedir. Terim ağırlıklarının belirlenmesinde en çok bilinen ve yaygın olarak kullanılan yaklaşım ise *tf x idf* biçiminde ağırlıklandırmadır. *tf x idf* biçiminde ağırlıklandırma Salton ve Buckley (1988) tarafından aşağıdaki biçimde ortaya konulmuştur.

$$w_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times \log N/df_j$$

tf x idf biçiminde ağırlıklandırmada, terim ağırlıklarının ($w_{i,j}$) belirlenmesinde iki faktör vardır. Bu faktörlerden ilki *terim sıklığı*dır (term frequency). Terim sıklığı, *j* teriminin *i* dokümanındaki geçiş sıklığını ($tf_{i,j}$) ifade etmektedir. Dokümanda 5 defa geçen bir terim ile 100 defa geçen bir terimin söz konusu doküman için farklı önem taşıması gerekmektedir. Dolayısıyla, terimlerin ağırlıklandırılmasında sadece dokümanda geçen terimlerin sıklıkları kullanılabilmektedir. Öte yandan, sağlıklı bir ağırlıklandırma yapabilmek için terimlerin dokümandaki geçiş sıklıkları tek başına yetersiz kalabilmektedir. Bir dokümanda yüksek sıklıkla geçen terimlerin tüm dermede yüksek sıklıkla geçmesi durumunda, bahsi geçen terimle oluşturulmuş sorgunun neredeyse tüm dermeyle ilgili olması gibi bir sonuç ortaya çıkabilmektedir. Bu istenmeyen

sonucu ortadan kaldırmak üzere ikinci bir faktör olan *devrik doküman sıklığı* (*idf, inverse document frequency*) devreye girmektedir (Spärck-Jones, 1972).

tf x idf biçiminde ağırlıklandırmada, *idf* ($\log(N/df_j)$) faktörü logaritmik bir fonksiyondur ve terimin doküman sıklığının artması durumunda azalma özelliği göstermektedir. Sonuç olarak, *tf x idf* biçiminde ağırlıklandırma sayesinde dermedeki az sayıda dokümanda geçen terimlere yüksek ağırlıklar atanabilmektedir. Tablo 5'te bir milyon dokümanın olduğu varsayılan bir dermedeki *idf* parametresi örneklendirilmektedir.

Tablo 5. Idf Parametresi Örneği

Terim	df_j	idf_j
elma	1	6
armut	100	4
kiraz	1.000	3
vişne	10.000	2
bir	100.000	1
ve	1.000.000	0

VUBEM'de dermede yer alan dokümanların uzunlukları göz önünde bulundurulduğunda, sadece *tf x idf* biçiminde ağırlıklandırmanın da yetersiz kaldığı durumlar ortaya çıkabilmektedir. Bu sorunun üstesinden gelebilmek üzere doküman uzunlukları *normalize* edilmektedir. Normalizasyon yapmanın temelinde yatan gerekçeler ise şunlardır (Singhal, Salton, Mitra, ve Buckley, 1995; Singhal, Buckley ve Mitra, 1996):

- **Yüksek Terim Sıklıkları:** Uzun dokümanlar aynı terimleri tekrarlı olarak kullanmaktadır. Sonuç olarak, uzun dokümanlar için terim sıklığı faktörleri kısa dokümanlara göre geniş olabilmekte ve bu durum uzun doküman terimlerinde sorgu-doküman benzerliğinin artmasına neden olabilmektedir.
- **Fazla Terim:** Uzun dokümanlar pek çok farklı/ayrık terimi bünyelerinde barındırmaktadır. Bir başka ifadeyle, uzun dokümanlar fazla sayıda konuyla ilgilidir. Bu durum, kısa dokümanlarda işlenen az sayıda konuya ait terimlerle fazla konuyu işleyen dokümanlardaki terimlerin bir

tutulmasına neden olmaktadır. Sonuç olarak, uzun dokümanların bulunduğu bir dermede yapılan arama sonuçlarında farklı konularla da ilgili olan dokümanlara erişilmektedir.

Terim ağırlıklarının doküman uzunluk normalizasyonu, uzun dokümanların kısa dokümanlara göre avantajını ortadan kaldırmak üzere kullanılmaktadır. Bilgi erişim sistemlerinde birçok[1] normalizasyon tekniği kullanılmaktadır. Aşağıdaki biçimde hesaplanan kosinüs normalizasyonu, hem yüksek terim sıklığını hem de fazla terim kullanım sorununu tek bir adımda gidermeyi hedefleyerek VUBEM'de en yaygın kullanılan normalizasyon tekniği olmuştur (Salton, Wong, Yang, 1975).

$$\sqrt{{w_1}^2 + {w_2}^2 + \ldots + {w_t}^2}$$

Öte yandan, Singhal, Buckley ve Mitra'nın (1996) yaptığı bir çalışmanın bulguları, kosinüs normalizasyonun erişimde kısa dokümanlara iltimas göstermeye meyilli olduğunu göstermiştir. Aynı çalışmada, bu sorunun üstesinden gelmek üzere eksen doküman uzunluğu normalizasyonu (pivot document length normalization) geliştirilip, TREC dermesiyle test edilmiştir ve klasik kosinüs normalizasyonuna göre %18,3 gelişme elde edilmiştir. Geliştirilen normalizasyon formülü ise aşağıdaki gibidir.

$$\frac{\frac{1 + \log(tf)}{1 + \log(ortalama\ tf)}}{(1 - eğim) \times eksen + eğim \times \ tekil\ terim\ sayısı}$$

[1] Normalizasyon tekniklerinden birkaçı aşağıdaki gibidir:

- **Maksimum *tf* normalizasyonu:** SMART için: $0{,}5 + 0{,}5 \times \frac{tf}{\max\_tf}$ (Salton ve Buckley, 1988), INQUERY için: $0{,}4 + 0{,}6 \times \frac{tf}{\max\_tf}$ (Turtle ve Croft, 1989).
- **Byte Uzunluk Normalizasyonu:** Okapi sistemi için (b, genel olarak 0,75 gibi bir sabittir):

$$\frac{tf}{2 \times \left(1 - b + b \times \frac{doküman\ uzunluğu}{ortalama\ doküman\ uzunluğu}\right) + tf}$$

(Robertson, Walker, Jones, Hancock-Beaulieu ve Gatford, 1995).

Doküman uzunluklarının bilgi erişim performansını olumsuz yönde etkilemesi, VUBEM'in önde gelen kısıtlamalarından sayılmaktadır. Bu sorunu çözüme kavuşturmak amacıyla yapılan çalışmalarda önemli ilerlemeler sağlanmış olsa da halen sorun üzerinde çalışılmaktadır. VUBEM'in önemli sorunlarından bir başkası da kullanıcıların bilgi ihtiyaçlarını sisteme doğru aktararak bilgi ihtiyacına cevap verebilecek nitelikteki sonuçların alınmasıdır. Bu amaçla, sorgu formülasyonunda sorgu genişletme ve ilgililik geribildirimi çözümleri üretilmiştir. İlgililik geribildiriminde, kullanıcılar sorgularını sisteme gönderdikten sonra dönen sonuçlar arasından (10 veya 20 sonuç arasından) ilgili olanları seçmektedir. Kullanıcılara ilgili sonuçların seçtirilmesindeki amaç, dokümanlarda geçen ilgili terimlerin belirlenmesidir. Bu sürecin sonunda beklenen etki ise yeniden oluşturulacak olan sorgu ile olabildiğince ilgili dokümanlara erişim sağlamak, ilgisiz dokümanları elemektir. Sorgu genişletmede ise ilgili olarak belirlenmiş dokümanlardaki önemli terimler sorguya eklenmektedir. Bunlara ek olarak, ilgili bulunan terimlerin ayrıca ağırlıklandırılmasıyla sonuç kümesinin ilgililiğinin artırılması da ilgililik geri bildirimi ile elde edilebilecek sonuçlardan bir başkasıdır. Küçük dermelerle yapılan deneylerde bu yöntemlerin duyarlılığı artırdığı tespit edilmiştir (Baeza-Yates ve Ribeiro-Neto, 1999, s. 30, 118).

2.3.3. Genişletilmiş (Extended) Boole Modeli

Genişletilmiş Boolean bilgi erişim modeli (GBBEM), 1983 yılında Salton, Fox ve Wu tarafından duyurulmuştur. GBBEM temel olarak, BBEM'in ilgililiğe dayalı sıralamalı sonuç verememe sorununa VUBEM ile çözüm getirmeyi amaçlamıştır. Böylelikle, sorgu oluşturmada Boole işleçlerinin sağladığı esneklikle VUBEM'in ilgililiğe dayalı sıralamalı sonuç kümesi döndürme özelliği birleştirilerek her iki modelden daha üstün bir model ortaya konulmuştur.

BBEM'de olduğu gibi GBBEM'de de, ilk etapta sorguda geçen terimlerin dokümanda yer alıp almadığı kontrol edilmektedir. Sorgu kriterlerinin sağlandığı dokümanlar BBEM'den farklı olarak 0 ile 1 arasında ilgililik değeri almaktadır.

Sorgu kriterlerini sağlayamayan dokümanlar ise 0 değerini alarak sonuç kümesine girememektedir. Ağırlık ise aşağıdaki biçimde hesaplanmaktadır.

$$w_{i,j} = tf_{norm\ i,j} \times idf_{norm\ i}$$

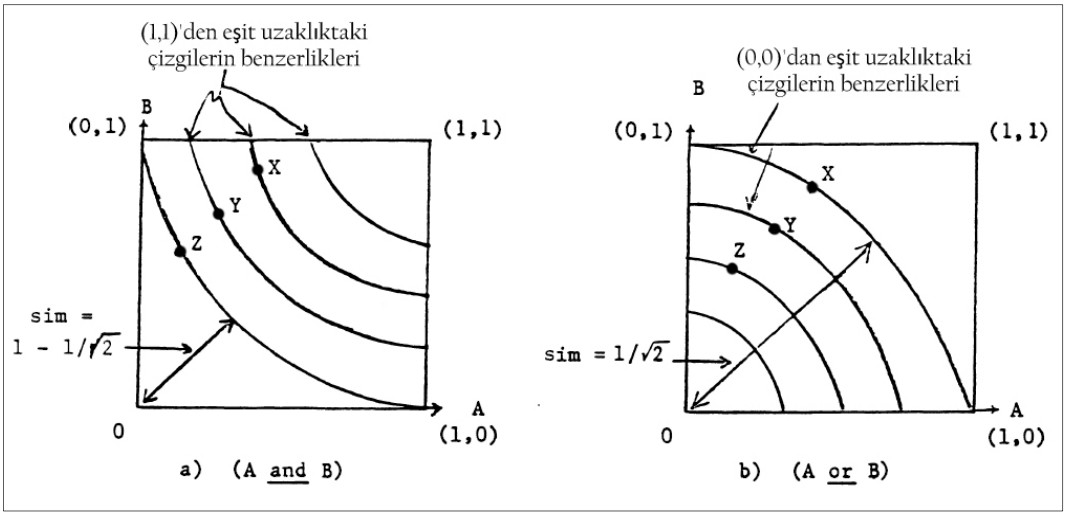
Yukarıdaki denklemin bileşenleri aşağıdaki gibidir.

- $w_{i,j} =$ *j* dokümanda yer alan *i* teriminin ağırlığı.
- $tf_{norm\ i,j} = \frac{tf_{\ i,j}}{tf_{max\ i,j}}$
  - $tf_{norm\ i,j}$: *j* dokümanda yer alan *i* teriminin normalize edilmiş sıklığı.
  - $tf_{\ i,j}$: *j* dokümanda yer alan *i* teriminin sıklığı.
  - $tf_{max\ i,j}$: *j* dokümanda yer alan *i* teriminin maksimum sıklığı.
- $idf_{norm\ i} = \frac{idf_i}{idf_{max\ g}}$
  - $idf_{norm\ i}$: *c* dermesinde yer alan *i* teriminin normalize edilmiş devrik doküman sıklığı.
  - idf_i: *c* dermesinde yer alan *i* teriminin devrik doküman sıklığı.
  - $idf_{max\ g}$: *c* dermesinde yer alan genel *g* teriminin maksimum devrik doküman sıklığı.

AND ve OR işleçlerinin iki boyutlu bir haritada gösterimi ise Şekil 8'deki gibidir. Şekil 8 (a)'daki AND sorguları için $(1,1)$ noktası; terimlerin belgede geçtiğini, Şekil 8 (b)'deki OR sorguları için $(0,0)$ noktası ise terimlerin belgede geçmediğini ifade etmektedir. Bu durumda, ağırlıklı *x* ve *y* terimlerine sahip D belgesinin $(0,0)$ ve $(1,1)$ noktalarının Öklid (Euclidan) uzaklıkları; $(0,0)$ için $\sqrt{(x-0)^2+(y-0)^2}$, $(1,1)$ için $\sqrt{(1-x)^2+(1-y)^2}$ olmaktadır. Bu uzaklıklar da $(0,0)$ ve $(1,1)$'in köşegeni olan $\sqrt{2}$ değerine (en büyük uzaklık değeri) bölünerek normalleştirilmektedir. Bunun sonucunda AND ve OR sorguları için benzerlikler aşağıdaki gibi olmaktadır.

$$sim(q\_OR, d) = \sqrt{\frac{x^2+y^2}{2}}$$

$$sim(q_{AND}, d) = 1 - \sqrt{\frac{(1-x)^2+(1-y)^2}{2}}$$



Şekil 8. GBBEM'de AND İçin $(1,1)$ Noktasından ve OR İçin $(0,0)$ Noktasından Eşit Uzaklık Çizgileri (Kaynak: Salton, Fox ve Wu,1982, s. 47)

Şekil 8'e göre sorgu ağırlıkları 0 ve 1 ise, doküman dört köşenin ($(0,0)$, $(0,1)$, $(1,0)$ veya $(1,1)$) birinde yer almak zorundadır. Bu köşegenler; OR sorguları için 0, $1/\sqrt{2}$, 1 değerleriyle, AND sorguları için ise 0, $1-1/\sqrt{2}$, 1 değerleriyle sınırlıdır (Baeza-Yates ve Ribeiro-Neto, 1999, s. 39). Böylece, AND sorgularında $(1,1)$ noktasından başlamak üzere artan sıralı sonuç kümesi dönmektedir. OR sorgularında ise $(0,0)$ noktasından başlamak üzere azalan sıralı sonuç kümesi döndürülmektedir.

GBBEM'de Öklid uzaklıkları ile birlikte, sorgu ile belirtilmiş p-Norm ($1 \leq p \leq \infty$) parametresi de önemlidir. *p* parametresi ile AND ve OR sorgularındaki, sorgu-doküman benzerlikleri şu sekilde hesaplanmaktadır.

$$sim(q_{OR}, d) = \sqrt[p]{\frac{x_1^p + x_2^p + \cdots + x_n^p}{n}}$$

$$sim(q_{AND}, d) = 1 - \sqrt[p]{\frac{(1 - x_1)^p + (1 - x_2)^p + \cdots + (1 - x_n)^p}{n}}$$

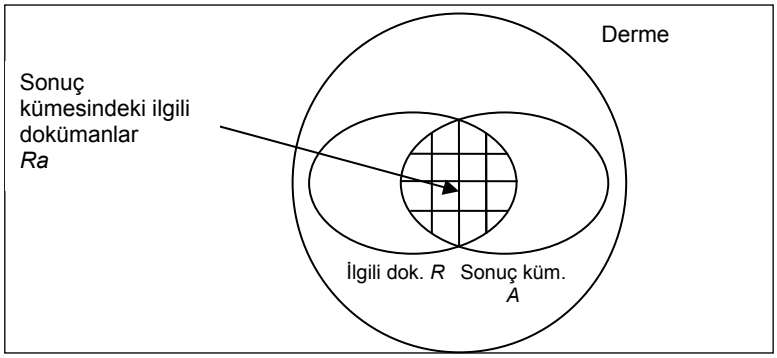
p parametresi 1 olduğunda AND ve OR uzaklığı yok sayılarak vektör uzayı oluşturulmaktadır. *p* parametresi sonsuz olduğunda ise AND ve OR sorguları BBEM'deki gibi kısıtlayıcı olmaktadır. Dolayısıyla, *p* parametresinin 1 ile sonsuz arasında olması durumunda GBBEM modeli Boole işleçlerini de kullanarak VUBEM gibi davranmaktadır. Ayrıca, p-Norm aşağıdaki örnekte olduğu gibi AND ve OR sorgularının gruplanmasında da kullanılabilmektedir. Sonuç olarak, kullanıcılar bilgi ihtiyaçlarını formüle etmede istedikleri kadar terimi AND ve OR işleçleriyle bağlayabilmektedir.

$$sim\left(q_{(k1\ AND\ k2)OR\ k3}, d\right) = 1 - \sqrt[p]{\frac{(1 - \sqrt[p]{\frac{(1 - w_{k1})^p + (1 - w_{k2})^p}{2}})^p + w_{k3}^p}{2}}$$

2.4. PERFORMANS DEĞERLENDİRME

Bilgi erişim sistemlerinin performanslarının değerlendirilmesi, "ilgililiğin" ön plana çıkarıldığı anma (recall) ve duyarlılık (precision) kriterlerine dayanmaktadır (Kent ve diğerlerinden (1955) aktaranlar: Saracevic, 1995; Cen, Lui, Zhang, Ru ve Ma, 2009). Şekil 9'dan yola çıkılarak, anma ve duyarlılık aşağıdaki biçimde ifade edilebilmektedir (Baeza-Yates ve Riberio-Neto, 1999, s. 75).

$$Anma = \frac{Ra}{R} \qquad Duyarlılık = \frac{Ra}{A}$$



Şekil 9. Duyarlılık ve Anmanın Görselleştirilmesi

Anma ve duyarlılık kriterleri tüm bilgi erişim sistemlerinin performansını ölçmede kullanılmasa da (örneğin, web arama motorlarında veya geniş hacimli dermelerde test edilen bilgi erişim sistemlerinde) küçük dermeler için kullanılabilmektedir.

Bilgi erişim sistemlerinin performanslarının değerlendirmesinde birden fazla soru sorulduğu için 11 anma basamağında (0, 0.1, 0.2, ..., 1) ortalama (average) duyarlılık aşağıdaki biçimde hesaplanmaktadır (Baeza-Yates ve Riberio-Neto, 1999, s. 77).

$$P(r) = \sum_{i=1}^{Nq} \frac{Pi(r)}{Nq}$$

Yukarıdaki formülde:

P(r) : *i.* anma basamağındaki ortalama duyarlılığı,

Pi(r) : *i.* sorgu için *r* anma basamağındaki duyarlığı,

Nq : Kullanılan sorgu sayısını ifade etmektedir.

Ayrıca, her sorgu için anma 11 standart anma basamağından ayrık olabilmektedir. Bu durumda, söz konusu anma basamağındaki duyarlılık değerine; kendisi ile kendinden sonra gelen anma basamağındaki maksimum değer atanarak interpolasyon (interpolation) uygulanmaktadır (Baeza-Yates ve Riberio-Neto, 1999, s. 77). İnterpolasyon sonucunda, 11 adet anma basamağı elde edildiği gibi anma-duyarlılık grafiği daha anlamlı hale gelmektedir (Manning, Raghavan ve Schütze, 2008).

Performans değerlendirmede kullanılan başka bir teknik de kullanıcı merkezli değerlendirme tekniğidir. İngilizce literatürde “anma normalizasyonu” (recall normalization) olarak geçen bu değerlendirme tekniği, Türkçe literatürde (Tonta, Bitirim ve Sever, 2002, s. 26) “normalize sıralama” olarak geçmektedir. Bu teknikle, ilgisiz dokümanların ilgili dokümanların önüne geçmesi durumunda, performans anma ve duyarlığa nazaran daha olumsuz etkilenmektedir. Normalize sıralamada (R_{norm}), sonuç kümesindeki tüm dokümanların ilgili olması

durumunda sonuç 1 değerini, tümü ilgisizse 0 değerini almaktadır. 0 ve 1 arasındaki değerleri hesaplamak için aşağıdaki formülden yaralanılmaktadır (Bollmann, 1983; Yao, 1995).

$$R_{norm} = \frac{1}{2}\left(1 + \frac{C^{+} - C^{-}}{C^{max}}\right)$$

Yukarıdaki formülde:

R_{norm} = Normalize sıralama,

C^{+} = erişim çıktısında ilgili belgelerin ilgisiz belgelerin önünde yer aldığı belge çiftleri sayısı,

C^{-} = erişim çıktısında ilgisiz belgelerin ilgili belgelerin önünde yer aldığı belge çiftleri sayısı,

C^{max} = mümkün olan en fazla C^{+} sayısıdır.

3 . BÖLÜM

APACHE LUCENE

3.1 . GİRİŞ

Apache Lucene, bilgi erişim alanında VUBEM ve BBEM temel alınarak geliştirilmiş yüksek performanslı ve ölçeklenebilir bir “uygulama programlama arayüzü”dür (API). Lucene, tek başına bir bilgi erişim sistemi veya arama motoru değildir. Uygulamalara dizinleme ve arama fonksiyonları katmaktadır (Gospodnetić ve Hatcher, 2005, s. 7). Açık kaynak kodlu olarak Java ile geliştirilen Lucene; Delphi (Mutis[1]), Perl (Plucene[2]), C# (Lucene.Net[3]), C++ (CLucene[4]), Phyton (PyLucene[5]), Ruby (Ferret[6]) ve PHP (Zend_Search_Lucene[7]) dillerine de çevrilmektedir. Lucene versiyonu aynı olmak koşuluyla, herhangi bir programlama dili ile oluşturulmuş standart Lucene dizini farklı diller tarafından yazılıp okunabilmektedir. Ekim 2010 tarihi itibariyle en yüksek Lucene Java versiyonu ise 3.02'dir.

Lucene, Doug Cutting tarafından arama motoru geliştirme amacıyla SorceForce'da açık kaynak kodlu bir proje olarak geliştirilmeye başlanmış ve ilk sürümü (0.01) 2000 yılında yayınlanmıştır. 2001 yılında, popüler Apache Jakarta ailesi projelerinden biri olarak ticari kullanıma izin veren Apache Software License ile lisanslanarak geliştirilmeye devam edilmiştir. Şubat 2005'de ise Apache Software Foundation'ın üst düzey bir projesi haline gelmiştir. Proje, başarılı uygulama örneklerini temel alarak bilgi depolama ve erişim konusunu kapsayan birçok alt proje üretmiştir. Daha sonra, alt projeler gelişerek başlı başına üst düzey projeler haline gelmiştir.

---

1 Mutis web sitesi: http://sourceforge.net/projects/mutis/
2 Plucene web sitesi: http://search.cpan.org/~tmtm/Plucene-1.25/lib/Plucene.pm
3 Lucene.Net web sitesi: http://lucene.apache.org/lucene.net/
4 Clucene web sitesi: http://clucene.sourceforge.net/
5 PyLucene web sitesi: http://lucene.apache.org/pylucene/
6 Ferret web sitesi: http://ferret.davebalmain.com/
7 Zend_Search_Lucene web sitesi: http://framework.zend.com/manual/en/zend.search.lucene.html

Lucene projelerinden en önemlisi ise Nutch'dur. Nutch, web gezgini (crawler) olan, HTML veya diğer doküman türlerini derleyebilen (parser) ve temelinde Lucene olan açık kaynak kodlu bir “arama motoru”dur (*About Nutch*, 2010). Nutch'un arama motoru olarak geliştirilmesi; bilgi keşfetme, metin ve üstveri çıkarma, bilgi depolama, bilgi işleme ve bilgi erişim araçlarının geliştirilmesini zorunlu kılmıştır. Bu bağlamda, metin ve üstveri çıkarma amacıyla Apache Tika[8], Google File Sytem'den esinlenilerek dağıtık depolama dosya sistemi olarak Hadoop, Google Big Table'dan esinlenilerek geniş veri kümelerini gerçek zamanlı okuyup yazabilen kolon-yönelimli veri tabanı Hbase, Lucene projesinin altında geliştirilmiştir (White, 2009, s. 9, 343). Her üç araç da şu an başlı başına üst düzey birer projedir ve Hbase'in temelinde yatan Hadoop; Yahoo![9] ve Facebook[10] gibi iki büyük şirket tarafından da geliştirilerek kullanılmaktadır. Bu iki araç büyük ölçekli bilgi erişim sistemlerinin geliştirilmesinde sıkça kullanılmaktadır.

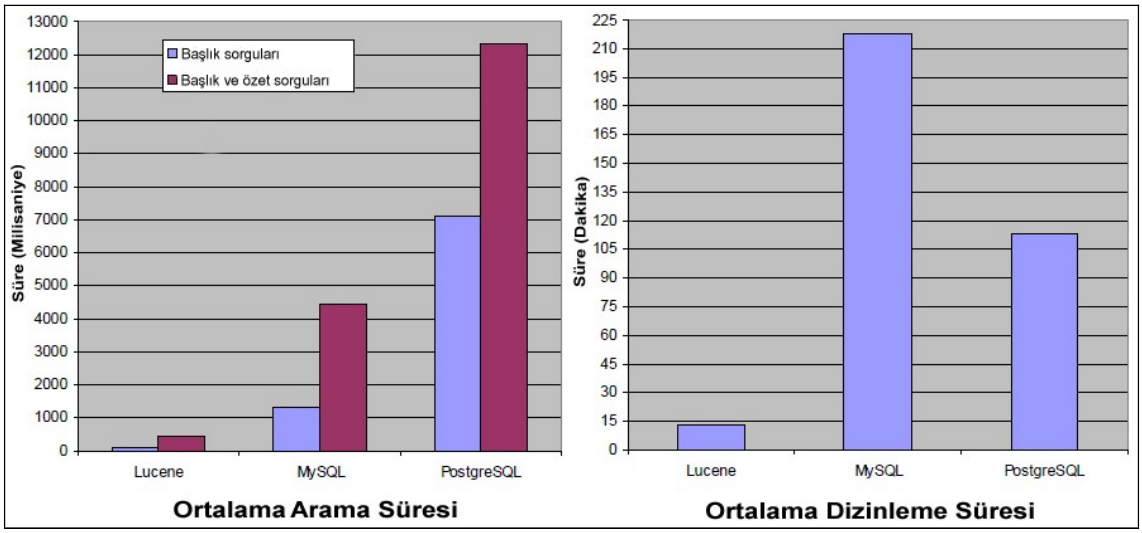
Nutch'un önemli bileşenlerinden bir diğeri de bilgi erişim araçlarıdır. Nutch, bilgi erişim araçları bütünü olarak Solr'u kullanmaktadır. Sorl ise, temelinde Lucene olan kurumsal arama sunucusudur (Smiley ve Pugh, 2009, s. 7). Apache Tomcat gibi herhangi bir Java taşıyıcısı (container) üzerinde çalışabilen Solr, web tabanlı bilgi erişim sistemi geliştirmede kullanılmaktadır. Sonuç filtreleme (faced search) özelliği bulunan Solr, kullanıcı sorgularına karşılık JSON veya XML formatında sonuç kümesi döndürmektedir. Solr, modern alış veriş sitelerinden, kütüphane otomasyonlarına kadar çok geniş kullanım alanlarına sahiptir. Örneğin; Türkiye'de ULAKBİM tarafından geliştirilen TO-KAT (Ulusal Toplu Katalog) projesi Ağustos 2009 tarihinden bu yana Sorl kullanmaktadır (*Toplu Katalog*, 2010). Açık erişim konusunda Solr'un güzel bir uygulama örneği ise Europeana[11] projesidir (*Lucid Imagination,* 2010). Ayrıca, Solr temel alınarak özel ihtiyaçlara yönelik araçlar da geliştirilebilmektedir. Öneğin; Solrmarc, MARC kayıtlarının normalizasyonu ve dizinlenmesi için Solr projesinden türemiş önemli bir projedir (*Solrmarc: Indroduction*, 2010). “Gelecek nesil katalog”

8 Apache Tika web sitesi: http://tika.apache.org/
9 Y! Hadoop dağıtımı: http://developer.yahoo.com/hadoop/
10 Facebook Hadoop geliştiricileri: http://developers.facebook.com/opensource/
11 Europeana web sitesi: http://www.europeana.eu/portal/

olarak nitelenen Blacklight[12] ve VuFind[13] iskeletlerinde Solrmarc kullanmaktadır.

Gerçek hayat projelerine ek olarak, Lucene birçok bilimsel deneyde de kullanılmaktadır. Ayrıca, Lucene'in bilimsel deneylerde kullanımını yaygınlaştırmak amacıyla performans ölçüm araçları geliştirmek üzere “Open Relevance”[14] projesi başlatılmıştır. Projenin tamamlanmasının ardından Lucene'in bilimsel deneyde de yaygın olarak kullanılması muhtemeldir.

Bilgi erişim sistemi tasarımlarında yaygın olarak kullanılan İVTYS'lerle Lucene kıyaslandığında ise Lucene'in hem dizinleme ve sorgulama hızında hem de bilgi erişim performansında MySQL ve PostgreSQL gibi tam metin dizinleme yeteneğine sahip olan rakiplerine karşı üstünlük sağladığı görülmektedir. Konuya ilişkin 2008 yılında 408.305 adet haberin yer aldığı Milliyet koleksiyonu ile yapılan bir araştırmanın (Arslan ve Yılmazel) bulguları Şekil 10 ve Şekil 11'de yer almaktadır.

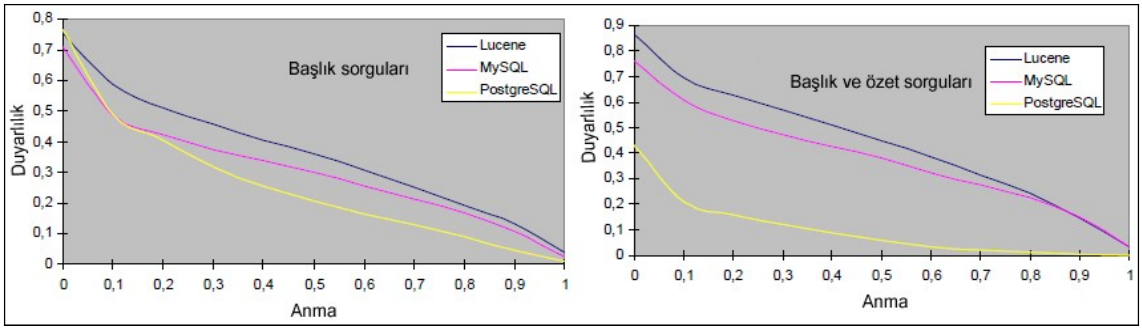


Şekil 10. Lucene ile MySQL ve PostgreSQL'in Dizinleme ve Arama Sürelerinin Karşılaştırılması (Kaynak: Arslan ve Yılmazel, 2008)

12 Ayrıntılı bilgi için http://projectblacklight.org/ adresi ziyaret edilebilir

13 Ayrıntılı bilgi için http://vufind.org/ adresi ziyaret edilebilir.

14 Open Relevance projesine http://lucene.apache.org/openrelevance/ adresinden erişebilmek mümkündür.



Şekil 11. Lucene ile MySQL ve PostgreSQL'in Bilgi Erişim Performansının Karşılaştırılması (Kaynak: Arslan ve Yılmazel, 2008)

Lucene API dokümantasyonu[15] incelendiğinde, Lucene'in hızlı uygulama geliştiricileri bilgi erişimin karmaşık sayılabilecek konularından olabildiğince uzak tutmayı hedeflediği dikkat çekmektedir. Birçok faklı dil için yüksek kalitede geliştirilmiş metin analizcileri, gövdeleme algoritmaları, benzerlik algoritmaları ve çeşitli sorgu modelleri kullanıma hazır biçimde API bünyesinde yer almaktadır. Diğer taraftan, bilgi erişim konusunda derinlemesine bilgi sahibi olan yazılım geliştiricilere de sofistike bilgi erişim sistemleri geliştirme fırsatı yaratmaktadır.

Lucene, içerisinden metin veya metin formatında üstveri çıkartılabilen her türlü elektronik bilgi kaynağını (html, pdf, xml, .doc, VTYS, vd.) dizinleyebilmektedir. Ancak, bilgi kaynaklarından metin veya üstveri çıkartabilecek herhangi bir araç Lucene API bünyesinde bulunmamaktadır. Dosya formatına göre metin veya üstveri çıkartabilen özel araçların ayrıca edinilmesi veya geliştirilmesi gerekmektedir. Çalışma kapsamında kullanılan Tika bu sorunun üstesinden gelebilmektedir. Tika, farklı formatlardaki kaynaklardan metin ve üstveri çıkarmak amacıyla açık kaynak kodlu birçok API'yı bünyesinde toplamış bir API'dır. Tika'nın kulandığı API'lar ve desteklediği formatlar Tablo 6'da yer almaktadır.

Lucene ile geliştirilmiş bir bilgi erişim sistemini uygulama katmanı ve Lucene katmanı olarak Şekil 12'deki gibi iki katmana ayırabilmek mümkündür. Uygulama katmanında; bilgi toplama (metin ve üstveri çıkarma), kullanıcı

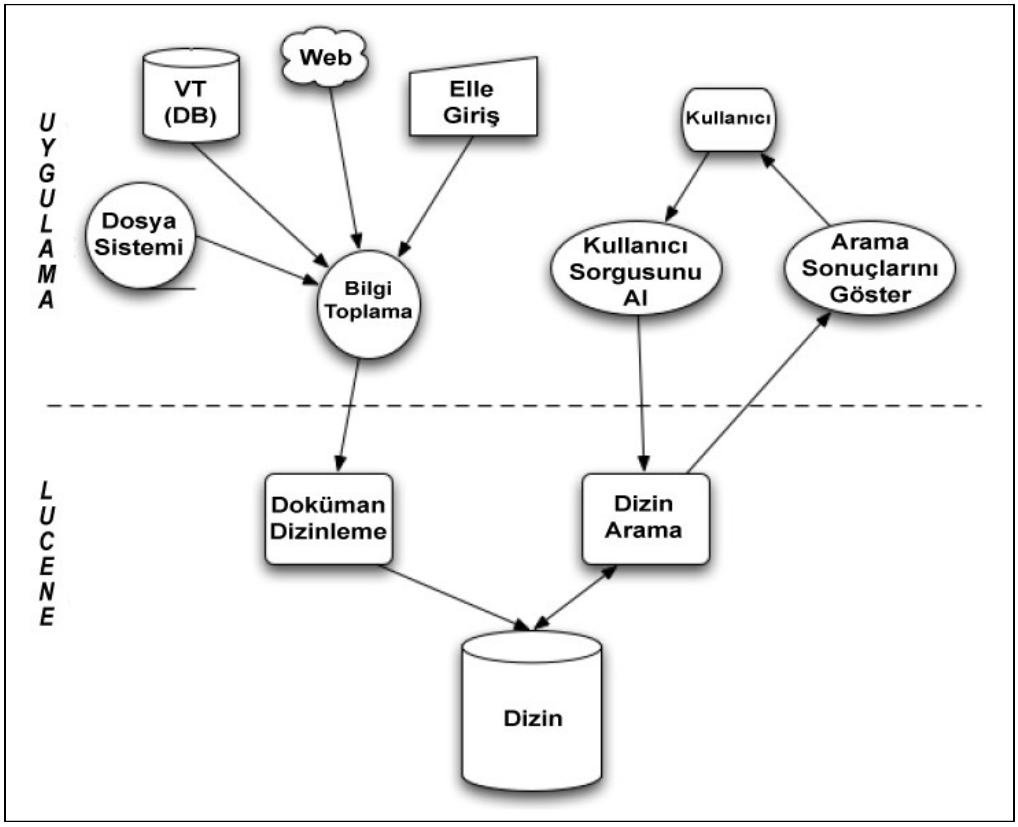
15 Lucene 3.0.2 dokümantasyonuna http://lucene.apache.org/java/3_0_2/api/all/index.html adresinden erişim sağlamak mümkündür. Ayrıca, çalışmanın Lucene API ile ilgili kaynak gösterilmeyen bölümleri bu dokümantasyona dayandırılmıştır.

sorgularının alınması ve sonuç kümesinin görüntülenme süreçleri ele alınmaktadır. Lucene katmanında ise dizinleme ve dizin arama süreçleri ele alınmaktadır. Lucene katmanında yer alan süreçler 3.2, 3.3, 3.4, 3.5 ve 3.6 bölümlerinde ayrıntılı olarak işlenmektedir.

Tablo 6. Apache Tika'nın Metin veya Üstveri Çıkartabildiği Dosya Türleri ve Kullandığı API'lar

Format	Kütüphane
Microsoft's OLE2 dokümanları (Excel, Word, PowerPoint, Visio, Outlook)	Apache POI
Microsoft Office 2007 OOXML	Apache POI
Adobe Portable Document Format (PDF)	PDFBox
Rich Text Format (RTF)—Gövde metni (üstveri yok)	Java Swing API (RTFEditorKit)
Düz metin karakter kümesi bulma	ICU4J
HTML	CyberNeko
XML	Java'nın javax.xml sınıfları
ZIP Arşivleri	Java'nın zip sınıfları, Apache Commons Compress
TAR Arşivleri	Apache Ant, Apache Commons Compress
AR Arşivleri	Apache Commons Compress
CPIO Arşivleri	Apache Commons Compress
GZIP sıkıştırma	Java'nın sınıfları (GZIPInputStream) , Apache Commons Compress
BZIP2 sıkıştırma	Apache Ant, Apache Commons Compress
İmaj formatları (sadece üstveri)	Java'nın javax.imageio sınıfları
Java sınıf (class) dosyaları	ASM (JCR-1522)
Java JAR dosyaları	Java'nın zip sınıfları ve ASM, Apache Commons Compress
MP3 (ID3v1 etiketleri)	Direkt gerçekleştirim
Diğer ses formatları (wav, aiff, au)	Java'nın sınıfları (javax.sound.*)
OpenDocument	XML doğrudan derlenmektedir (parse)
Adobe Flash	FLV dosyalarından üstveri doğrudan derlenmektedir (parse)
MIDI dosyaları (gömülü metin, örn. Şarkı sözleri)	Java'nın kendi sınıfları (javax.audio.midi.*)
WAVE (üstveri)	Java'nın kendi sınıfları (javax.audio.sampled.*)

Kaynak: McCansless, Gospodnetić ve Hatcher, 2010, s. 237



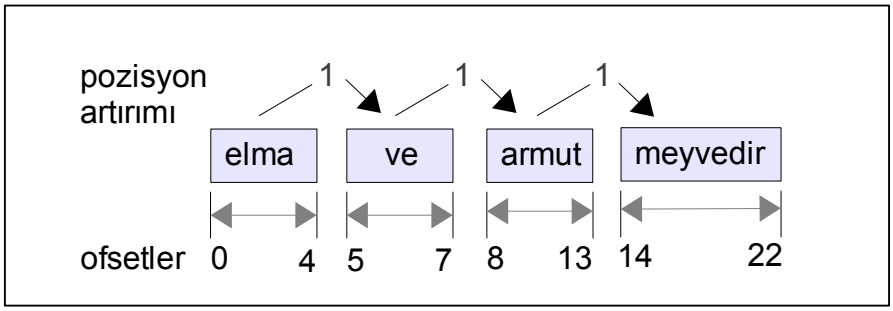
Şekil 12. Lucene'in Uygulamalara Entegrasyonu
(Kaynak: Gospodnetić ve Hatcher, 2005, s. 8)

3.2 . ANALİZCİLER

Analiz, dizine aktarmak amacıyla metinleri terimlerine (token) ayırma süreçlerini kapsamaktadır. Lucene ile metnin analiz edilmesinde kullanılan sınıf ise analizci (analyzer) sınıfıdır. Analizci, metin içerisinden terimleri/kelimeleri çıkarma, noktalama işaretlerini silme, karakterlerden aksanları silme, harflerin tamamını küçük harflere dönüştürme, ortak kelimeleri atma (dur kelimeleri) ve kelimeleri gövdeleme süreçlerini yürütmektedir (Gospodnetić, ve Hatcher, 2005, s. 103). Analizcilerin ürettiği terimler ise sorgu-doküman veya doküman-doküman benzerliğinin belirlenmesinde kullanılmaktadır. Bu bağlamda, analizciler hem dizinleme hem de sorgu esnasında kullanılmaktadır.

Lucene'de “Token” sınıfı “terim” anlamında kullanılmaktadır. Metnin analiz edilmesiyle meydana getirilen terimler pozisyon ve ofset (başlangıç-bitiş karakterleri) üstverilerine sahip olmaktadır. Şekil 13'te bir örneği bulunan üst

veriler, ihtiyaç halinde yakınlık-uzaklık sorgularında, deyim (phrase) sorgularında veya arama sonuçlarının gösterildiği kullanıcı arayüzlerinde sorgu terimlerinin vurgulanmasında kullanılmak üzere dizinde depolanmaktadır.



Şekil 13. Terimlerin Pozisyon ve Ofset Üstverileri

Lucene'de “TokenStream” terimi ise metnin bütününü ifade etmekte ve “Tokenizer” ve “TokenFilter” olmak üzere iki türü bulunmaktadır. Tokenizer karakter bazında işlemlerde, TokenFilter ise kelime bazında işlemlerde kullanılmaktadır. Latin harfleri için Lucene ile hazır gelen TokenStream mimarisi Şekil 14'te, sık kullanılan Tokenizer türleri Tablo 7'de ve TokenFilter türleri Tablo 8'de yer almaktadır.

TokenStream

- **Tokenizer**
  - CharTokenizer
    - LetterTokenizer
      - LowerCaseTokenizer
    - WhitespaceTokenizer
  - NgramTokenizer
  - EdgeNGramTokenizer
  - KeywordTokenizer
  - StandardTokenizer

TokenStream

- **TokenFilter**
  - CachingTokenFilter
  - CollationKeyFilter
  - DelimitedPayloadTokenFilter
  - EdgeNGramTokenFilter
  - ASCIIFoldingFilter (ISOLatin1AccentFilter)
  - LengthFilter
  - LowerCaseFilter
  - NGramTokenFilter
  - NumericPayloadTokenFilter
  - PositionFilter
  - ReverseStringFilter
  - ShingleFilter
  - SnowballFilter
  - StandardFilter
  - StopFilter
  - SynonymTokenFilter
  - TeeSinkTokenFilter
  - TokenOffsetPayloadTokenFilter
  - TypeAsPayloadTokenFilter

Şekil 14. Latin Harfleri İle Yazılmış Metinlerde Kullanılabilecek TokenStream Mimarisi

Tablo 7. Sık Kullanılan Tokenizer Türlerinin Özellikleri

Tokenizer Türü	Özellikler
KeywordTokenizer	Girilen metni tek bir terim dizisine çevirmektedir (Örn; [bilgi] [erişim] [sistemleri] ---> [bilgi erişim sistemleri]).
CharTokenizer	Karakter kontrolü yapmaktadır. Ayrıca tüm karakterleri küçük harfe de dönüştürebilmektedir. Terimlerin en fazla 255 karakter olması gerekir.
WhitespaceTokenizer	Terimlerin arasındaki boşlukları işlemektedir.
LetterTokenizer	Karakterlerin harf olup olmadığını kontrol etmektedir.
LowerCaseTokenizer	Tüm karakterleri küçük harfe dönüştürmektedir.
SinkTokenizer	Terimlerin tekrar kullanılabilmesi için liste biçimde saklanmalarını (cache) sağlamaktadır.
StandardTokenizer	Gramer temelli olarak e-posta adresleri ve URL gibi özel tipleri çıkarmada kullanılmaktadır.

Kaynak: Gospodnetić ve Hatcher, 2005, s. 110

Tablo 8. Sık Kullanılan TokenFilter Türlerinin Özellikleri

TokenFilter Türü	Özellikler
LowerCaseFilter	Terimlerde geçen harflerin tamamını küçük harfe çevirmektedir.
StopFilter	Önceden sağlanmış olan dur listesinde yer alan kelimelerin metinde geçmesi durumunda, söz konusu kelimeleri metinden çıkartmaktadır.
PorterStemFilter	Porter gövdeleme algoritması ile İngilizce kelimeleri gövdelemektedir.
TeeSinkTokenFilter	SinkTokenizer ile birlikte metni tekrar kullanılabilir terimlerine ayırmaktadır.
ASCIIFoldingFilter	Aksanlı karakterleri aksansız hale dönüştürmektedir.
CachingTokenFilter	Metinde geçen terimleri kaydedip tekrarlama özelliğine sahiptir.
LengthFilter	Metinde geçen terimlerin belirli karakter uzunluğunda olup olmadığını kontrol edip, kısa terimleri silmektedir.
TokenFilter	Giriş olarak TokenStream (metin) kabul eden TokenStream'dir.
StandardFilter	Kısaltmalardan noktaları, kelimelerden kesme işaretlerini ve kesme işaretlerinden sonra gelen karakterleri silmektedir.

Kaynak: Gospodnetić ve Hatcher, 2005, s. 110-111

Yukarıdaki Latin harfleri ile uyumlu olan analiz araçlarına ek olarak, Lucene API bünyesinde dillere özgü olan ve deneysel aşamada olan birçok “Tokenizer” ve “TokenFilter” bulunmaktadır. Bunlara ek olarak, Lucene API bünyesinde “Tokenizer” ve “TokenFilter” araçları ile geliştirilmiş, kullanıma hazır çeşitli analizciler bulunmaktadır. Tablo 9'da yer alan temel analizciler Türkçe dahil olmak üzere Latin harfleri ile yazılmış dokümanların analiz edilmesinde kullanılabilmektedir. Ayrıca, analiz işlemlerinin farklı dillerde kolaylıkla yapılabilmesi için Lucene API bünyesinde çeşitli diller (Çince, Arapça, Almanca, Fransızca vd.) için geliştirilmiş hazır analizciler de bulunmaktadır. Bu analizciler, dillere özgü gövdeleme algoritmalarına ve dur listelerine de sahiptir. Ancak, Türkçe için kullanıma hazır bir analizci bulunmamaktadır. Türkçe metin analizine yönelik olarak sadece Snowball[16] dili ile üretilmiş morfolojik analiz yapabilen bir gövdeleme algoritması mevcuttur.

Tablo 9. Analizci Türlerinin Özellikleri

Analizci	Özellikler
WhitespaceAnalyzer	Metindeki boşluklara göre terimleri ayırmaktadır.
SimpleAnalyzer	Harf olmayan karakterlere göre metni terimlerine ayırıp, terimlerde geçen tüm karakterleri küçük harfe dönüştürmektedir.
StopAnalyzer	SimpleAnalyzer özelliklerine ek olarak dur listesinde geçen sözcükleri terim listesinden silmektedir.
KeywordAnalyzer	Girilen metni tek bir terim dizisine çevirmektedir.
StandardAnalyzer	StopAnalyzer özelliklerine ek olarak, e-posta adresleri ve URL gibi özel tipleri çıkarma yeteneğine sahiptir.

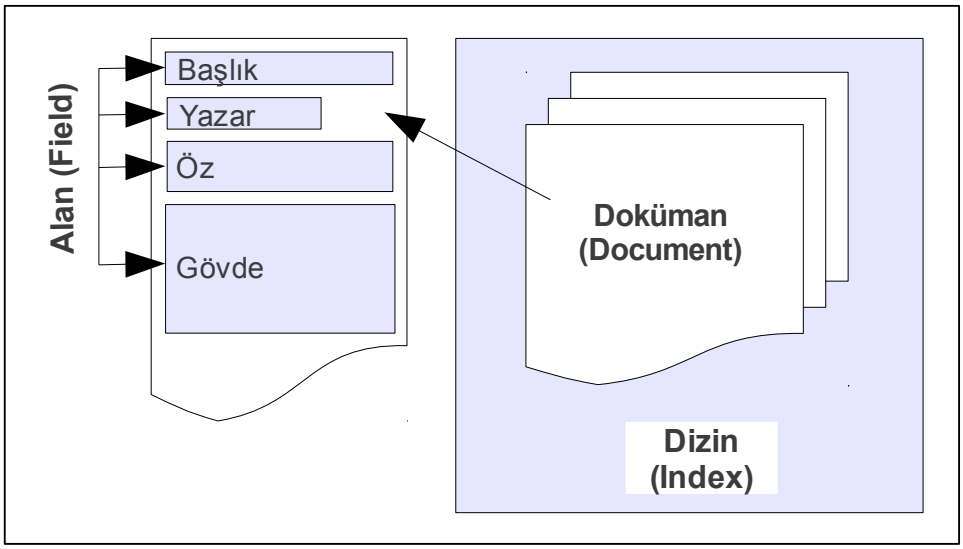
Kaynak: Gospodnetić ve Hatcher, 2005, s. 110

3.3 . DOKÜMAN VE ALAN

Doküman (Document), Lucene'in dizinlediği ve eriştirdiği her bir tekil bilgi kaynağını ifade eden sınıftır. Gerçek hayatta üretilen dokümanlar özniteliklerine göre başlık, yazar, öz ve gövde gibi biçimsel bir semantik yapıdadır. Lucene dokümanı da bu semantik yapı ile kurgulanmıştır. Dokümanı oluşturan her bir

16 Türkçe gövdeleme algoritmasına ilişkin detaylı bilgi için bakınız: Çilden, 2006

semantik birim “alan” (field) sınıfıdır. Alan(lar) dokümana, doküman(lar) ise dizine eklenmektedir. Dizin, doküman ve alanların ilişkileri Şekil 15'te yer almaktadır.



Şekil 15. Lucene'in Dizin, Doküman ve Alan İlişkisi

Kullanıcı ihtiyaçları doğrultusunda geliştirilmiş bir bilgi erişim sisteminde, erişilmek istenen dokümanların önemi birbirinden farklı olabilmektedir. Örneğin, bir doküman kümesi içerisinde Türkçe dokümanların diğer dillerdeki dokümanlardan daha önemli olduğu varsayılsın. Bu senaryoya göre kullanıcı sorgularına karşı döndürülen sonuç kümesinde Türkçe dokümanların üst sırada yer alması gerekmektedir. Lucene, bu senaryoyu gerçekleştirmek üzere ilgililiğin hesaplanmasında önem belirleme (boosting) fonksiyonuna sahiptir. Lucene, doküman dizinlemesi esnasında bu fonksiyona varsayılan değer olarak 1.0 değerini atamaktadır. 1.0 değeri doküman önemi belirlemede etkisiz bir değerdir. Dizinleme esnasında dokümanın önem değerinin 1.0'ın üzerinde atanması durumunda söz konusu dokümanın önemi diğer dokümanlara göre artmaktadır. Bu senaryonun tam tersi bir senaryo da oluşturabilmek mümkündür. Bazı dokümanların diğer dokümanlardan daha az öneme sahip olması istenildiğinde önem değeri 1.0'ın altında atanabilmektedir.

Lucene, dokümanlarda olduğu gibi dokümanı oluşturan alanlarda da önem belirleme fonksiyonuna sahiptir. Doküman öneminin artırılmadığı durumlarda

sadece belirli alanlarının önemi artırılabilmektedir. Öte yandan, dizinleme esnasında doküman öneminin artırılması durumunda, dokümanda yer alan tüm alanlara, dokümana atanmış olan önem değeri atanmaktadır. İhtiyaç dahilinde bu değerler artırılıp azaltılabilmektedir.

Diğer taraftan, alanlarda herhangi bir önem artırma değeri belirlenmese bile Lucene'in benzerlik algoritmasında uzunluk normalizasyonunu sağlayan *lengthNorm* metodu kısa alanlara yüksek önem atamaktadır. Bu özellik bazı senaryolar için avantaja dönüşebilmektedir. Örneğin, tez çalışması kapsamında üstverilerin yer aldığı kısa alanlar, hiçbir önem artırma değeri belirlenmeden tam-metin alanına göre daha önemli hale gelmiştir. Lucene'in normalizasyon metodunun kısa alanlara yüksek önem ataması bazı senaryolarda da dezavantaja dönüşebilmektedir. Örneğin, sadece tek bir alanı olan dokümanların bulunduğu bir dizinde veya dizinde birçok alanı bulunan dokümanların sadece bir alanında arama yapıldığında kısa alanlar uzun alanlara göre daha ilgili olmasalar bile sonuç kümesinde üst sırada yer alacaktır. Bu durumun dezavantaj olarak görülmesi durumunda özel normalizasyon algoritmaları geliştirilebileceği gibi, Lucene API bünyesinde bulunan SSS (Sweet Spot Similarity) algoritmasında yer alan doküman uzunluk normalizasyonu da kullanılabilmektedir. TREC 2007'de yapılan bir araştırmanın (Cohen, Amitay ve Carmel, 2007) bulguları, SSS algoritmasının Lucene'in varsayılan benzerlik algoritmasına göre daha iyi sonuç verdiğini ortaya koymaktadır.

Bilgi erişim performansını etkilemesi bakımından, Lucene dokümanını oluşturan “alan” sınıfı Lucene'in en önemli sınıflarından biridir. Alana gönderilen metnin analiz edilip edilmeyeceği, terimlerin depolanıp depolanmayacağı, terimlerin dizinlenip dizinlenmeyeceği, dizinlenecekse hangi özelliklerin göz önünde bulundurularak dizinleneceği alan üzerinde belirlenmektedir. Alan sınıfının dizinlemeye yönelik sahip olduğu özellikler Tablo 10'da, depolamaya yönelik sahip olduğu özellikler Tablo 11'de ve terim vektörlerine yönelik sahip olduğu özellikler Tablo 12'de yer almaktadır.

Tablo 10. Dizinlemeye Yönelik Olarak Alanın Sahip Olduğu Özellikler

Seçenek	Özellikler ve kullanım alanları
Index.ANALYZED	Alana gönderilen metnin analiz edilerek aranabilir ayrık terimler halinde dizinlenmesini sağlamaktadır. Bu seçenek; başlık, öz ve gövde gibi normal metin alanlarında kullanılmaktadır.
Index.NOT_ANALYZED	Alana gönderilen metnin analiz edilmeden olduğu gibi tek bir terim halinde aranabilir olmasını sağlamaktadır. Bu seçenek; anahtar kelimeler, yazar adları, telefon numaraları ve e-posta adresleri gibi “kesin eşleştirme” (exact match) aramalarını sağlamak amacıyla kullanılmaktadır.
Index.ANALYZED_NO_NORMS	index.ANALYZED seçeneğinden farklı olarak norm bilgilerinin (önem belirleme değeri, uzunluk) dizinde depolanmamasını belirtmektedir. Bu seçenek; bilgi erişim sisteminin üzerinde koştuğu donanımın bellek ve işlem gücünün az olması durumunda seçilmektedir.
Index.NOT_ANALYZED_NO_NORMS	Index.NOT_ANALYZED seçeneğinden farklı olarak norm bilgilerinin dizinde depolanmamasını belirtmektedir.
Index.NO	Alanda yer alan bilginin aranabilir olmayacağını belirtir. Bu seçenek; doküman numarası veya oluşturulmuş senaryoya göre URL gibi dolaylı olarak erişilecek veya görüntülenecek alanlarda kullanılmaktadır.

Tablo 11. Depolamaya Yönelik Olarak Alanın Sahip Olduğu Özellikler

Seçenek	Özellikler ve kullanım alanları
Store.YES	Alana gönderilen ham metnin olduğu gibi dizine depolanmasını sağlamaktadır. Bu seçenek; sorgu sonuçlarında başlık, URL veya yazar adları gibi bilgilerin kullanıcılara gösterilmek istenmesi durumunda kullanılmaktadır.
Store.NO	Alana gönderilen ham metnin dizine depolanmamasını sağlamaktadır. Bu seçenek; makalelerin gövde bölümleri gibi çok fazla bilginin yer aldığı alanlarda kullanılmaktadır.

Not: Lucene, depolama alanının yetersiz olduğu senaryolarda ham metnin dizinde depolanmasında sıkıştırma algoritması kullanımına izin vermektedir. Ancak, sıkıştırma algoritması kullanımı depolama alanı kazandırırken işlemci gücü kaybettirmektedir.

Tablo 12. Terim Vektörlerine Yönelik Olarak Alanın Sahip Olduğu Özellikler

Seçenek	Özellikler
TermVector.YES	Dokümanda geçen tekil terimlerin ve terimlerin geçiş sıklıklarının dizinde depolanmasını sağlamaktadır.
TermVector.WITH_ POSITIONS	Dokümanda geçen tekil terimlerin, terimlerin geçiş sıklıklarının ve pozisyonlarının dizinde depolanmasını sağlamaktadır.
TermVector.WITH_ OFFSETS	Dokümanda geçen tekil terimlerin, terimlerin geçiş sıklıklarının ve ofset bilgilerinin (karakter başlangıç-bitiş pozisyonları) dizinde depolanmasını sağlamaktadır.
TermVector.WITH_ POSITIONS_OFFSETS	Dokümanda geçen tekil terimlerin, terimlerin geçiş sıklıklarının, pozisyonlarının ve ofset bilgilerinin dizinde depolanmasını sağlamaktadır.
TermVector.NO	Terim vektör bilgilerinin dizinde depolanmamasını sağlamaktadır.

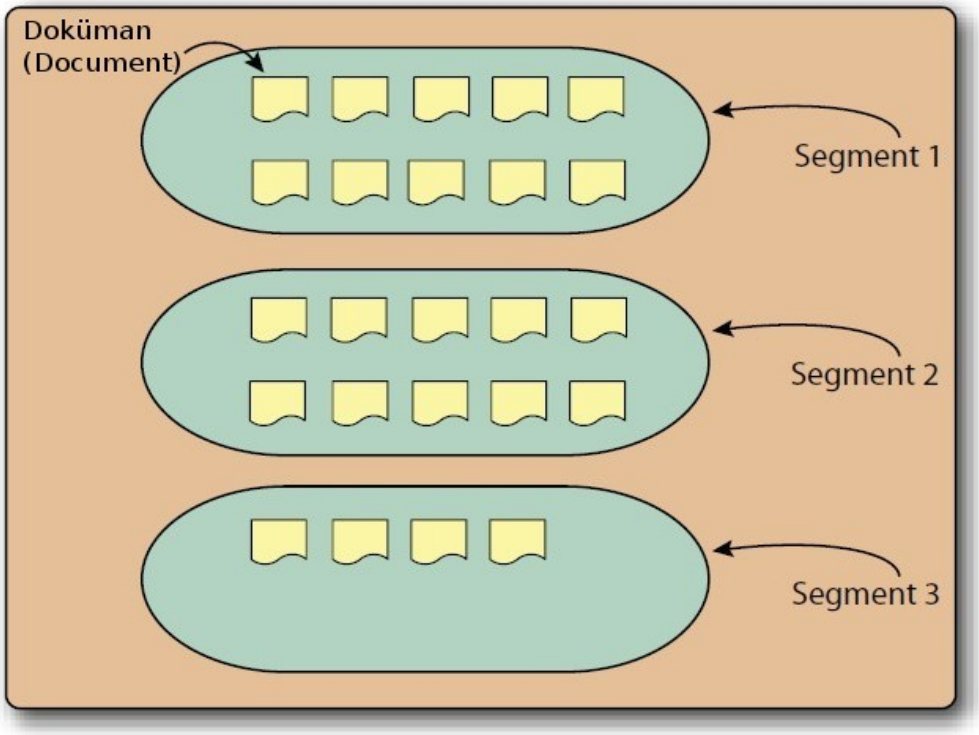
3.4 . DİZİN

Dizin (index), terimlerin üstveri ile birlikte depolandığı, güncellendiği, silindiği ve kullanıcı sorgularının yapıldığı ortamı ifade etmektedir. Lucene API bünyesinde, “FSDirectory” ve “RAMDirectory” olmak üzere iki temel dizin ortamı bulunmaktadır. Bunlardan RAMDirectory, dizin dosyalarının hızlı işlenebilmesi için RAM (Random Access Memory) üzerine kaydedilmesini sağlamaktadır. FSDirectory ise dizin dosyalarının işletim sisteminin dosya sistemi üzerine kaydedilebilmesini sağlamaktadır.

Lucene dizini, performansı en yüksek seviyede tutarken sistem kaynaklarını en az seviyede kullanabilecek bir yapıda geliştirilmiştir. Arslan ve Yılmazel'in (2008) yaptığı çalışmanın bulguları da bunu kanıtlar niteliktedir. Ayrıca, Lucene'in INQUERY bilgi erişim sistemi ile gündeme gelen “artımlı dizinleme” (incremental indexing) tekniğini kullanması, Lucene'i rakiplerine göre üstün kılmaktadır. Artımlı dizinleme öncesi bilgi erişim sistemlerinin oluşturduğu devrik dizine yeni bir dokümanın eklenmesi, silinmesi veya güncellenmesi durumunda tüm dizinin yeniden yaratılması gerekmekteydi (Brown, Callan ve Croft, 1994).

Artımlı dizinleme ile bu olumsuz özellik ortadan kaldırılmıştır. Böylelikle, hem sistem kaynaklarının en az oranda kullanılması hem de dizine yeni girmiş dokümanların anında erişilebilir olması sağlanmıştır.

Lucene dizini bir veya birden fazla segmentten oluşmaktadır. Dizine belirli aralıklarla eklenmiş dokümanlar Şekil 16'daki gibi birçok segment üretmektedir. Bu yapı dizine eklenmiş dokümanların anında erişilebilir olmasını sağlamaktadır. Öte yandan, segmentlerin ayrık olması sorgu hızını yavaşlatmaktadır. Sistemin iş yükünün az olduğu zamanlarda bu segmentleri birleştirerek dizini optimize etmek mümkündür.



Şekil 16. 3 Segmentli Optimize Edilmemiş Lucene Dizini

Kaynak: McCansless, Gospodnetić ve Hatcher, 2010, s. 436

Lucene, “çoklu dosya” (multifile) ve “bileşik” (compound) olmak üzere iki dizin yapısı kullanmaktadır. Çoklu dosya yapısında dizini oluşturan dosyalar/segmentler ayrık biçimdedir. Bileşik yapıda ise dizin dosyaları/segmetler .zip dosyalarına benzer biçimde tek bir dosyada toplanmıştır (MCCansless, Gospodnetić ve Hatcher, 2010, s. 434). Sistemin çoklu dosya dizin yapısında çok fazla sayıda açılmış segmenti işleyememesi

durumunda bileşik dizin yapısı kullanılmaktadır. Ayrıca, büyük dizinlerde bileşik dizin yapısı çoklu dosya dizin yapısına göre daha hızlı çalışmaktadır. Lucene, her iki dizin yapısını da ihtiyaç duyulduğunda birbirine çevirebilmektedir. Lucene dizininde kullanılan dosya uzantıları ve açıklamalarına ilişkin bilgi ise Tablo 13'te yer almaktadır.

Tablo 13. Lucene Dizininde Kullanılan Dosyalar, Dosya Uzantıları ve Açıklamaları

Dosya adı	Uzantı	Açıklamalar
Segment dosyası	segments.gen segments_N	Segmentler hakkında bilgi depolamaktadır.
Kilitli dosya	write.lock	Aynı dosya üzerine farklı kullanıcıların aynı anda yazmaması sağlanmaktadır.
Bileşik dosya	.cfs	Sistemin aynı anda çok sayıda dizin dosyasını işleyememesi durumunda dizin dosyalarının sanal biçimde birleştirilmiş halidir.
Alanlar	.fnm	Alanlar hakkında bilgi depolamaktadır.
Alan dizini	.fdx	Alan verileri için işaretçi (pointer) barındırmaktadır.
Alan verisi	.fdt	Dokümanlar için kaydedilmiş alan verilerini tutmaktadır.
Terim bilgileri	.tis	Terim bilgilerinin depolandığı, terim sözlüğünün bir parçasıdır.
Terim bilgi dizini	.tii	Terim bilgileri dosyasının dizinidir.
Sıklıklar	.frq	Terimlerin sıklıklarıyla birlikte geçtikleri doküman bilgilerini depolamaktadır.
Pozisyonlar	.prx	Terimlerin pozisyon bilgilerini depolamaktadır.
Normlar	.nrm	Dokümanların ve alanların, uzunluk ve önem belirleme (boosting) değerlerini depolamaktadır.
Terim vektör dizini	.tvx	.tvd ve .tvf dosyalarına ofset bilgilerini depolamaktadır.
Terim vektör dokümanları	.tvd	Her dokümanın sahip olduğu terim vektörlerinin bilgisini depolamaktadır.
Terim vektör alanları	.tvf	Alanlardaki terim vektörlerinin bilgisini depolamaktadır.
Silinmiş dokümanlar	.del	Silinmiş dokümanların bilgilerini depolamaktadır.

Tablo 13'te yer alan dizin dosyalarının tümü Lucene'in devrik dizinini oluşturmaktadır. Lucene dokümanını oluşturan alanlar henüz İVTYS'lerdeki gibi ilişkisel bir yapıda olmasa da devrik dizini oluşturan dosyalarda ilişkisel bir yapı kurulmuştur. Bu özellik Lucene'in erişim hızının en üst seviyeye çıkmasında rol oynamıştır.

3.5 . ARAMA BİLEŞENLERİ

Lucene'in yarattığı dizin yapısı ve bünyesinde barındırdığı çeşitli algoritmalar, kullanıcıların bilgi ihtiyaçlarını ifade etmelerinde zengin seçenekler sunmaktadır. Kullanıcılar arama yaparken aşağıda yer alan Lucene özelliklerinin tamamını kullanabilmektedir.

- **Terim araması:** Lucene, doğal dille ifade edilmiş bir veya birden çok sorgu terimini işleyebilmektedir. Örneğin, [elma], [elma şekeri], [“portakal ağacı”] veya [“portakal ağacı” “mandalina ağacı”]

- **Boole işleçleri ile arama:** Lucene**, “**AND”, “OR” ve “NOT” İşleçleri ile arama cümlesi oluşturmaya imkân tanımaktadır. İşleçlerin sorguda büyük harflerle yazılması gerekmektedir. Ayrıca, AND işleci yerine “&&”, OR işleci yerine “||” ve NOT işleci yerine “!” işaretini kullanabilmek mümkündür. Oluşturulan sorgu cümlelerindeki terimlerin herhangi bir boolean işleci ile birleştirilmemesi durumunda, terimler varsayılan olarak OR işleci ile birleştirilmektedir. Örneğin, [“elma şekeri” yapımı] ifadesi [“elma şekeri” OR yapımı] şekline dönüşmektedir. Ayrıca, arama cümlesinde geçen terimlerin erişilmek istenen dokümanlarda geçmesini zorunlu tutmak amacıyla “+” işaretini, geçmemesini zorunlu tutmak amacıyla da “-” işaretini terimden veya deyim terimden önce ekleyebilmek mümkündür. Örneğin, “dokümanlarda 'elma şekeri' geçsin, 'elma bahçesi' geçmesin” biçimindeki bir bilgi ihtiyacı [+“elma şekeri” -“elma bahçesi”] biçiminde ifade edilmektedir.

- **Alan araması:** Alanlara ayrılmış Lucene dokümanlarında sadece belirli

alanlarda Boole işleçlerini de sorgu cümlesine katarak arama yapabilmek mümkündür. Örneğin; "'Nihat Genç'in yazdığı 'deneme' türündeki eser(ler) getirilsin" biçimindeki bir bilgi ihtiyacı [yazar:"Nihat Genç" AND tür:deneme] biçiminde "yazar" ve "tür" alanları kullanılarak ifade edilmektedir.

- **Joker (wildcard) ile arama:** Tek karakterlik joker aramalarında "?" işareti, çok karakterli arama ise "*" işareti kullanılmaktadır. Örneğin, [s?z] ifadesi "s*ö*z", "s*a*z" veya "s*i*z" terimleriyle ilgili sonuçları döndürmektedir. [bil*] ifadesi ise "bil*gi*", "bil*gisayar*" veya "bil*im*" gibi "bil" ile başlayan terimlerin sonuçlarını döndürmektedir.

- **Bulanık (fuzzy) arama:** Lucene, Levenshtein Distance veya Edit Distance algoritmaları ile bulanık aramaya imkan tanımaktadır. Bulanık aramalarda tek bir terimden sonra "~" işareti kullanılmaktadır. Ayrıca, Lucene 1.9 sürümüyle birlikte bulanık aramalarda 0 ile 1 arasında benzerlik değerleri de atayabilmek mümkün hale gelmiştir. Değer atanmaması durumunda ise benzerlik değeri varsayılan olarak 0.5 atanmaktadır. Örneğin, [bilgi~0.8] ifadesi "bilgi"ye benzeyen "ilgi", "bilge" ve "belge" gibi "bilgi"nin yazımına yakın terimlerin sonuçlarını döndürebilmektedir.

- **Yakınlık (proximity) araması:** Birbirlerine belirli uzaklıktaki terimlerin aranmasında da deyimlerden (phrase) sonra "~" işareti kullanılmaktadır. Örneğin; "bilgi" teriminden önce veya sonra en fazla 5 terim uzaklıkta "erişim" teriminin aranması durumunda ["bilgi erişim"~5] ifadesi kullanılmalıdır.

- **Aralık (range) araması:** Küçükten büyüğe doğru belirli aralıktaki sayılar ve tarihler veya sayılar "TO" işleci ile aranabilmektedir. Örneğin; tarih alanında 2009.01.01 tarihi 2010.01.01 tarihi arasında arama yapmak için [tarih: [20090101 TO 20100101]] ifadesi kullanılmaktadır.

- **Terim önemi belirlenmiş (boosting) arama:** Lucene, dokümanın veya alanın dizinleme esnasında öneminin belirlenmesine imkân sağladığı gibi

arama esnasında kullanıcının terim önemini belirlemesine de imkân sağlamaktadır. Terim önemi belirlenmemiş bir sorguda varsayılan terim önem değeri 1.0'dır. Terim öneminin artırılmak veya azaltılmak istenmesi durumunda tek terim veya deyim terimden sonra "^" işareti ile 1.0'ın üstünde veya altında değerler atanabilmektedir. Örneğin, ["elma şekeri"^2 tatlı] ifadesi "elma şekeri" teriminin "tatlı" teriminden daha önemli olduğunu belirtmektedir. Lucene bu arama ifadesine karşılık, "tatlı" terimine göre daha fazla öneme sahip olan "elma şekeri" ile ilgili dokümanları daha üst sırada listelemektedir.

- **Gruplama araması:** Lucene, parantez işaretleriyle ve Boole işleçleri ile alt sorguların gruplanmasına imkân tanımaktadır. Örneğin, [(elma AND şeker) OR (elma AND tatlı)]
- **Alan gruplama araması:** Lucene, alanlar üzerinde de gruplama araması yapmaya imkân tanımaktadır. Örneğin, [tür:((elma AND şeker) OR "elma şekeri")]

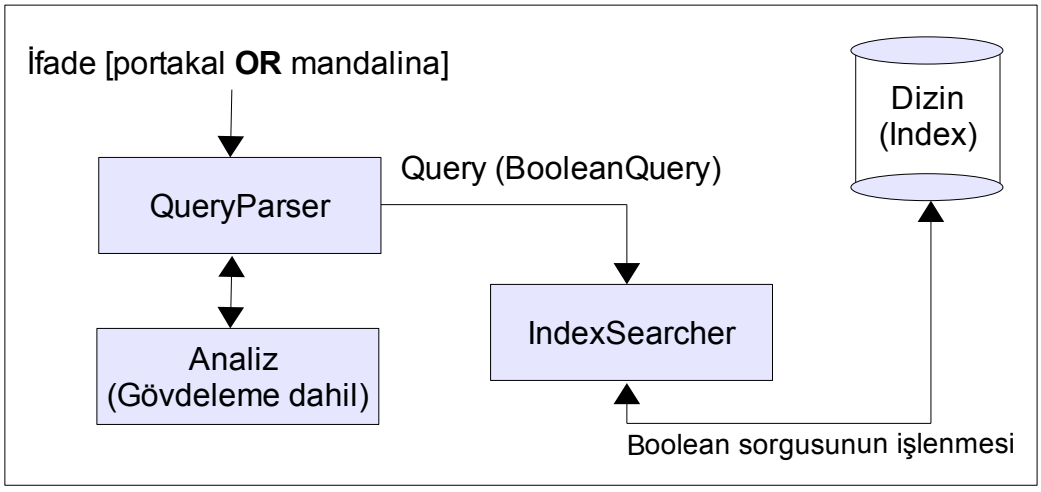
Yukarıda yer alan arama seçeneklerini gerçekleştirmek amacıyla Lucene API bünyesinde birçok sorgu sınıfı bulunmaktadır. Bu sınıflar, Tablo 14'ün Query satırında yer almaktadır.

Tablo 14. Lucene API'ın Arama Konusundaki Temel Sınıfları

Sınıf	Amaç
IndexSearcher	Dizini kullanarak aramaların ele alındığı temel sınıftır. "search" metoduna farklı türlerden sorguları gönderilerek aramanın bütünü bu sınıf üzerinden gerçekleştirilir.
Query (alt sınıflar)	Sorgu türlerinin belirlendiği soyut sınıftır. Sorgunun gerçekleştirildiği alt sınıfları şunlardır: TermQuery, MultiTermQuery, BooleanQuery, WildcardQuery, PhraseQuery, PrefixQuery, MultiPhraseQuery, FuzzyQuery, TermRangeQuery, NumericRangeQuery, SpanQuery, RegexQuery, SpanRegexQuery, MoreLikeThis, FuzzyLikeThisQuery
QueryParser	İnsanın girdiği ve okuyabildiği sorguları Query nesnelerine dönüştürmektedir.
TopDocs	IndexSearcher sınıfının yaptığı aramalara karşı erişilen dokümanları yüksek benzerlik sıralamasına göre tutmaktadır.
ScoreDoc	TopDoc sınıfındaki her bir arama sonucuna erişimi sağlamaktadır.

Tablo 14'te yer alan QueryParser sınıfı, sıkça kullanılan, önemli bir sınıftır. Bu sınıf, tek bir metin kutusu ve arama butonu olan kullanıcı arayüzlerinde kullanıcıların sorgu cümlelerini ifade etmelerinde sıkça kullanılmaktadır. Bu sınıf sayesinde, kullanıcıların girmiş oldukları arama ifadeleri/sorgu cümleleri en uygun arama sınıflarınca ele alınıp, uygun analizci ile analiz edilip işlenmek üzere IndexSearcher sınıfına gönderilmesini sağlamaktadır. Örneğin; Şekil 17'de görselleştirilmiş olan [portakal OR mandalina] sorgusu ele alındığında:

- Birinci adımda, sorgu derlenir (parsed).
- İkinci adımda, Boole işleci olan OR algılanır ve sorgu türü "BooleanQuery" olarak belirlenir.
- Üçüncü adımda, terimler gövdeleme dahil olmak üzere analiz edilir.
- Son aşamada ise sorgu için BooleanQuery sınıflarının örnekleri yaratılarak, işlenmek üzere IndexSearcher sınıfına gönderilir.

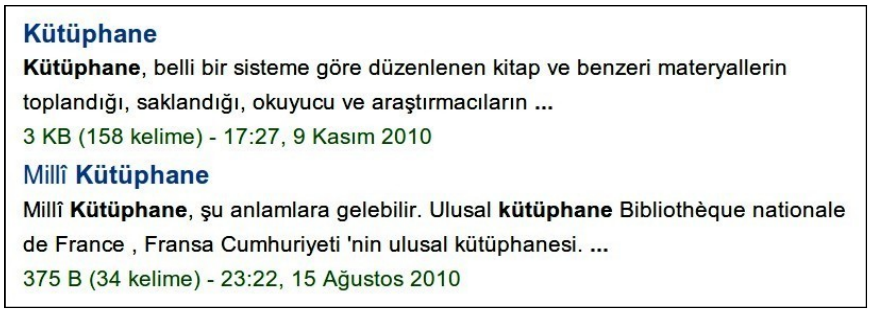


Şekil 17. QueryParser Sınıfının İşleyişi

Lucene ile oluşturulmuş bir bilgi erişim sisteminde aramalara karşı döndürülen sonuç kümesinin sıralaması varsayılan olarak ilgililiğe göredir. Ancak, ihtiyaç halinde alanlara göre artan azalan biçiminde sıralama da yaptırılabilmektedir. Ayrıca, sonuç filtreleme (örneğin, tarih aralığına göre) ve alan gruplamasıyla alan içi ilgililiğe dayalı sıralamalar da yaptırılabilmektedir.

Lucene ile arama konusunda değinilmesi gereken önemli bir sınıf da “MoreLikeThis” sınıfıdır. Bu sınıf ile kullanıcılardan alınan ilgililik geribildirimi basitçe işlenebilmektedir. Kullanım yöntemi ise sorgu sınıflarıyla aynıdır. Sorguya karşılık döndürülen sonuç kümesinden, kullanıcı ilgili bulduğu dokümanları bilgi erişim sistemine bildirmektedir. Bildirim bazen tek bir dokümanla yapılabileceği gibi (örneğin, Google'ın “Benzer” bağlantısı) oluşturulmuş senaryoya göre birden fazla doküman ile de yapılabilmektedir. Bildirim kullanıcıdan alındıktan sonra oluşturulacak senaryoya göre ilgili dokümanların belirli terimleri (örneğin minimum terim sıklığının belirlenmesi) ile yeni bir sorgu oluşturulmaktadır. Bu sorgu ile sonuç kümesinin duyarlılığının artırılacağı varsayılmaktadır.

Lucene sorgu sonuçlarının görüntülenmesinde veya kullanıcı arayüzü geliştirme konusunda da çeşitli sınıflar içermektedir. Bu sınıflar içerisinden yaygın olarak kullanılanı “Highlighter” sınıfıdır. Bu sınıf, arama terimlerinin sonuç kümesinde yer alan dokümanların ön izleme bölümlerinde vurgulanmasını sağlamaktadır. Başlık ve yazar gibi kısa alanlar için ağırlıklandırmaya ihtiyaç duymayan vurgulandırıcılar üretilebileceği gibi, metin gövdesi gibi uzun dokümanlarda ağırlıklandırma ile ilgili olabilecek genişliklerde yer alan terimler de vurgulanabilmektedir. Bu özelliğin kullanılabilmesi için terimlerin pozisyon ve ofset üstverilerinin dizine yazılmış olması gerekmektedir. Şekil 18'de Lucene ile geliştirilmiş Wikipedia bilgi erişim sisteminin vurgulama örneği yer almaktadır.



Şekil 18. Wikipedia'nın Kütüphane Terimini Vurgulama Örneği

3.6 . BENZERLİK

Lucene dokümantasyonunda ve Lucene hakkında yazılmış olan iki kitapta (Gospodnetić ve Hatcher, 2005; MCCansless, Gospodnetić ve Hatcher, 2010) Lucene'in modeli hakkında “Lucene bilgi erişim modellerinden hem BBEM'i hem de VUBEM'i kullanmaktadır” bilgisi verilmektedir. Ayrıca, Boole operatörleri ile oluşturulmuş sorgularla uyuşan dokümanların benzerliklerinin VUBEM'e göre belirlendiği bilgisi yer almaktadır.

Lucene'in temel olarak BBEM ve VUBEM kullanması nedeniyle kimi araştırmacılar Lucene'in modelinin GBBEM olduğunu, kimi araştırmacılar ise VUBEM olduğunu öne sürmüştür. Lucene API ve yaratabildiği dizin ile sadece BBEM'e, VUBEM'e ve GBBEM'e dayalı bilgi erişim sistemleri geliştirebilmek mümkündür. Ancak, Lucene'in dayandığı model, üç modelin de bazı özelliklerini kullanmaktadır. Filtreleme ve sıralama gibi özelliklerde sadece BBEM, bilgi ihtiyaçlarının temsilinde BBEM veya GBBEM, eşleşme fonksiyonunda ise önce BBEM, daha sonra VUBEM temel alınmıştır. Örneğin, [elma AND şeker] sorgusuna karşılık, öncelikle dokümanlarda “elma” ve “şeker” terimlerinin her ikisinin de yer alıp almadığı BBEM ile kontrol edilmektedir. Her iki terimin de yer aldığı dokümanların benzerlikleri VUBEM'e göre hesaplanmaktadır.

Lucene'in benzerlik formülü ve formülde yer alan bileşenler aşağıdaki biçimdedir:

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum_{t\ in\ d} (tf(t\ in\ d) \cdot idf(t)^2 \cdot tBoost \cdot norm(t,d))$$

- *score (q,d)*: *q* sorgusunun *d* dokümanına benzerliğini belirtmektedir.
- *coord(q,d)*: Sorgulama foktörlerinden biridir. Bu faktörle, sorgu terimlerinin hepsinin geçtiği bir doküman daha az sayıda sorgu teriminin geçtiği dokümana göre daha yüksek skor almaktadır.
- *queryNorm(q)*: Sorgu skorlarının karşılaştırılmasında kullanılan normalizasyondur. Öklid normalizasyonu şu şekilde sağlamaktadır:

$$queryNorm(q) = \frac{1}{\sqrt{SumOfSquaredWeights}}$$

Boolean sorgularında;

$$sumOfSquaredWeights = qBoost^2 \cdot \sum_{t\ in\ q} (idf(t) \cdot tBoost)^2$$

- *tBoost*: Sorgu esnasında kullanıcının belirlediği terimin önemini ifade etmektedir.
- *tf(t in d)*: *t* teriminin *d* dokümanındaki geçiş sıklığını ifade etmektedir. Yüksek sıklığa sahip terimler daha yüksek skor almaktadır.
- *idf(t)*: Devrik doküman sıklığını ifade etmektedir. Idf aşağıdaki biçimde hesaplanmaktadır.

$$idf(t) = 1 + \log\left(\frac{doküman\ sayısı}{doküman\ sıklığı + 1}\right)$$

- *norm(t,d)*: Önem belirleme değerleriyle birlikte alan uzunluklarını dizinleme esnasında aşağıdaki biçimde hesaplamaktadır.

$$norm(t,d) = docBoost \cdot lenghtNorm(f) \cdot \prod_{field\ f\ in\ d\ named\ as\ t} fBoost$$

  - *docBoost*: Dizinleme esnasında doküman için belirlenen önem değerini belirtmektedir.
  - *lenghtNorm(f)*: Alanın uzunluk normalizasyonunu belirtmektedir. Ayrıca, kısa dokümanlara iltimas göstermektedir.
  - fboost: Dizinleme esnasında alan için belirlenen önem değerini belirtmektedir.

4 . BÖLÜM

AÇIK ARŞİVLER

4.1 . GİRİŞ

Bu bölümde; bilginin depolanması, organizasyonu ve paylaşımı konularında ortaya çıkmış bir model olan “açık erişim” konusu ve açık arşivlerde depolanan bilgi kaynaklarının organizasyonunun ve birlikte işlerliğinin sağlanabilmesinde kullanılan yöntemler hakkında mevcut durum ortaya konulmaktadır.

Günümüz toplumları “post-kapitalist”, “post-endüstriyel” ya da “bilgi toplumu”, “enformasyon toplumu”, “sibernetik kapitalizm” veya “ağ toplumu” gibi, birbirlerinden farklı gibi görünen, ama aynı toplumsal yapıyı (durumu, özelliği) tanımlamak amacıyla kullanılan terimlerle (kuram da diyebiliriz) tanımlanmaktadır (Yamaç, 2009, s. 11). Günlük yaşamda yoğun bilgi kullanımı, üretimde kas ya da makine gücüne oranla bilginin gücünden daha fazla yararlanılması, bilginin işlenmesinde bilgi ve iletişim teknolojilerinin ağırlıklı olarak kullanılması “bilgi toplumu”nun başat özellikleri olarak ortaya çıkmaktadır (Tonta ve Küçük, 2005).

Bilgi teknolojileri başta olmak üzere diğer tüm alanlardaki teknolojilerin gelişiminde ise başrolü II. Dünya Savaşı şemsiyesi altında bilim oynamıştır. Sadece Amerika Birleşik Devletleri'nde önde gelen 6.000 bilim insanı, bilimin savaşa uygulanması konusunda çalışmalar yürütmüştür (Bush, 1945). Savaş ve uzay teknolojileri alanında çıktı vermeyi hedefleyen bu dönemin insanlığa en büyük katkılarından biri, bilgisayar ve hiper metin veya www (word wide web) öngörüsünün yapılarak, gerçekleştirimi için çaba sarf edilmesidir.

II. Dünya Savaşı ve ardından başlayan Soğuk Savaş döneminde savunma veya askeri amaçlı yapılan araştırmaların en önemli çıktılarından bir diğeri de bilimsel bilgidir. 19. yüzyıl ortalarına kadar bilimsel dergi sayısı 10 iken, 1971 yılında bilimsel dergi sayısı 70.000'e ulaşmıştır (Çelik, 1987). Bilgi üretiminde meydana gelen büyük artış literatüre “bilgi patlaması” olarak geçmiştir. Bu durum, bilginin

organizasyonunu güçleştirdiği gibi araştırmacıların literatürü takip etmekte güçlük çekmesine de neden olmuştur.

1980'li yıllara gelindiğinde ise bilimsel dergi yayıncılığında bilgi ve iletişim teknolojilerinin kullanımı başlamıştır. Bilgi ve iletişim teknolojilerinin yayıncılık alanında kullanılmasıyla ortaya çıkan "bilimsel elektronik dergiler" yeni bir yayıncılık modeli olan "elektronik yayıncılık" modelini de beraberinde getirmiştir. Yeni yayın türü ve yayıncılık modeli hem yayıncılara hem de araştırmacılara önemli avantajlar sağlamıştır. Söz konusu avantajları veya fırsatları şu başlıklar altında toplamak mümkündür (Lancaster, 1995):

- Editoryal sürecin hızlanması.
- Seçimli bilgi yayım hizmetinin verilebilmesi.
- Bilginin farklı formlarda sunulabilmesi (metin içerisine hareket, ses, hiper bağ vd. gömebilme).
- Okuyucu yorumları ile kamudan hızlıca değerlendirme veya geri bildirim alabilme.
- Düşük maliyet.
- Yayın hızı ve kolay iletişim/erişim.

Bu avantajlara ek olarak, tam metin arama ve zaman-mekân bağımlılığı olmaksızın bilgi kaynaklarına erişebilme avantajlarını da saymak mümkündür.

1983 yılında Amerikan Kimya Kurumu (American Chemical Society) paralel yayıncılığa (hem basılı hem elektronik yayıncılık) (Tonta, 1997), 1992 yılında ise *The Online Journal of Current Clinical Trials* ilk "sadece elektronik dergi" yayıncılığına başlamıştır (Küçük, Al ve Olcay, 2008). 1990'lı yıllardan itibaren bilgi teknolojilerinin ve İnternet kullanımının yaygınlık kazanması, ayrıca dergilerin elektronik sürümlerinin araştırmacılar tarafından kabul görmeye başlaması ve elektronik yayıncılığın getirdiği avantajların hem yayıncılar hem de araştırmacılar tarafından cazip görülmesi, elektronik dergi üretiminin ve kullanımının hızla yaygınlaşmasına zemin hazırlamıştır.

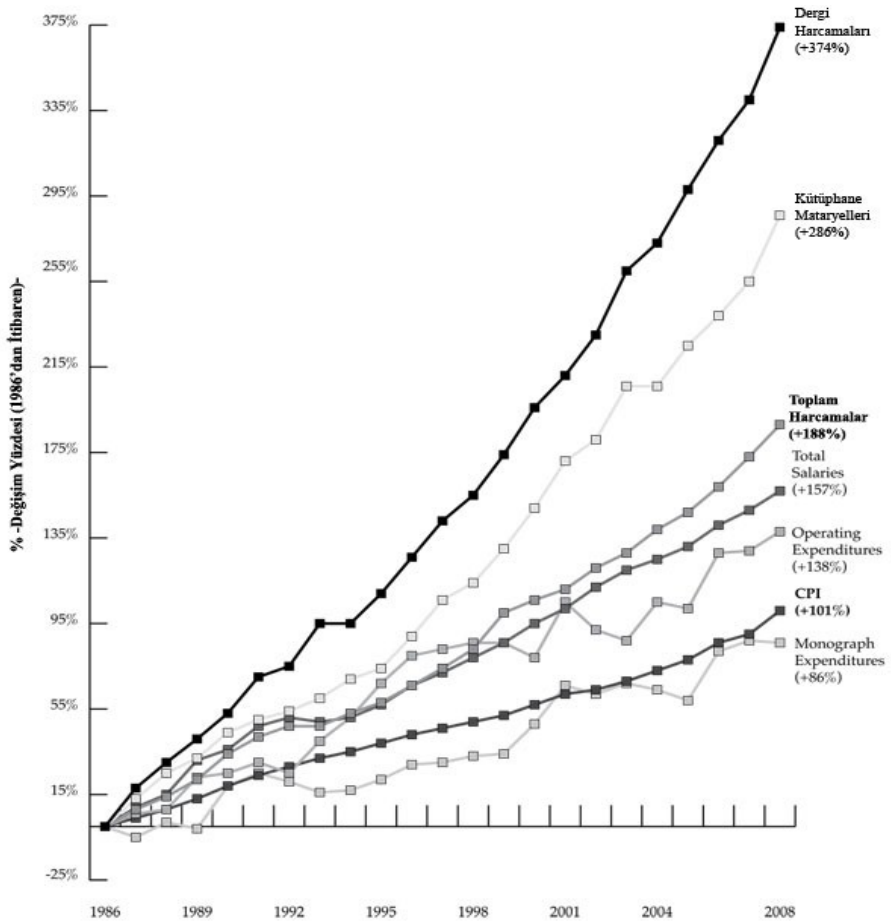
1980 ve 1990'lı yıllarda, özellikle Amerika Birleşik Devletleri'nde, elektronik yayıncılığın avantajlarını görebilen çeşitli kurumlar, üniversiteler ve yayın evleri yoğun bir biçimde dijitalleştirme faaliyetinde bulunmuştur (Karasözen, 1996; Atılgan ve Yalçın, 2009). Ayrıca, bu dönemde dijitalleştirme faaliyetlerinin yanı sıra, yayınların doğrudan elektronik ortamda üretimi de başlamıştır. ARL'nin yıllık olarak yayınladığı raporlar incelendiğinde ise, dijitalleştirme ve elektronik ortamda yayın üretme faaliyetlerinin sonucu olarak, bu dönemden sonra bilgi kaynaklarının bibliyografik kayıtlarına erişiminden ziyade, kendilerine (tam metin) erişimin önem kazandığı görülmektedir. Nitekim, ARL'nin 2009 yılında yayınladığı raporda (s. 18, 20-21) araştırma kütüphanesi bütçelerinin %51'inin elektronik kaynaklara ayrıldığı ve en büyük payı tam metin erişime imkân tanıyan elektronik kaynakların aldığı görülmektedir.

Elektronik yayıncılıkla birlikte özellikle baskı ve dağıtım gibi maliyeti artıran geleneksel yayıncılığın fiziki unsurlarının ortadan kalkması, çok küçük bütçelerle bile elde edilebilecek altyapı araçlarıyla özel veya tüzel kişilerin kolaylıkla elektronik yayıncı olabilmesine imkân tanımıştır. Öte yandan, paralel/elektronik yayın ve yayıncı sayısının zaman içerisinde artış göstermesi büyük ticari yayıncıların hoşnutsuzluğuna yol açmıştır. Büyük yayıncılar etki faktörü yüksek olan dergileri satın alarak, yayınladıkları basılı dergilerini kısmen ya da tamamen elektronik ortamda belirledikleri tek bir fiyat üzerinden pazarlayarak ya da elektronik dergi ile basılı kopyasının birlikte alınmasını zorunlu hale getirerek rekabeti ortadan kaldırmaya yönelmişlerdir (Kayaoğlu, 2006; Ertürk, 2008, s. 34). Bu durumun kütüphanelere iki önemli yansıması ise şu şekilde olmuştur:

- Büyük ticari yayıncıların yayınlarını paketler haline getirerek, belirledikleri paket satış fiyatı üzerinden pazarlaması, bilgi merkezlerini gereksinimleri dışındaki bilgi kaynaklarını da satın almaya zorlamıştır (Kayaoğlu, 2006; Ertürk, 2008, s. 34-35).
- Elektronik ortamda üretimi yapılan bilgi kaynaklarının dağıtımının İnternet aracılığı ile yapılmasının maliyeti düşüreceği, maliyetin düşmesiyle birlikte de satış fiyatlarının düşmesi öngörülürken,

öngörülerin tersine, fiyatların artması bilgi merkezlerini olumsuz yönde etkilemiştir (Tonta, 2010b).

Büyük yayıncıların fiyatlandırma ve pazarlama politikalarının sonucunda kütüphaneler bütçelerini artırmak zorunda kalmıştır. Artırılan bütçelerin 1986-2008 yılları arasında ARL üyesi kütüphanelerde yapılan harcama trendlerine yansıması ise Şekil 19'daki gibi olmuştur. 22 yıl içerisinde kütüphane materyallerine yapılan harcamaların %286 artması, özellikle de dergi harcamalarının %324 artması dikkat çekmektedir. Öte yandan, bütçeleri artan kütüphanelerin daha fazla sayıda dergiyi sağlayabilmiş olması beklenirken ya da en azından sahip oldukları dergi sayısını koruyabilmiş olmaları beklenirken, daha az sayıda dergiye abone olabilmişlerdir (Ertürk, 2008, s.36).



Şekil 19. 1986-2008 Yılları Arasında ARL Üyesi Kütüphanelerde Harcama Trendleri

(Kaynak: ARL, 2009, s. 15)

Süreli yayın fiyatlarının yüksek oranda artması, kütüphanelerin aboneliklerini yenileyememesine, yayıncıların kaybettikleri abonelerden bekledikleri gelirleri mevcut abonelerden sağlamak için fiyatları yükseltmelerine neden olmuştur (Tonta, 2006). Kısır döngüye dönüşen bu süreç literatüre "süreli yayın krizi" veya "dergi krizi" (serials crisis) olarak geçmiştir (Parks, 2002; Tonta, 2006; Kayaoğlu, 2006). Kriz olarak nitelendirilen dönemde akademik dergi fiyatları astronomik düzeyde artmıştır. Örneğin, 1998-2003 yılları arasında İngiltere'de enflasyon artışı %11 iken, akademik dergi fiyatları %58 oranında artmıştır (Steele, 2005). Tonta (2006), son otuz yıldır bilimsel ve teknik dergi fiyatlarının enflasyon oranının çok üzerinde artmasını, bir makalenin telif hakları kısıtlamalarından dolayı farklı dergilerde yayınlanamamasına ve dergi yayıncılığında yeterince rekabetin olmamasının bazı yayıncılara kolayca "tekel gücü" kazandırabilmesine dayandırmıştır.

Yayıncıların "keyfi" olarak fiyatları artırabilmelerinin sonucunda kütüphaneler sağlamakta güçlük çektikleri yayınlara daha uygun fiyatlarla erişebilmek amacıyla konsorsiyumlar oluşturmuş olsalar da, bu girişimler de sorunun çözümüne yeterince katkı sağlamamış ve bunun sonucu olarak "açık erişim" modeli ortaya atılmıştır.

4.2 . AÇIK ERİŞİM

Açık erişim (open access), "bilimsel literatürün İnternet aracılığıyla finansal, yasal ve teknik bariyerler olmaksızın, erişilebilir, okunabilir, kaydedilebilir, kopyalanabilir, yazdırılabilir, taranabilir, tam metne bağlantı verilebilir, yazılıma veri olarak aktarılabilir ve her türlü yasal amaç için kullanılabilir biçimde kamuya ücretsiz açık olması" biçiminde tanımlanmaktadır (BOAI, 2002; ANKOS, 2010a).

Açık erişim modeli bilimsel bilginin üretiminden erişimine giden süreçte meydana gelen çarpıklığın ortadan kaldırılması konusunda atılmış en somut adımdır. Bilimsel araştırmaların büyük bir kısmının kamu kaynaklarıyla finanse edilmesi, araştırma sonuçlarının kalite kontrolünde yine kamu kaynaklarıyla finanse edilmiş bilim insanlarından faydalanılarak üretimi tamamlanan bilimsel

yayınların pek de karşılık beklemeksizin, cömertçe ticari yayıncılara sunulması ve büyük ölçüde kamu kaynaklarıyla üretilen yayınların ticari yayıncılar tarafından yüksek fiyatlarla kütüphanelere satılması suretiyle kamunun erişimine açılması çarpıklığı en dikkat çeken noktadır (Tonta, Ünal ve Al, 2007; Tonta, 2010b). Açık erişim tartışmalarının ve çözüm önerilerinin referans noktasının büyük ölçüde bu çarpıklık olduğu görülmektedir.

Açık erişim konusunda farkındalığın artırılması amacıyla yapılan önemli çalışmalar ise 2000'li yıllarda düzenlenen toplantılar ve bu toplantılardan çıkan önemli bildirgelerle başlamıştır. George Soros'un kurduğu Açık Toplum Enstitüsü (Open Society Institute) tarafından 2001 yılında Macaristan'da düzenlenen bir çalıştayın sonucunda Budapeşte Açık Erişim Bildirgesi yayınlanmıştır. Bildirgede, kalite kontrolünden geçmek şartıyla, yazarların yayınlarını kendi kendine arşivlemesi (self archiving) ya da açık erişim dergilerde (open access journal) yayınlaması önerilmiştir (BOAI, 2002). Bu bildirge, açık erişimin genel hatlarını çizmesi ve diğer girişimlere öncü olması bakımında önemlidir. Nitekim, 2003 yılında Howard Hughes Tıp Enstitüsü'nde (ABD) ekosistemi (bilim insanları, yayıncılar, dernekler ve kütüphaneciler) kapsayan toplantılar, 2003 yılında Çevrimiçi Avrupa Kültür Mirası (European Culture Heritage Online) ve Max Planck Kurumu'nun Berlin'de düzenlediği çalıştay ve sonrasında IFLA gibi uluslararası kuruluşların da gerçekleştirdiği birçok faaliyet ve bu faaliyetlerden çıkan sonuçlar ve çözüm önerileri Budapeşte Açık Erişim Bildirgesi'ne paraleldir.

Öte yandan, açık erişim modelinin bilim insanlarına önerdiği; ticari yayıncılarda yayınladıkları araştırmaları kendi kendine arşivleme veya kurumsal açık arşivlerde yayınlama yaklaşımı, önlem alınmaması durumunda kişilerin veya kurumların zaman zaman yayıncılarla telif hakları bağlamında karşı karşıya gelmelerine neden olabilmektedir. Rekabete kapalı bir biçimde tekel gücüne sahip ticari yayıncıların ayrıca telif hakları sözleşmeleriyle ellerinde yasal haklar bulundurması uzlaşma zemininden rahatlıkla uzaklaşabilmelerine olanak tanıyarak açık arşivlerin önünde büyük engeller oluşturmuştur. Nitekim, PLoS'un (Public Library of Science) 2000 yılında tıp alanında sınırlı erişimli

dergilerde yayınlanan makalelerin gecikmeli olarak altı ay içerisinde PMC'de (PubMed Central) yayınlanması önerisine bilim insanlarının olumlu yaklaşmasına karşın yayınların telif haklarını ellerinde tutan yayıncılarının ticari kaygılar gütmesi, girişimin başarısızlıkla sonuçlanmasına neden olmuştur (Tonta, 2007). Bu durum, hem telif hakları konusunun hem de yayıncılarla uzlaşma yollarının bulunmasının ne derece önemli olduğunu göstermesi bakımından kayda değerdir.

Yayıncılarla uzlaşının sağlanabilmesinde kuşkusuz yayıncıların tekel gücünün kırılabilmesi önem taşımaktadır. Bu amaçla, açık erişim modelinin gündemde tutulup desteklenmesi ve ticari yayıncıların yayınladıkları etki faktörü yüksek dergilere rakip olabilecek dergilerin farklı finansman modelleriyle çok daha uygun fiyatlarla (örneğin, 1/3 oranında) satılabileceğinin ortaya konulması ticari yayıncıların açık erişime mesafeli duruşunu değiştirmesine neden olmuştur. 2007 yılı itibariyle ticari yayıncıların büyük bir kısmı (%92'si) araştırmacıların hakemli ticari dergilerde yayınlanan makalelerinin ön baskı (preprint) ya da son baskılarını (postprint) kendi web sayfaları ya da kurumsal arşivler aracılığıyla herkesin erişimine açmasına izin vermiştir (Tonta, 2007).

Rekabete dayalı biçimde yayıncılarla uzlaşı sağlamaya ek olarak telif hakları devrine ilişkin sözleşmelerin gözden geçirilmesinin ve açık erişime olanak tanıyan farklı telif anlaşmalarının düzenlenmesinin de önemli olduğu görülmektedir, çünkü Dünya Fikri Mülkiyet Örgütü'nün (World Intellectual Property Organization) uluslararası imzaya açtığı telif hakkı sözleşmeleri[1] ve bu sözleşmelere dayalı olarak yapılan ulusal yasalar her ne kadar yayınların eğitim veya araştırma amaçlı erişime açılabilmesine olanak sağlamış olsa da yayıncılar ile yazarlar arasında telif haklarının devrine ilişkin yapılan sözleşmelerin yasalardan önce gelmesi yayıncıların yasaları atlatabilmelerine olanak tanımıştır. Bu soruna getirilen çözüm önerisi ise ticari yayıncılarla açık erişim modelini destekleyen ek sözleşmelerin yapılmasıdır. SPARC (Scholarly Publishing and Academic Resources Coalition) Yazar Ek Sözleşmesi (SPARC

1 WIPO Telif Hakkı Sözleşmesi (WIPO Copyright Treaty): http://www.wipo.int/export/sites/www/treaties/en/ip/wct/pdf/trtdocs_wo033.pdf adresinden erişilebilir.

Author Addendium)[2] bu konuda en bilindik sözleşmedir.

Açık erişimli dergilerde, kişisel arşivlerde ve kurumsal açık arşivlerde yayınlanan eserlerin telif hakları ihlaline uğramadan daha geniş çevrelerce kullanılabilmesi amacıyla da Creative Commans (CC) lisansları gündeme getirilmiş ve kullanımı teşvik edilmiştir. CC'ın amacı, telif yasalarına bakışı ve oluşturduğu lisans sözleşmelerinin dayanağı aşağıdaki biçimde ifade edilmektedir:

> Bireylerin İnternet üzerinde bilgiye erişme ve bilgiden yararlanma olanaklarının, telif yasalarıyla kısıtlanmakta olduğunu dikkate alan Creative Commons; sanatçıların ve tüm diğer telif hakkı sahiplerinin, yasanın kendilerine tanıdığı kimi haklardan kamu lehine feragatını sağlayan, özel telif lisansı sözleşmelerini hazırlayarak yaygınlaşması için çaba göstermektedir. Böylelikle, bireylerin özgürce yararlanabilecekleri kamusal bilgi alanının korunması ve geliştirilmesi amaçlanmaktadır.
>
> Free Software Foundation tarafından daha önce meydana getirilen GNU General Public License (GNU GPL) metinlerini temel alan bu sözleşmelerin özelliği, yaratıcı kişilerin, telif haklarından tamamen feragat etmeksizin, eserlerini serbestçe paylaşıma açmalarına ve böylelikle diğer sanatçılarca da kullanılabilmesine imkân tanımasıdır (*Creative Commons*, 2010).

Açık erişim modeliyle örtüşen CC yaklaşımı, CC lisanslarının açık arşivlerde sıkça kullanılmasına neden olmuştur. Eser sahipleri, Tablo 15'te yer alan dört koşulu temel alan Tablo 16'daki altı lisanstan birini eserlerine CC lisansı olarak atayabilmektedir. Tüm lisansların ortak noktası ise eserlerin kopyalanabilir, dağıtılabilir ve görüntülenebilir olmasıdır.

2 SPARC Yazar Ek Sözleşmesi: http://www.arl.org/sparc/bm~doc/Access-Reuse_Addendum.pdf adresinden erişilebilir.

Tablo 15. Temel CC Koşulları

Koşul Adı	Logo	Özellikler
by: Atıf Gerekli (Attribution)		Eser kopyalanabilir, dağıtılabilir, görüntülenebilir ve üzerinde değişiklik yapılıp yenisi üretilebilir (Atıf yapmak şartıyla).
sa: Aynen Paylaşım Şartlı İşleme (Share Alike)		Eserin kopyalarında ya da eserden üretilmiş eserlerde de aynı lisans kullanılmalıdır.
nc: Gayri Ticari (Non-Commercial)		Ticari olmamak koşuluyla eser kopyalanabilir, dağıtılabilir, görüntülenebilir ve üzerinde değişiklik yapılıp yenisi üretilebilir.
nd: İşlemeye Kapalı (No Derivatives)		Eser üzerinde değişiklik yapılmadan kopyalanabilir, dağıtılabilir ve görüntülenebilir fakat eserden eser üretilemez.

Kaynak: *Creative*, 2010

Tablo 16. CC Lisans Türleri ve Açıklamaları

Lisans Adı	Lisans Logosu	Lisans Açıklaması
by		**İzinler:** Eser kopyalanabilir, üzerinde değişiklik yapılıp yenisi üretilebilir ve ticari amaçla kullanılabilir. **Şart:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi.
by-sa		**İzinler:** Eser kopyalanabilir, üzerinde değişiklik yapılıp yenisi üretilebilir ve ticari amaçla kullanılabilir. **Şartlar:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi ve eserin kopyalarında veya eserden üretilmiş yeni eserlerde de aynı lisansın kullanılması.
by-nc		**İzinler:** Eser kopyalanabilir ve üzerinde değişiklik yapılıp yenisi üretilebilir. **Şartlar:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi ve eserin hiçbir kopyasının veya eserden üretilmiş yeni eserlerin hiçbirinin ticari amaçla kullanılmaması.
by-nd		**İzinler:** Eser kopyalanabilir, ve ticari amaçla kullanılabilir. **Şartlar:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi ve eserin orijinalliğinin korunması.
by-nc-nd		**İzinler:** Eser kopyalanabilir, üzerinde değişiklik yapılıp yenisi üretilebilir. **Şartlar:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi, eserin hiçbir kopyasının ticari amaçla kullanılmaması ve eserin orijinalliğinin korunması.
by-nc-sa		**İzinler:** Eser kopyalanabilir, üzerinde değişiklik yapılıp yenisi üretilebilir. **Şartlar:** Eserin tüm kopyalarında orijinal sahibinin belirtilmesi, eserin hiçbir kopyasının veya eserden üretilmiş yeni eserlerin hiçbirinin ticari amaçla kullanılmaması ve eserin kopyalarında veya eserden üretilmiş yeni eserlerde de aynı lisansın kullanılması.

Kaynak: *Creative*, 2010

Açık erişim modeline sorun teşkil eden lisans sözleşmeleri, yayıncıların açık erişime yaklaşımları ve bilim insanlarının bilinçlendirilmesi konularında 10 yıl gibi kısa bir süre içerisinde büyük ilerleme sağlanmıştır. Ayrıca, bilimsel ve teknik yayınlara erişimin önündeki bariyerleri kaldırılması düşüncesiyle ortaya atılan bu model çok kısa bir zaman içerisinde tüm bilgi kaynaklarına erişimin önündeki bariyerlerin kaldırılması konusunda önemli çalışmalara da öncülük etmiştir. Avrupa Birliği ülkelerinde başlatılan "Europeana" adlı proje ile imaj, metin, ses ve video gibi farklı formlarda bilgi taşıyan 6 milyon dijital kaynağa erişim fırsatı yaratılmıştır (*Europeana*, 2010). Ekim 2010 tarihinde açık erişimli dergileri ve kurumsal arşivleri listeleyen ve bu ortamlarda yer alan yayınlara erişimi sağlayan önemli servisler incelendiğinde ise şu sonuçlara ulaşılmıştır:

- ROAR'ın (Registry of Open Access Repositories)[3] dünya çapında listelediği 1642 adet açık erişim arşivinin 1127'si kurumsal açık arşivdir (geri kalan arşivler ise elektronik dergi, tez, eğitim öğretim materyalleri, makaleler ve diğerleridir). ROAR'da kayıtlı olan Türkiye adresli 15 (11 adet kurumsal açık arşiv, 4 adet elektronik dergi) adet açık arşiv bulunmaktadır.

- DOAJ (Directory of Open Access Journals)[4] ise 5514 dergi ve bu dergilerde yayınlanmış 460590 makaleye erişim sağlamaktadır. 1134 adet dergi ile Amerika Birleşik Devletleri DOAJ'da birinci sırada yer alırken, 172 dergi ile yedinci sırada yer alan Kanada'dan sonra 151 dergi ile Türkiye sekizinci sırada yer almaktadır.

- Open J-Gate[5] ise 4511'i hakemli olmak üzere 7566 adet dergiye erişim sağlamaktadır.

Yukarıdaki verilerden yola çıkarak açık erişim modelinin bilimsel ve teknik bilgide olduğu kadar diğer alanlarda da bilgiye erişimin karşısında duran bariyerlerin kaldırılması konusunda ilerleme kat ettiği görülmektedir. Türkiye ile ilgili veriler göz önünde bulundurulduğunda ise açık erişimin sadece elektronik

3 ROAR web sitesi: http://roar.eprints.org/
4 DOAJ web sitesi: http://www.doaj.org/
5 Open J-Gate web sitesi: http://www.openj-gate.com/

dergi yayıncılığında ilerleme sağlanabildiği fakat kurumsal açık arşivler konusunda henüz yeterli düzeyde yol alınamadığı görülmektedir.

4.3 . AÇIK ARŞİVLER GİRİŞİMİ VE BİRLİKTE İŞLERLİK

Açık erişim modelinin, uygulama alanında işlevlerini sağlayabilmesi için yayınların nasıl organize edilip eriştirileceği hakkında uygun standartların ve protokollerin geliştirilmesi gerekmektedir. Bu amaçla, 1990 yılında Santa Fe'de yapılan bir toplantının (Universal Preprint Service) ardından Açık Erişimler Girişimi (OAI) kurulmuş (Van de Sompel ve Lagoze, 2000) ve OAI, standartların ve protokollerin geliştirilmesinde en aktif organizasyon olmuştur. Andrew W. Mellon Foundation, Coalition for Networked Information, Digital Library Federation ve National Science Foundation gibi önemli kuruluşlar OAI için destek sağlamaktadır.

OAI, standart ve protokollerin seçimi veya geliştirilmesi için 1999 yılında çalışmalarına başlamış, referans model (Reference Model for an Open Archival Information System - OAIS) ise 2002 yılında CCSDS (Consultative Committee for Space Data Systems) ve ISO tarafından oluşturulmuştur. OAIS, bilginin üretiminden başlamak üzere; üstveri oluşturma, birlikte işlerlik mekanizmalarının bütününü, göç ve koruma gibi sistemi ilgilendiren tüm konuları ele almaktadır (OAIS, 2002).

Açık arşivlerin yapılanış modeli olarak belirlenen kendi kendine arşivleme, açık erişimli dergilerde yayın yapma ve yayınların kurumsal açık arşivler aracılığı ile erişime sunulması modelleri, açık arşivlerin dağıtık bir yapıda olmasını gerekli kılmıştır. İstisnasız tüm dağıtık yapılarda öne çıkan en önemli konu ise birlikte işlerliğin sağlanmasıdır. Bunun için kayıtların belirli standartlara göre hazırlanması ve kayıtların iletiminde standartlara dayanan çeşitli protokollerin kullanılması veya geliştirilmesi gerekmektedir. Örnek olarak, kütüphanecilik alanında bibliyografik kayıtların MARC üstveri standardına göre hazırlanması ve MARC kayıtlarının değişiminde istemci ve sunucunun Z39.50 protokolünü kullanması verilebilir.

OAI, 2001 yılında birlikte işlerlik konularının bütününü kapsayan OAI Üstveri Harmanlama Protokolü'nün (OAI-PMH) 1.1 sürümünü yayınlamış, ancak 2002 yılında OAIS'in piyasaya sürülmesi ile kesin referans modelin belirlenmesi OAI-PMH 2.0 sürümünün çıkartılmasına neden olmuştur.

OAI-PMH şartnamesinin[6] (specification) 2.0 versiyonu incelendiğinde en dikkat çeken nokta, protokolün mümkün olabildiği ölçüde kolay uygulanabilir ve esnek bir yapıda oluşturulmuş olmasıdır. Şartnamenin birçok alanında “zorunlu” (must) ifadeleri yer almamaktadır. OAI-PMH, kütüphanecilik alanında birlikte işlerliği sağlamaya yönelik geliştirilmiş olan Z39.50 protokolüyle karşılaştırıldığında uygulaması oldukça basit bir protokoldür. Ancak, OAI-PMH'nin basit bir protokol olması fonksiyonlarının da kısıtlı olduğunu göstermektedir.

OAI-PMH, genel olarak veri sağlayıcı (data provider) ve servis sağlayıcı (service provider) arasındaki birlikte işlerliği sağlamak amacıyla yapılandırılmıştır. Veri sağlayıcı, üstverilerini sağlamaktan sorumludur. Servis sağlayıcı ise veri sağlayıcısından verileri toplayıp erişimi sağlama gibi katma değerli hizmetleri sağlamaktan sorumludur.

OAI-PMH iletişim kurmak için özel bir port ve iletişim protokolü kullanmamaktadır. OAI-PMH, iletişim protokolü olarak http (Hypertext Transfer Protocol) kullanmakta, port için ise herhangi bir sınırlama getirmemektedir. Veri sağlayıcı, talep (request) için http'nin “GET” veya “POST” metotlarını/fiillerini kullanmaktadır.

OAI-PMH, üstveri standardı olarak DC kullanımını zorunlu tutmaktadır. Diğer üstveri şemalarının kullanımı ise isteğe bağlıdır. Kullanıcılar ihtiyaç halinde kendi oluşturdukları üstveri şemalarını da kullanabilmektedir. Ancak, kullanılacak olan üstveri kayıtlarının metin biçiminde XML[7] olarak

6 Şartname, http://www.openarchives.org/OAI/openarchivesprotocol.html adresinden erişilebilmektedir. Ayrıca çalışmanın OAI-PMH bölümünde kaynak gösterilemeyen ifadeler bu şartnameye dayandırılmıştır.

7 XML (eXtensible Markup Language – Genişleyebilir İşaretleme Dili): Bilginin yapılandırılması, depolanması ve taşınması için yaratılan standart bir işaretleme dilidir. XML'i HTML'den (Hyper Text Markup Language – Hiper Metin İşaretleme Dili) ayıran en önemli özellik işaretlerin (tag) önceden tanımlı olmamasıdır. Kullanıcı, ihtiyaç duyduğu etiketleri kendi tanımlamak zorundadır. (Kaynak: w3schools, 2010)

yapılandırılması ve doğrulama/onaylama için XML şemalarının oluşturulması zorunlu tutulmaktadır.

OAI-PHM, arşivde yer alan bilgi kaynaklarının harmanlanması için tanımlanmış ve XML biçiminde kodlanmış veri kümesine “kayıt” (record) adını vermektedir. Şekil 20'de bir örneği olan standart kayıt 3 bölümden oluşmaktadır:

1. **Başlık (header):** Üstveri şemasından bağımsız olarak harmanlamaya yardımcı olan bilgileri vermektedir. Bu bölümde, kaydın tanımlanmasında kullanılan tekil kimlikleyici, tarih bilgisi (kaydın oluşturulma, değiştirilme veya silinme tarihi) ve seçimli harmanlama için kaydın bağlı olduğu küme (konu veya kurum olarak. Örneğin, matematik, edebiyat, TKD, Gazi Üniversitesi) bilgileri yer almaktadır.
2. **Üstveri (metadata):** Veri sağlayıcı DC üstveri standardını sağlamak zorundadır. Diğer üstveri şemaları seçime bağlıdır.
3. **Hakkında (about):** Zorunlu olmayan ve tekrarlanabilir alanlardır. Kaynak/köken ve haklar (rights) olmak üzere iki kolondan oluşmaktadır. Telif bildirimi yapmak için haklar kolonu sıklıkla kullanılmaktadır.

```
<record>
<header>
        <identifier>oai:tk.kutuphaneci.org.tr:article/2061</identifier>
```

```
        <datestamp>2008-08-31T21:00:00Z</datestamp>
        <setSpec>tk:HY</setSpec>
</header>
<metadata>
<oai_dc:dcxsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
        <dc:title xml:lang="tr_TR">
        Karma Kütüphane: Dijital ve Geleneksel Kütüphanelerin Odak Noktası
        </dc:title>
        <dc:creator>Mehdi Afzali</dc:creator>
        <dc:subject xml:lang="tr_TR">
        Karma kütüphane; dijital kütüphane; geleneksel kütüphane
        </dc:subject>
        <dc:description xml:lang="tr_TR">
        Günümüzde kütüphaneler geleneksel basılı dermelerden hem basılı hem de dijital kaynakları içeren
        “karma kütüphaneler”e dönüşmektedir...
        </dc:description>
        <dc:publisher xml:lang="tr_TR">
        Türk Kütüphaneciler Derneği / Turkish Librarians' Association
        </dc:publisher>
        <dc:date>2008-09-01</dc:date>
        <dc:type xml:lang="tr_TR">Hakemli Yazılar</dc:type>
        <dc:format>application/pdf</dc:format>
        <dc:identifier>
        http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2061
        </dc:identifier>
        <dc:source xml:lang="tr_TR">
        Türk Kütüphaneciliği; Cilt 22, Sayı 3 (2008): Türk Kütüphaneciliği; 266-278
        </dc:source>
        <dc:language/>
</oai_dc:dc>

<about>
        <dc xmlns="http://www.openarchives.org/OAI/dc.xsd">
        <rights>
        Yazarlar, ilk yayın hakkı ...
        </rights>
        </dc>
</about>
</metadata>
</record>
```

Şekil 20. Örnek Bir OAI-PHM Kaydı

Önceden de belirtildiği gibi OAI-PMH geliştirilirken uygulanabilme kolaylığı göz önünde bulundurulmuştur. Bu yüzden seçimli harmanlama imkânı da kısıtlı düzeyde yapılabilmektedir. Üstverilerin harmanlanmasında sadece *tarih* (*from* ve *until* parametreleriyle; eklenme, güncellenme ve silinme tarihi) ve *küme* (konu ve kurum kümeleri) bazında seçim yapılabilmektedir.

OAI-PMH'de harmanlama yapmaya yönelik kullanılabilecek toplam 6 adet talep/fiil vardır. Bunlar:

1. **GetRecord:** Arşivdeki üstveri kayıtları içerisinden sadece birine erişmek için kullanılan fiildir. Fiil kullanılırken tekil kimlikleyici ve talep edilen

üstveri türünün belirtilmesi gerekmektedir.

2. **Identify:** Arşiv hakkındaki bilgilere ulaşmada kullanılan fiildir. Aşağıdaki bilgilerin bir kısmını döndürmektedir.

    - Aşağıdaki elementlerden en az birinin döndürülmesi zorunludur.
        - repositoryName : İnsanın okuyabileceği biçimde arşivin adı.
        - baseURL : Arşivin temel URL'i.
        - protocolVersion : Arşivin desteklediği OAI-PMH versiyonu.
        - earliestDatestamp : Arşivde meydana gelen değişim, silinme veya güncelleme tarihlerinin en küçüğünü döndürür. Arşiv kayıtlarında bu tarihten daha küçük bir tarih olmamalıdır.
        - deletedRecord : Silinmiş kayıtlar.
        - granularity: ISO9601'e göre arşivin desteklediği tarih formatı (Örneğin, YYYY-MM-DD veya YYYY-MM-DDThh:mm:ssZ).
    - Aşağıdaki elementin döndürülmesi zorunludur.
        - AdminEmail : Arşiv yöneticisinin e-posta adresi.
    - Aşağıdaki elementlerin döndürülmesi isteğe bağlıdır.
        - compression: Web sunucusunun sıkıştırılmış cevap gönderebilme durumu.
        - description: Arşivin tanımı.
3. **ListIdentifiers**: Arşivdeki üstveri kayıtlarını kimliklendiren tekil kimlikleyicileri döndürmektedir. Silinmiş kayıtların belirlenmesinde sıkça kullanılmaktadır.

4. **ListMetadataFormats:** Arşivin desteklediği üstveri şemasını/şemalarını listelemektedir.

5. **ListRecords:** Arşivdeki üstveri kayıtlarını döndürmektedir. MetadataPrefix (üstveri şeması) argümanının kullanımı zorunludur. İsteğe bağlı olarak *until* ve *from* tarih seçim argümanları ve konu-kurum

seçimine olanak tanıyan *set* argümanı kullanılabilmektedir.

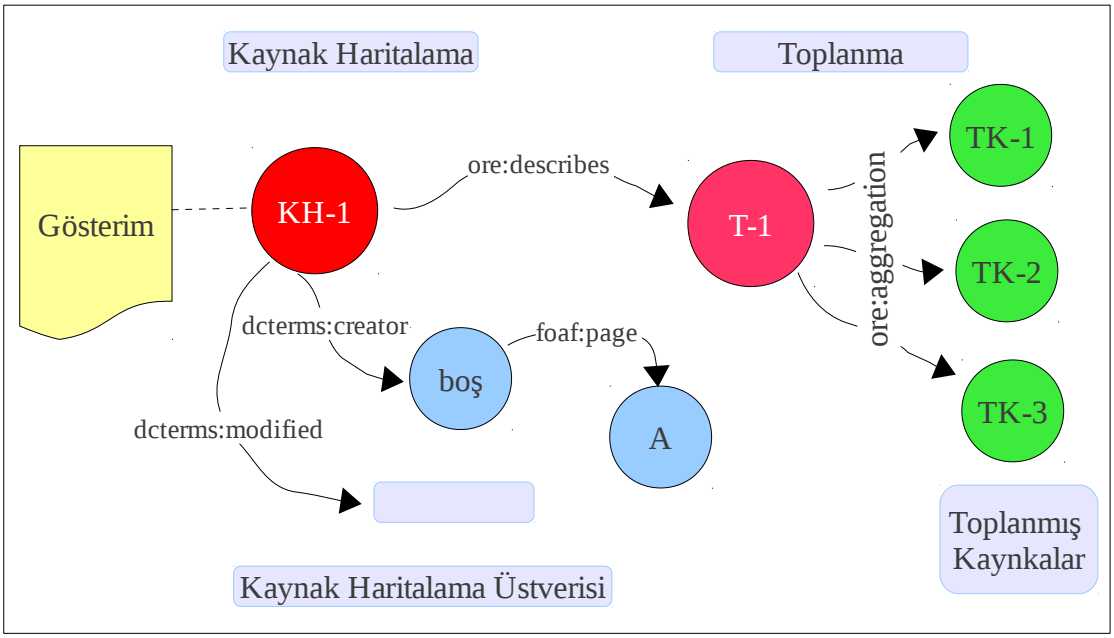
6. **ListSets:** Arşivdeki *kümeleri* (set) (konular veya kurumlar) listelemektedir. Seçimli harmanlamada kullanılmaktadır.

Açık erişim modelinde birlikte işlerliği sağlamak üzere kullanılan OAI-PMH'in yetenekleri değerlendirildiğinde ise önemli eksikliklerinin olduğu görülmektedir. Açık erişim tanımında “yasal ve teknik bariyerlerin kaldırılması” konusuna ve bilgi kaynaklarının “kaydedilebilir, kopyalanabilir ve yazılıma veri olarak aktarılabilir“ olma özelliklerine vurgu yapılmaktadır. Ayrıca, açık erişimin yasal altyapısı için önerilen CC lisansları ile de bu özellikler yeniden vurgulanmıştır. Ancak, OAI-PMH şartnamesinde bilgi kaynaklarına doğrudan erişimin şart koşulmamış olması, OAI-PMH'in söz konusu açık erişim özelliklerini tam olarak destekleyememesine neden olmuştur. Açık kaynak kodlu ve ücretsiz olan tüm açık erişim yazılımlarının, DC elementlerinden “identifier” elementine ham kaynak URL'ini atamaması, ham bilgi kaynaklarının OAI-PMH aracılığı ile kopyalanamamasına neden olmaktadır. Bu durum, bilgi kaynaklarının uzun süreli korunabilmesine yönelik programların geliştirilememesine neden olduğu gibi açık arşivleri tam-metin dizinleyebilecek bilgi erişim sistemlerinin tasarlanamamasına da neden olmaktadır.

Açık arşiv yazılımlarını üreten topluluklar, OAI-PMH'in tekil ve ham bilgi kaynaklarını işlemekte yetersiz kalmasından ötürü farklı farklı çözüm yolları üretmiştir. OAI de 2008 yılının sonunda standardı belirlemek üzere “Nesne Yeniden Kullanımı ve Değişimi” (Object Reuse and Exchange - ORE) standardının 1.0 sürümünü yayınlamıştır. OAI-ORE, temel olarak toplu web kaynaklarının tanımlanmasında ve değişiminde standartları belirlemektedir (OAI, 2010).

OAI-ORE, OAI-PMH ile kıyaslandığında OAI-ORE'nin semantik WEB mimarisine daha uygun olduğu görülmektedir. Semantik Web'in veri modeli ise XML formatındaki RDF (Resource Description Framework) ile sağlanmaktadır. RDF içerisinde DC elementleri ile kaynağın tanımlayıcı bilgileri verilmektedir. Ayrıca, kaynağın ilişkili olduğu diğer kaynaklar da RDF ile haritalanmaktadır.

Şekil 21 OAI-ORE'nin işleyişini örneklendirmektedir.



Şekil 21. OAI-ORE Örneği: Kaynak Haritalama KH-1 ile Tanımlanmış Toplanma T-1'in Üç Kaynağı Toplaması (Kaynak: OAI, 2008)

OAI-ORE kaynak haritalandırma ise üç farklı biçimde gerçekleştirilmektedir. Gerçekleştirme türlerinden ATOM ve RDF; XML biçiminde, RDFa ise insanın okuyabileceği xHTML biçiminde gerçekleştirilmektedir. Şekil 22'de OAI-ORE'nin ATOM biçimindeki örneği bulunmaktadır.

Kaynak haritalarının keşfedilmesi ve harmanlanmasına yönelik olarak sadece Atom ve RDF kullanılabileceği gibi OAI-PMH aracılığı ile harmanlanmak üzere OAI-ORE de çağrılabilmektedir. Açık arşiv yazılımının OAI-ORE'yi desteklemesi durumunda, AOI-PMH'in “metadataPrefix” fiiline üstveri türü olarak “oai_rem_atom” parametresinin atanmasıyla Şekil 23'te yer alan örnekte olduğu gibi AOI-PMH'den OAI-ORE'ye bağ kurulabilmektedir.

```
<?xml version="1.0" encoding="UTF-8"?>
<atom:entry
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:atom="http://www.w3.org/2005/Atom"
   xmlns:ore="http://www.openarchives.org/ore/terms/"
   xmlns:foaf="http://xmlns.com/foaf/0.1/">
   <atom:id>tag:arxiv.org,2008:astro-ph:0601007</atom:id>
   <!-- ArXiv dokümanlarının toplanması -->
   <atom:link href="http://arxiv.org/aggregation/astro-ph/0601007"
      rel="http://www.openarchives.org/ore/terms/describes"/>
            <atom:category term="http://www.openarchives.org/ore/terms/Aggregation"
            scheme="http://www.openarchives.org/ore/terms/" label="Aggregation"/>
   <!-- Kaynaklar ... -->
   <atom:link href="http://arxiv.org/abs/astro-ph/0601007"
            rel="http://www.openarchives.org/ore/terms/aggregates"
            title="Makale başlığı" type="text/html" />
   <atom:link href="http://arxiv.org/ps/astro-ph/0601007"
      rel="http://www.openarchives.org/ore/terms/aggregates"
           title="Makale başlığı" type="application/postscript"
      hreflang="en"/>
   <atom:link href="http://arxiv.org/pdf/astro-ph/0601007"
      rel="http://www.openarchives.org/ore/terms/aggregates"
            title="TMakale başlığı " type="application/pdf"
      hreflang="en"/>
   <!-- Üstveri: başlık ve yazarlar -->
         <atom:title> Makale başlığı</atom:title>
   <atom:author>
      <atom:name>Yazar adı</atom:name>
      <atom:email>yazar@domain</atom:email>
   </atom:author>
   <atom:author>
      <atom:name>İkinci yazar adı</atom:name>
   </atom:author>
   <atom:author>
      <atom:name>Üçüncü yazar adı</atom:name>
   </atom:author>
   <!-- Kaynak haritalandırma -->
   <atom:link href="http://arxiv.org/rem/atom/astro-ph/0601007"
      rel="self"
      type="application/atom+xml"/>
   <!-- Kaynak haritalandırma üstverisi: tarih, haklar ve yazar-->
   <atom:updated>2008-10-03T07:30:34Z</atom:updated>
   <atom:published>2008-10-01T18:30:02Z</atom:published>
      <atom:rights>Bu kaynak haritası CC ile lisanslanmıştır</atom:rights>
      <atom:link href="http://creativecommons.org/licenses/by-nc/2.5/rdf" rel="license"
      type="application/rdf+xml"/>
   <atom:source>
      <atom:author>
         <atom:name>arXiv.org e-Print Repository</atom:name>
         <atom:uri>http://arxiv.org</atom:uri>
      </atom:author>
   </atom:source>
   <!-- Son kullanıcının okumaya başlayacağı sayfa-->
   <atom:link href="http://arxiv.org/abs/astro-ph/0601007" rel="alternate"/>
</atom:entry>
```

Şekil 22. OAI-ORE ATOM Uygulama Örneği

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2007-02-08T08:55:46Z</responseDate>
  <request verb="GetRecord" identifier="oai:domain:nesne1"
         metadataPrefix="oai_rem_atom">http://domain/oai2</request>
  <GetRecord>
   <record>
    <header>
     <identifier>oai:domain:nesne1</identifier>
     <datestamp>2007-01-06</datestamp>
    </header>
    <metadata>
     <!-- Nesne 1 için kaynak haritası -->
     <entry xmlns="http://www.w3.org/2005/Atom"
          xsi:schemaLocation="http://www.w3.org/2005/Atom
          http://www.openarchives.org/OAI/2.0/atom_entry.xsd">
      <title>Nesne 1 için kaynak haritalama verisi</title>
      ...
     </entry>
    </metadata>
   </record>
  </GetRecord>
</OAI-PMH>
```

Şekil 23. “oai_rem_atom” Fili ile AOI-PMH'den OAI-ORE'ye Bağ Kurma Örneği

Sonuç olarak, açık arşivlerde yaygın olarak kullanılan protokol OAI-PMH'dir. Ayrıca, açık kaynak kodlu ücretsiz açık arşiv yazılımlarının tamamı OAI-PMH'i desteklemektedir. OAI-ORE ise kullanımı zorunlu olmayan ve farkındalığı daha düşük bir standarttır. Açık erişimli bilgi kaynaklarının uzun süreli korunması için ve bilgi kaynakları için katma değerli servislerin üretilebilmesi için OAI-ORE standardı hakkında farkındalığın artırılması gerekmektedir.

4.4 . AÇIK ARŞİV YAZILIMLARI

Açık arşiv yazılımlarını, kurumsal açık arşiv yazılımları ve açık dergi yazılımı olarak ikiye ayırabilmek mümkündür. Kurumsal açık arşiv yazılımları da ücretli ve ücretsiz olarak ikiye ayrılmaktadır. Eprints[8], Dspace[9], Fedora[10] ve Greenstone[11] yaygın olarak kullanılan ücretsiz yazılımlardır. Bepress,

8 Eprints web sitesi: http://www.eprints.org/
9 Dspace web sitesi: http://www.dspace.org/
10 Fedora web sitesi: http://fedora-commons.org/
11 Greenstone web sitesi: http://www.greenstone.org/

CONTENTdm, Open Repository ve Digital Commons yazılımları ise ücretlidir. Open Journal Systems (OJS)[12] ise açık dergi yönetim sistemi olarak açık kaynak kodlu ve ücretsiz bir yazılımdır.

Eprints ve Dspace en yaygın kullanılan ücretsiz kurumsal açık arşiv yazılımıdır. Yazılımlar hakkında ayrıntılı bilgi aşağıda yer almaktadır.

- **Eprints:** Southampton Üniversitesinde geliştirilmiştir.
  - Perl programlama dili ile geliştirilmiştir
  - MySQL veri tabanın üzerine inşa edilmiştir.
  - Yazılım, GNU (General Public License) ile lisanslanmıştır.
  - Rol tabanlıdır ve iş akış özelliği bulunmaktadır. Örneğin, editör rolündeki bir kullanıcı iş akışı içerisinde yazardan düzeltme isteyebilmektedir. Yayına hazır eserler editör tarafından dizgici rolüne sahip kişilere yönlendirilebilmektedir.
  - OAI-PMH'i desteklemektedir.
  - Tam-metin arama özelliği bulunmaktadır.
  - RSS desteği bulunmaktadır.
  - Yazılımın indirilebileceği adres: http://software.eprints.org
- **DSpace:** MIT ve HP tarafından geliştirilmiştir.
  - Java programlama dili ile geliştirilmiştir.
  - PostgreSQL veya diğer SQL standardını destekleyen diğer VTYS'ler üzerine inşa edilebilmektedir.
  - BSD (Berkeley Software Distribution) ile lisanslanmıştır.
  - Eprints'de olduğu gibi rol tabanlıdır ve iş akış özelliği bulunmaktadır.
  - OAI-PMH ve AOI-ORE desteği bulunmaktadır.
  - Lucene ile tam metin arama özelliğine sahiptir.
  - RSS (Really Simple Syndication) ve OAI-ORE Atom desteği bulunmaktadır.
  - Yazılımın indirilebileceği adres: http://sourceforge.net/projects/dspace

12 OJS web sitesi: http://pkp.sfu.ca/ojs/

- Web grubu adresi: http://dspace.org

Eprints ve Dspace yazılımları farklı türlerdeki bilgi kaynaklarının (tez, kitap, makale, ders notu vd.) arşivlenmesi amacıyla kullanılabilmektedir. Ancak, OJS sadece bilimsel dergiler için kurgulanmış bir yazılımdır. Tez kapsamında derme olarak kullanılmak üzere üretilen Türk Kütüphaneciliği Açık Arşivi de OJS ile gerçekleştirilmiştir. Aşağıda OJS hakkında ayrıntılı bilgi bulunmaktadır.

- Public Knowledge Project altında geliştirilmiştir.
- PHP programlama dili ile geliştirilmiştir.
- MySQL, PostgreSQL veya Oracle VTYS üzerine inşa edilebilmektedir.
- GNU ile lisanslanmıştır.
- Rol tabanlıdır ve iş akış özelliği bulunmaktadır.
- AOI-PHM desteği bulunmaktadır.
- BBEM'e dayanan tam-metin arama özelliği vardır.
- RSS ve Atom desteği bulunmaktadır.
- Dergilerin geriye dönük sayılarının dizinlenebilmesi için tanımlanmış bir XML şeması bulunmaktadır.
- Kullanıcı arayüzlerinin tasarımında Smarty tema motoru kullanmaktadır.
- Yazılımın indirilebileceği adres: http://sourceforge.net/projects/dspace

5. BÖLÜM

BULGULAR

5.1. GİRİŞ

Bu araştırmada, açık arşivler için geliştirilebilecek bilgi erişim sistemlerinin performanslarının ortaya konulması amaçlanmıştır. Bu amaçla, açık arşivlerin özellikleri göz önünde bulundurularak üç farklı bilgi erişim sistemi geliştirilmiştir. Bunlardan birincisi; sadece insana dayalı olarak üretilmiş üstverinin kullanıldığı bilgi erişim sistemi, ikincisi; sadece tam-metnin kullanıldığı otomatik dizinleme yapan bilgi erişim sistemi, üçüncüsü ise, hem insana dayalı olarak üretilmiş üstverinin kullanıldığı hem de tam-metnin kullanıldığı bilgi erişim sistemidir.

Araştırmada, üstveriye dayalı bilgi erişim sistemi; “üstveri bilgi erişim sistemi” (ÜBES), tam-metne dayalı bilgi erişim sistemi; “tam-metin bilgi erişim sistemi” (TBES) ve hem üstveriye hem de tam-metne dayalı bilgi erişim sistemi; “karma bilgi erişim sistemi” (KBES) olarak adlandırılmıştır.

Söz konusu bilgi erişim sistemlerinin; erişebildikleri ilgili doküman sayısı, erişim kümesi içerisinde yer alan ilgisiz doküman sayısı ve erişilen ilgili dokümanların erişilen ilgisiz dokümanların önünde yer alıp almaması performans değerlendirme kriterlerini oluşturmuştur. Bu doğrultuda, araştırma kapsamına giren üç bilgi erişim sisteminin dizinleme tekniğinin bilgi erişim performansına etkisini ortaya koyabilmek amacıyla; tüm bilgi erişim sistemleri Lucene ile tasarlanıp, anma-duyarlılık ve normalize sıralama sonuç performansları değerlendirilmiştir.

Araştırma kapsamında geliştirilen bilgi erişim sistemleri, dermenin özellikleri, soruların ilgililiği, soruların formülasyonu, performans değerlendirme ölçütleri ve performans değerlendirme sonuçları alt başlıklarda ayrıntılı olarak işlenmiştir. Ayrıca, araştırmanın daha iyi anlaşılabilmesi için bulguların ortaya çıkmasını sağlayan hazırlık aşaması çalışmaları hakkında da bilgi verilmiştir.

5.2. HAZIRLIK AŞAMASI

Hazırlık aşaması, bilgi erişim sistemlerinin performanslarının değerlendirebilmesi için gerçekleştirilen ön çalışmaları kapsamaktadır. Dolayısıyla bu bölüm, işlenmek üzere dermenin elde edilişini (dizinlenecek olan metin tabanlı PDF dokümanlarının ve dokümanlara ait üstverinin elde edilişi), metin analizini ve tasarlanan bilgi erişim sistemlerinin özelliklerini kapsamaktadır.

Hazırlık aşamasında ve performans değerlendirme süreçlerinde; işletim sistemi olarak Linux dağıtımlarından CentOS-5, geliştirme ortamı olarak NetBeans IDE 6.9, programlama dili olarak Java (JDK 6), İVTYS olarak PostgreSQL 8.4 ve bilgi erişim API'ı olarak Lucene 3.02 kullanılmıştır. Sonuçların istatistiksel değerlendirmesinde SPSS 11.5 kullanılmıştır.

5.2.1. Dermenin Elde Edilişi

Bilgi erişim sistemleri tasarlanmadan önce, ilk etapta, açık arşivde yer alan tüm üstveri 1952 yılından başlamak üzere yıllara göre OAI-PMH aracılığı ile talep edilmiştir. *Türk Kütüphaneciliği* açık arşivinin kullandığı OJS 2.2.4 sürümü OAI-ORE standardını desteklememektedir. Bu sebepten ötürü, açık arşivde yer alan metin tabanlı PDF dokümanlarına DC setinin “identifier” elementiyle doğrudan ulaşılamamıştır. OJS'nin kaynak kodu incelenerek söz konusu dokümanlara ulaşmanın sistematik bir yolu bulunmuştur. Daha sonra, OAI-PMH ile döndürülen sonuçlar içerisinde her bir dokümana ait “identifier” elementine metin tabanlı PDF dokümanının URL'i eklenmiştir. Sonraki aşamada, üstveri DC elementlerine göre ayrıştırılıp (parse), PostgreSQL veri tabanında hazırlanmış olan bir tabloya kaydedilmiştir. En son aşamada, PDF dokümanlarında yer alan metin Tika aracılığı ile çıkartılarak, işlenip veri tabanına kaydedilmiştir.

Metnin işlenmesi aşamasında ise, satır sonundaki uygun heceden kesilmiş kelimeler ele alınmıştır. Satır sonunda yer alan kelimelerin uygun heceden kesilip alt satırdan yazımının devam etmesi durumunda, Tika'nın PDF

dosyalarını işlemede kullandığı PDFBox, söz konusu kelimedeki kesme işaretini silip heceleri birleştirmemektedir. Bunun sonucunda, analiz aşamasında söz konusu kelimeler kaybedilmektedir. Bu sorunun üstesinden gelebilmek üzere satır sonu ve sayfa sonu göz önünde bulundurularak, kesme işaretiyle ayrılmış hecelerin birleştirilmesinin Türkçe'ye uygun olup olmadığı Zemberek[1] ile kontrol ettirilip, birleştirme veya ayrıştırma (örneğin, tarihlerde) yapılmıştır.

Sonuç olarak, tasarlanacak üç bilgi erişim sisteminde dizinlenmek üzere ihtiyaç duyulan üstveri ve üstveri ile ilişkili tam-metin bilgi yerel bilgisayara depolanmıştır.

5.2.2. Metin Analizi

Bilgi erişim sistemlerinin gerçekleştirimindeki önemli noktalardan biri de metin analizidir. Metin analizinde dizine girecek olan terimler işlenmektedir. Terimlerin işlenme sürecinde ise, terimlerin gövdelenip gövdelenmeyeceği, deyim olarak alınıp alınmayacağı, en az kaç karakter uzunluğunda olacağı, tarihlerin veya rakamların dizine alınıp alınmayacağı belirlenmektedir.

Türkçe terimlerin gövdelenmesinin bilgi erişim performansına etkisi konusunda yapılmış araştırmalar; dermenin büyüklüğüne, sorgularda yer alan terimlere ve gövdeleme algoritmalarının yeteneklerine göre farklı sonuçlar vermektedir. Sezer'in (1999, s. 65) 92 adet tez özü gibi küçük bir derme üzerinde Duran'ın (1997) geliştirdiği Gövdebul algoritmasıyla yaptığı araştırmada gövdelemenin performansı yaklaşık %20 iyileştirdiği, Eroğlu'nun (2000, s. 88-89) aynı gövdeleme algoritmasını kullanarak 2468 doküman ile yaptığı araştırmada ise gövdelemenin erişim performansını %25 iyileştirdiği ortaya konulmuştur. Ancak, her iki araştırmada da gövdelemeli ve gövdelemesiz dizinleme ile elde edilen duyarlılık değerlerinin arasında istatistiksel olarak anlamlı bir farkın olup olmadığı ortaya koyulmamıştır. Can ve diğerlerinin (2008) 408.305 adet dokümana sahip TREC benzeri bir derme üzerinde 72 adet doğal dil sorgusuyla

1 Zemberek, Türk dilleri (Türkiye Türkçesi, Azeri Türkçesi vd.) için geliştirilmiş bir doğal dil işleme kütüphanesidir. OpenOffice için yazım kontrolü yapmak amacıyla geliştirilmiştir. Ayrıca, sözlüğe dayalı morfolojik analiz yapabilen gövdeleme algoritmasına da sahiptir. Ayrıntılı bilgi için, http://code.google.com/p/zemberek/ adresi ziyaret edilebilir.

yaptığı araştırmada ise, gövdelemeli dizinlemenin gövdelemesiz dizinlemeye göre istatistiksel olarak anlamlı bir fark yarattığı ortaya konulmuştur. Aynı araştırmada, hem sözlük kullanan hem de morfolojik analiz yapan gövdeleme algoritmasının en iyi performansı sergilediği de tespit edilmiştir. Bu sonuçlardan yola çıkarak, gövdelemenin sağlayabileceği performans iyileştirmelerinden faydalanmak üzere analiz aşamasında gövdeleme algoritmasının kullanımına karar verilmiştir.

Araştırmada tasarlanan bilgi erişim sistemlerinde, analiz aşamasında, alanlarda yer alan terimler deyim olarak dizinlenmemiştir (örneğin, deyim olarak “İrfan Çakın” biçiminde değil, ayrık olarak “İrfan” ve “Çakın” biçiminde alınmıştır). Bunun sebebi, terimlerin deyim olarak dizinlenmesi durumunda, deyimi oluşturan terimlerden biri ile yapılan sorgularda herhangi bir sonucun döndürülmemesidir. Ayrıca, dizine girecek olan terimlerin minimum veya maksimum karakter uzunlukları dikkate alınmamıştır. Son olarak, dizine rakamlar ve tarihler de alınmıştır.

Bu doğrultuda, üç bilgi erişim sisteminin de ihtiyaç duyduğu “Türkçe analizci” geliştirilmiştir. Türkçe analizcide kullanılmak üzere; “Türkçe küçük harf filtresi”, “Türkçe gövdeleme filtresi” ve “Türkçe dur filtresi” geliştirilmiştir. Türkçe küçük harf filtresinin geliştirilmesinin nedeni, Lucene API ile gelen LowerCaseFilter'ın “I” ve “i” harflerini İngilizce'deki gibi algılamasıdır (örneğin, “Ihlamur” ifadesini “ihlamur” ifadesine çevirmesi). Türkçe gövdeleme filtresinde kullanılmak üzere Türkçe gövdeleme algoritması seçiminde sözlüğe dayalı birkaç seçenek (Gövdebul (Duran, 1997) ve Zemberek (Akın ve Akın, 2010)) olsa da Lucene API'nin sunduğu olanakların dışına çıkılmadan, sadece morfolojik analiz yapan Snowball gövdeleme algoritması kullanılmıştır. Dur filtresinde kullanılmak üzere de dermede en sık geçen terimlerle birlikte, erişim açısından anlamlı olmayan; bir, ve, veya, için, ama, fakat, bu, şu, lakin, ki, de, da, ile, mı, mi, mu, mü, of, the, is, are ve and terimleri seçilmiştir.

5.2.3. Tasarlanan Bilgi Erişim Sistemlerinin Özellikleri

ÜBES, TBES ve KBES'in tasarımında hem BBEM'in hem de VUBEM'in özelliklerinden faydalanan Lucene API 3.02 kullanılmıştır. Üç bilgi erişim sistemi de ortak olarak Lucene'in varsayılan bilgi erişim modelini kullanmaktadır. Dizinleme ve sorgulama aşamasında ise bir üst bölümde özelliklerine değinilen "Türkçe Analizci" kullanılmıştır.

Tasarlanan bilgi erişim sistemlerinden ÜBES, DC elementlerinden "title", "author", "description", "subject" ve "type" elementleri ile oluşturulmuştur. Söz konusu elementler Lucene dokümanını oluşturan alanlara eşleştirilerek dizinlenmiştir[2]. TBES'de, dokümanın sadece tam-metnini barındıran "fulltext" alanı dizinlenmiştir. KBES ise DC elementlerinden "title", "author", "description", "subject" ve "type" elementleri ile birlikte, belgenin tam-metnini barındıran "fulltext" ile oluşturulup Lucene alanlarına eşleştirilerek dizinlenmiştir. Ayrıca, KBES'de yer alan DC alanlarında geçen terimlerin birçoğu tam-metin alanında da yer almaktadır. Yani, "title", "author", "subject" ve "decription" alanlarında yer alan terimler büyük ölçüde "fulltext" alanından elde edilmiş ve söz konusu terimler "fulltext" alanından silinmemiştir.

Sorgu kısmında ise, kullanıcının alan belirterek yaptığı sorgularda sadece kullanıcının belirttiği alanlarda arama yapılmıştır. Alan belirtmeden yapılan sorgularda ise, sorgu ifadesi OR işleci ile bağlanarak tüm alanlarda arama yapılmıştır. Böylece, olası tüm çakışmaların yakalanabilmesi ve üstveri alanlarındaki terimlerle sorgu ifadesinin çakışması durumunda doküman uzunluk normalizasyonunun dezavantajından faydalanılarak kısa olan üstveri alanlarının yüksek puan/skor alması beklenmiştir. Dolayısıyla, üstveri alanlarına özel olarak önem belirleme değeri veya ağırlıklandırma değeri atanmamıştır.

---

2 Dizinleme, terim vektörlerinin oluşturulması anlamında kullanılmıştır. Ayrıca, DC elementlerinden "tarih" elementi gibi diğer elementler de dizine depolanmıştır fakat terim vektörü oluşturulmamıştır. Terim vektörü oluşturulmamış alanlarda herhangi bir arama yapılmadığı için bu elementler araştırmada listelenmemiştir.

5.3. *TÜRK KÜTÜPHANECİLİĞİ* AÇIK ARŞİVİ

Türk Kütüphaneciliği (TK), Türk Kütüphaneciler Derneği'nin (TKD) resmi yayın organıdır. 1952 yılında *Türk Kütüphaneciler Derneği Bülteni* (TKDB) adıyla yayın hayatına başlamış, 1987 yılında *Türk Kütüphaneciliği* adını almış, 1995 yılında da hakemli statüde yayınlanmaya başlamıştır.

TK'nin içeriğini, özgün bilimsel (hakemli) yazılar, görüşler, okuyucu mektupları, tanıtım yazıları, çeviriler, haberler ve duyurular oluşturmaktadır. 1995 yılından itibaren hakemli yazılar bölümünde ve daha sonra görüşler bölümündeki makalelerde öz ve anahtar sözcükler de yer almıştır.

TKDB/TK, 2002 yılında Gürdal ve Ertürk (2002) tarafından imaj tabanlı PDF dosyaları biçiminde sayısallaştırılarak, TKD'nin web sitesi üzerinden erişime açılmıştır. Söz konusu çalışma belirli bir süre sonra TKD tarafından yaşatılamadığı için kaybedilmiştir.

Tez çalışması kapsamında kullanılmak üzere uygun bir açık arşiv bulunamadığı için TKD'nin web sitesi üzerinden yıllara ve sayılara göre erişilebilen TK'ya açık arşiv standartlarından OAI-PMH'i destekleyen OJS ile bir açık arşiv ve dergi yönetim sistemi oluşturulmasına ve gerçekleştirilecek çalışmanın tez kapsamında kullanılmasına karar verilmiştir.

Bu doğrultuda, TKD web sitesinde yer alan yaklaşık 2000 makale/doküman yıl, cilt ve sayı bilgilerine göre indirilmiş, kirli PDF dosyaları ayıklanıp temiz biçimde yeniden taranmıştır. Söz konusu dokümanlar Abbyy FineReader eğitilip, optik karakter tanımadan (OCR) geçirilerek imaj tabanlı formattan metin tabanlı formata dönüştürülmüştür. Ayrıca, OCR hatalarına karşı dokümanlar insan denetimden geçirilip, metin kalitesi artırılmıştır.

1952 yılından 2010 yılının 2. sayısı dahil olmak üzere toplam 2215 doküman DC elementlerine göre Türkçe ve İngilizce[3] olmak üzere dizinlenmiştir. DC alanlarından “subject” alanına, makalede/dokümanda geçen anahtar sözcükler atanmış, herhangi bir denetimli dil kullanılmamıştır.

3 Başlık, öz ve anahtar kelimeler İngilizce olarak makalede yer almışsa İngilizce olarak da dizinlenmiştir.

5.4. TEST DERMESİ

Test dermesi, Türk Kütüphaneciliği Açık Arşivi'ni oluşturan, 1952 yılının 1. sayısadan 2010 yılının 2. sayısını kapsayan toplam 2215 adet dokümandan oluşmaktadır. Bilgi erişim sistemlerinde dizinlenen alanlar ve alanların karakter uzunlukları Tablo 17'de, dermede en sık geçen 25 terim ise Tablo 18'de yer almaktadır.

Tablo 17'nin “Min” sütununda yer alan “Öz” satırlarının 0 olmasının nedeni; 1886 dokümanın (dokümanların %85'i) öz bölümüne sahip olmamasıdır. “Konu” satırının 0 olmasının nedeni; 22 dokümanın İngilizce olmasından ötürü Türkçe konu başlığı verilmemesidir. “Tam-metin” satırının 0 olmasının nedeni ise, 5 dokümanın imajdan oluşmasıdır (metin barındırmaması).

Tablo 17. Alanlarda Yer Alan Karakter Uzunlukları

Alan	Min.	Max.	Ortalama	Ortanca	Toplam
Başlık	5	269	46,2	41	102342
Yazar	3	108	14,2	14	31542
Öz	0	2356	109,3	0	242261
Konu (Anah. Kel.)	0	260	31,8	28	70549
Tam-metin	0	294930	18793,3	13050	41605165

Tablo 18. Dermede En Sık Geçen 25 Terim

Sıra	Terim	Sıklık	Sıra	Terim	Sıklık
1	ve	156321	14	çok	15891
2	olarak	97673	15	hizmet	15506
3	bir	92524	16	halk	15409
4	bu	72819	17	yer	15346
5	kütüphane	61213	18	the	14973
6	için	41196	19	kütüphanecilik	14829
7	yıl	31693	20	türk	13965
8	il	31238	21	eser	13901
9	de	31019	22	daha	13815
10	bilgi	29798	23	genel	13401
11	da	28590	24	üzere	12799
12	kitap	27519	25	of	12631
13	çalışma	21728			

Tablo 18'de yer alan terimler ve terim sıklıkları ise, gövdeleme algoritmasının hatalarından arındırılarak elde edilmiştir. Terimleri ve geçiş sıklıklarını olabildiğince doğru verebilmek amacıyla, gövdeleme ile elde edilen tüm terimler ve geçiş sıklıkları listelenmiştir. Ayrıca, OCR ve yazım hataları değerlendirmeye alınmadan gövdeleme algoritmasının gövdelemede başarısız olduğu terimler ile başarılı olduğu terimlerin geçiş sıklıkları toplanarak Tablo 18 elde edilmiştir.

Tabloda yer alan “il” terimi hem isim olan “il” (city) hem de bağlaç olan “ile” (with) olarak değerlendirilmelidir. Çünkü, gövdeleme algoritması “ile” bağlacının son karakteri olan “e” karakterini “il” isminin “-e hali” olarak algılamaktadır. Bunun sonucunda çekim eki olarak algıladığı “e” karakterini silerek “ile” bağlacını isim olan “il” gövdesine/köküne indirmektedir. Ayrıca, “il” terimine benzer biçimde, “yıl” terimi de hem isim olarak “yıl” (year) hem de fiil (yıl-mak) olarak değerlendirilmelidir.

Tablo 18'deki dermede en sık geçen 25 terim içerisine 2 adet İngilizce terimin (*the* ve *of*) girebilmesi de dikkat çekicidir. Dermede yer alan 22 adet İngilizce dokümana ek olarak, dokümanların bir kısmında başlık, öz, özet ve kaynakça bölümlerinde İngilizce terimler geçmektedir. İngilizce terimler tüm dermede sınırlı sayıda geçmesine rağmen en sık geçen 25 terim içerisine 2 adet İngilizce terim girebilmiştir.

5.5. TEST SORULARININ SEÇİMİ, FORMÜLASYONU, İLGİLİLİK VE NORMALİZE SIRALAMA BULGULARI

Bilgi erişim sistemlerinin performanslarının değerlendirilmesinde kullanıcıların bilgi ihtiyaçlarını temsil edebilecek sorguların oluşturulması amacıyla çeşitli sorular seçilmektedir. Soruların seçiminde bilgi erişim sisteminin hangi özelliklerinin değerlendirilmek istendiği önem taşımaktadır (Saracevic, 1995). Bu bağlamda, araştırma kapsamında seçilen test sorularının üç bilgi erişim sisteminin temel özelliklerinin bilgi erişim performansına etkisini ölçebilecek nitelikte olmasına dikkat edilmiştir. İnsana dayalı, otomatik ve karma dizinlemenin birbirine göre avantajını ortaya koyabilmek; gövdeleme

algoritmasının ve kullanıcıların yapabileceği muhtemel sorguların özelliklerinin değerlendirmesini sağlayabilmek amacıyla aşağıda yer alan 9 soru seçilmiştir.

1. İrfan Çakın'ın yazdığı tüm dokümanlar.
2. İrfan Çakın'a bilimsel/hakemli dokümanlarda yapılan atıflar.
3. “Bilgi arama davranışı”
4. AACR, AACR1 veya AACR2.
5. OPAC veya “çevrimiçi katalog”.
6. Engelliler veya özürlüler.
7. Engelli veya özürlü.
8. “Kullanıcılara eğitimler”, “okuyuculara eğitimler” veya oryantasyonlar.
9. “Kullanıcı eğitimi”, “okuyucu eğitimi” veya oryantasyon.

Ayrıca, test sorularının seçiminde objektif davranılmış, herhangi bir bilgi erişim sistemini öne çıkarabilecek sorulardan kaçınılmıştır. Örneğin, birinci ve ikinci sorularda en fazla ve en az dokümanın yazarlığından sorumlu bir yazar seçmek yerine, orta değerde dokümanın yazarlığından sorumlu olan bir yazar tercih edilmiştir. Çok sayıda dokümanın yazarlığından sorumlu bir yazarın seçilmesi durumunda, birinci soruda TBES, ikinci soruda ÜBES dezavantajlı duruma düşerek, KBES değerlerinin artmasına neden olunacağı düşüncesiyle araştırmada ortalama sayıda yayını olan bir yazar tercih edilmiştir.

Ayrıca, soruların seçiminde gerçek kullanıcıların *Türk Kütüphaneciliği* açık arşivi üzerinde yaptığı aramalar da değerlendirmeye alınmıştır. Örneğin, dördüncü ve beşinci soruda geçen AACR, AARC2 ve OPAC terimleri sırasıyla *Türk Kütüphaneciliği* açık arşivinde en sık arama yapılan terimlerdir (söz konusu aramalar, tek bir terimle yapılmıştır).

Sorulara karşı döndürülen sonuç kümesindeki dokümanlar değerlendirilirken “ilgili” ve “ilgisiz” olarak ikili (binary) değerlendirmeye tabi tutulmaktadır. İlgililik kararı ise kullanıcıdan kullanıcıya değişebilecek bir kriterdir. Örneğin, sorgu sonucu döndürülen doküman kümesi içerisinden bazı dokümanlar kullanıcı tarafından daha önceden erişilip kullanılmışsa, kullanıcı söz konusu dokümanlara sorgusuyla ilgili olsa bile önceden erişip kullandığı için “ilgisiz” kararı verebilmektedir. Bu tür ilgililik kararları öznel sayılmaktadır. Performans değerlendirmelerinde bu tür yaklaşımlardan kaçınılarak, ilgililik kararı nesnel olarak verilmektedir (Soydal, 2000, s. 45). Bu doğrultuda, araştırma kapsamında kullanılan tüm dokümanların ilgililik kriterleri sorularla birlikte açıkça ifade edilmiştir.

Araştırmada kullanılan performans değerlendirme tekniği (anma-duyarlılık) ile sonuç kümesi çeşitli aralıklarda (örneğin; ilk 5, 10, 25 veya 50 sonuç gibi) kesilmemiştir. Dolayısıyla, seçilen sorularla hangi dokümanların ilgili olduğuna dermedeki 2215 doküman tek tek incelenerek karar verilmiştir.

Bilgi erişim sistemlerine yöneltilen sorular, soruların ilgililik ölçütleri, sorularla bilgi erişim sistemlerinin test edilmek istenen özellikleri, aramaların gerçekleştirildiği alanlar ve sorgu sonuçlarının değerlendirilmesi alt başlıklarda yer almaktadır. Ayrıca, seçilen sorularla ilgili dermede kaç dokümanın yer aldığı ve her bir bilgi erişim sisteminin her bir soruya karşılık erişebildiği ilgili doküman sayısı, anmanın minimumdan maksimuma çıktığı aralıkta erişilen ilgisiz dokümanların sayısı Tablo 19'da, her bir bilgi erişim sisteminin sergilediği anma-duyarlılık ve normalize sıralama değerleri (R_{norm}) Tablo 20, 21 ve 22'de yer almaktadır.

Tablo 19. Sorgularla İlgili Tüm Dokümanların Sayısı, Bilgi Erişim Sistemlerinin Erişebildikleri İlgili ve İlgisiz Doküman Sayısı

Soru	Tüm ilgili	ÜBES		TBES		KBES	
		Erişilen ilgili	Erişilen İlgisiz	Erişilen ilgili	Erişilen İlgisiz	Erişilen ilgili	Erişilen İlgisiz
1	15	15	0	15	40	15	0
2	14	0	0	14	45	14	0
3	9	4	0	9	8	9	3
4	16	3	0	13	9	13	1
5	21	10	0	14	14	21	14
6	8	8	1	7	29	8	2
7	8	8	1	7	29	8	2
8	14	2	0	5	5	5	5
9	14	4	0	8	16	8	16
Toplam	119	54	2	92	195	101	43

Tablo 20: TBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı

Soru 1		Soru 2		Soru 3		Soru 4		Soru 5		Soru 6		Soru 7		Soru 8		Soru 9	
A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D
0,067	1,000	0,071	0,031	0,111	1,000	0,063	1,000	0,048	1,000	0,125	1,000	0,125	1,000	0,077	0,500	0,077	1,000
0,133	1,000	0,143	0,061	0,222	1,000	0,125	1,000	0,095	1,000	0,250	1,000	0,250	1,000	0,154	0,667	0,154	1,000
0,200	1,000	0,214	0,081	0,333	0,750	0,188	1,000	0,143	1,000	0,375	1,000	0,375	1,000	0,231	0,750	0,231	1,000
0,267	0,571	0,286	0,105	0,444	0,800	0,250	1,000	0,190	1,000	0,500	0,800	0,500	0,800	0,308	0,444	0,308	1,000
0,333	0,625	0,357	0,119	0,556	0,500	0,313	1,000	0,238	0,833	0,625	0,625	0,625	0,625	0,385	0,500	0,385	0,455
0,400	0,667	0,429	0,140	0,667	0,545	0,375	1,000	0,286	0,750	0,750	0,333	0,750	0,333			0,462	0,462
0,467	0,636	0,500	0,159	0,778	0,583	0,438	1,000	0,333	0,538	0,875	0,194	0,875	0,194			0,538	0,389
0,533	0,421	0,571	0,174	0,889	0,533	0,500	1,000	0,381	0,500							0,615	0,333
0,600	0,429	0,643	0,180	1,000	0,529	0,563	1,000	0,429	0,529								
0,667	0,455	0,714	0,189			0,625	0,769	0,476	0,455								
0,733	0,458	0,786	0,204			0,688	0,579	0,524	0,478								
0,800	0,353	0,857	0,218			0,750	0,600	0,571	0,500								
0,867	0,371	0,929	0,232			0,813	0,419	0,619	0,520								
0,933	0,389	1,000	0,246					0,667	0,500								
1,000	0,273																
Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm	
0,73		0,15		0,52		0,84		0,55		0,53		0,53		0,48		0,67	

Tablo 21: ÜBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı

Soru 1		Soru 2		Soru 3		Soru 4		Soru 5		Soru 6		Soru 7		Soru 8		Soru 9	
A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D
0,067	1,000	0,000	0,000	0,111	1,000	0,063	1,000	0,045	1,000	0,125	1,000	0,125	1,000	0,077	1,000	0,077	1,000
0,133	1,000	0,100	0,000	0,222	1,000	0,125	1,000	0,091	1,000	0,250	1,000	0,250	1,000	0,154	1,000	0,154	1,000
0,200	1,000	0,200	0,000	0,333	1,000	0,188	1,000	0,136	1,000	0,375	1,000	0,375	1,000			0,231	1,000
0,267	1,000	0,300	0,000	0,444	1,000			0,182	1,000	0,500	1,000	0,500	1,000			0,308	1,000
0,333	1,000	0,400	0,000					0,227	1,000	0,625	1,000	0,625	1,000				
0,400	1,000	0,500	0,000					0,273	1,000	0,750	1,000	0,750	1,000				
0,467	1,000	0,600	0,000					0,318	1,000	0,875	0,875	0,875	0,875				
0,533	1,000	0,600	0,000					0,364	1,000	1,000	0,889	1,000	0,889				
0,600	1,000	0,800	0,000					0,409	1,000								
0,667	1,000	0,900	0,000					0,455	1,000								
0,733	1,000	1,000	0,000														
0,800	1,000																
0,867	1,000																
0,933	1,000																
1,000	1,000																
Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm	
1,00		0,00		1,00		1,00		1,00		0,75		0,75		1,00		1,00	

Tablo 22: KBES'in Sergilediği Anma-Duyarlılık ve Normalize Sıralama Performansı

Soru 1		Soru 2		Soru 3		Soru 4		Soru 5		Soru 6		Soru 7		Soru 8		Soru 9	
A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D
0,067	1,000	0,071	1,000	0,111	1,000	0,063	1,000	0,048	1,000	0,125	1,000	0,125	1,000	0,077	1,000	0,077	1,000
0,133	1,000	0,143	1,000	0,222	1,000	0,125	1,000	0,095	1,000	0,250	1,000	0,250	1,000	0,154	1,000	0,154	1,000
0,200	1,000	0,214	1,000	0,333	1,000	0,188	1,000	0,143	1,000	0,375	1,000	0,375	1,000	0,231	0,750	0,231	1,000
0,267	1,000	0,286	1,000	0,444	1,000	0,250	1,000	0,190	1,000	0,500	1,000	0,500	1,000	0,308	0,444	0,308	1,000
0,333	1,000	0,357	1,000	0,556	1,000	0,313	1,000	0,238	1,000	0,625	1,000	0,625	1,000	0,385	0,500	0,385	1,000
0,400	1,000	0,429	1,000	0,667	1,000	0,375	1,000	0,286	1,000	0,750	1,000	0,750	1,000			0,462	1,000
0,467	1,000	0,500	1,000	0,778	1,000	0,438	1,000	0,333	1,000	0,875	0,875	0,875	0,875			0,538	0,389
0,533	1,000	0,571	1,000	0,889	0,889	0,500	1,000	0,381	1,000	1,000	0,800	1,000	0,800			0,615	0,333
0,600	1,000	0,643	1,000	1,000	0,818	0,563	1,000	0,429	1,000								
0,667	1,000	0,714	1,000			0,625	1,000	0,476	1,000								
0,733	1,000	0,786	1,000			0,688	0,579	0,524	1,000								
0,800	1,000	0,857	1,000			0,750	0,600	0,571	1,000								
0,867	1,000	0,929	1,000			0,813	0,419	0,619	0,929								
0,933	1,000	1,000	1,000					0,667	0,700								
1,000	1,000							0,714	0,652								
								0,762	0,667								
								0,810	0,586								
								0,857	0,600								
								0,905	0,613								
								0,952	0,625								
								1,000	0,600								
Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm		Rnorm	
1,00		1,00		0,88		0,87		0,85		0,77		0,77		0,56		0,79	

Not: A: Anma, D: Duyarlılık, Rnorm: Normalize sıralama

5.5.1. 1. Soru: İrfan Çakın'ın Yazdığı Tüm Dokümanlar

Birinci soruda, İrfan Çakın'ın yazdığı dokümanlar aranmıştır. Bu soru, üstverinin bilgi erişim performansını ne ölçüde etkilediğini, özellikle de üstveri kullanmayan TBES ile üstveri kullanan KBES'in erişim performanslarını karşılaştırmak üzere seçilmiştir. İlgililik değerlendirmesinde ise, İrfan Çakın'ın yazar sorumluluğu olmadığı dokümanlar (örneğin, yer aldığı etkinlikler, editör sunumları vd.) ilgisiz kabul edilmiştir. Bu soruda, üç bilgi erişim sisteminin de tüm alanları üzerinde arama yapılmıştır.

Formülasyon: İrfan AND Çakın

ÜBES formülasyon çıktısı: (+ creator:irfan + creator:çakın) OR (+ title:irfa + title:çak) OR (+ description:irfa + description:çak) OR (+ subject:irfa + subject:çak)

TBES formülasyon çıktısı: +fulltext:irfa +fulltext:çak

KBES formülasyon çıktısı: (+ creator:irfan + creator:çakın) OR (+ title:irfa + title:çak) OR (+ description:irfa + description:çak) OR (+ subject:irfa + subject:çak) OR (+ fulltext:irfa + fulltext:çak)

Birinci soruda üç bilgi erişim sistemi de tüm ilgili dokümanlara erişebilmiştir. Anmanın minimumdan maksimuma çıktığı aralıkta ÜBES ve KBES ilgisiz hiçbir dokümana erişmemiştir. TBES ise 40 adet ilgisiz dokümana erişmiştir. Bu durum normalize sıralama değerlerine de yansımış, ÜBES ve KBES için R_{norm}=1, TBES için R_{norm}=0,75 olmuştur.

Bu soru için ÜBES'in dermedeki “sadece tüm ilgili” dokümanlara erişebilmesi beklenen bir sonuçtur. TBES'in 40 adet ilgisiz dokümana ulaşmasının sebebi, genel olarak kısa olan editör sunumlarında veya haberlerde sorgu terimlerinin geçmesidir. KBES'in ÜBES ile aynı sonuçlara ulaşmasının nedeni, kullanıcının girdiği sorgunun OR işleci ile bağlanarak tüm alanlar üzerinde aranmasıdır. KBES'in “yazar” alanında yer alan terimlerle sorgu terimlerinin çakışması durumunda kısa olan “yazar” alanı uzun olan “tam-metin” alanına göre yüksek skorlar almaktadır. Bunun neticesinde, kısa alan çakışmaları sıralamada üst

sıraya yerleşmiştir. Bu durumun, KBES'in her bir anma basamağındaki duyarlılığını artırdığı gibi normalize sıralama değerlerini de olumlu yönde etkilediği görülmektedir.

Ayrıca, bu soruda (ve ikinci soruda) gövdeleme algoritmasının özel isimleri işlemede yetersiz kaldığı da görülmektedir. Bunun temel nedeni gövdeleme algoritmasının terim sözlüğü kullanmamasıdır. Ayrıca, gövdeleme algoritmasının terimlerin büyük harfle başlamasını da özel isim çıkarımı yapmak için kullanmadığı görülmektedir.

5.5.2. 2. Soru: İrfan Çakın'a Bilimsel/Hakemli Dokümanlarda Yapılan Atıflar

İkinci soruda, hakemli makalelerde yazara yapılan atıflara erişilmek istenmiştir. Bir başka ifadeyle, yazarın yazmadığı fakat içerisinde yazarın adının geçtiği bilimsel dokümanlara erişilmesi söz konusudur.

Kullanıcının bilgi ihtiyacının temsili bilgi erişim sistemlerinin en önemli üç bileşeninden biridir. Buradan hareketle, sorgular bilgi erişim sistemlerinin maksimum performans sergileyebileceği biçimde yapılandırılmıştır. Dolayısıyla, bu soru hem üstverinin hem de alanlara dayalı sorgunun bilgi erişim performansına katkısını ölçmek üzere seçilmiştir. İlgililik değerlendirmesinde ise, İrfan Çakın'ın yazar sorumluluğu olan dokümanlar, katıldığı etkinlikler, hakkında çıkan haberler ve editör sunumları ilgisiz kabul edilmiştir. Bu soruda KBES'de yapılan sorguda “yazar”, “tam-metin” ve “tür” alanları üzerinde, TBES'de yapılan sorguda ise “tam-metin” alanı üzerinde arama yapmıştır. ÜBES “tam-metin” alanına sahip olmadığı için bu soruda anlamlı bir sorgu üretilememiştir.

Formülasyon: "İrfan Çakın" OR "Çakın, İrfan" OR "Çakın, İ." ifadeleriyle her bilgi erişim sistemi için ayrı ayrı formülasyon üretilmiştir.

ÜBES formülasyon çıktısı: Anlamlı formülasyon yapılamamaktadır.

TBES formülasyon çıktısı: fulltext:"irfa çak" OR fulltext:"çak irfa" OR fulltext:"çak i"

KBES formülasyon çıktısı: -(creator:"irfan çakın" OR creator:"çakın,

irfan" OR creator:"çakın, i.") + (fulltext:"irfa çak" OR fulltext:"çak irfa" OR fulltext:"çak i") + (type:"Makaleler" OR type:"Hakemli Yazılar")

İkinci soruda ÜBES sonuç döndüremediği için R_{norm}=0 olmuştur. ÜBES tam-metin alanına sahip olmadığı için bu sonuç normaldir. KBES'de sorgu esnasında "yazar" ve "tür" alanlarında elemeye gidildiği için sadece tüm ilgili dokümanlara erişim sağlanmıştır ve R_{norm}=1 olmuştur. TBES'de sadece tam-metin alanı olduğu için herhangi bir özellik elemesi yapılamamıştır ve anmanın minimumdan maksimuma çıktığı aralıkta tüm ilgili dokümanlarla birlikte 45 adet ilgisiz dokümana erişilmiştir.

TBES için birinci soru ile ikinci soru arasında hem erişilen ilgili doküman hem de erişilen ilgisiz doküman sayısında çok az bir fark olmasına karşın, ikinci soruda birinci soruya göre normalize sıralama değerinin (birinci soru için R_{norm}=0,73, ikinci soru için R_{norm}=0,15) çok fazla düştüğü görülmektedir. Bunun sebebi büyük ölçüde TBES'in İrfan Çakın'ın kendi yazdığı dokümanları, İrfan Çakın'a yapılan atıflardan önce listelemesinden kaynaklanmaktadır.

5.5.3. 3. Soru: "Bilgi Arama Davranışı"

Üçüncü soruda, konu sorgusu üretilmiştir. Bu soru, insana dayalı dizinleme ile makineye dayalı dizinlemenin karşılaştırılması amacıyla, özellikle insana dayalı dizinleme ile dokümanı temsil etmek üzere atanmış terimlerin dokümanı temsil edip edemediğini saptamak amacıyla seçilmiştir. Ayrıca, bu sorgu deyim olarak yapılandırılmış ve kesin çakışmanın olması beklenerek sorgu daraltılmıştır. İlgililik değerlendirmesinde ise, salt "bilgi arama davranışı" hakkında yazılmış bilimsel dokümanlar ve konuyu tanımlayan veya öneminden bahseden dokümanlar ilgili kabul edilmiştir. Diğer dokümanlar ilgisiz kabul edilmiştir. Bu soruda ve bundan sonraki sorularda, sorgu ifadeleri ile üç bilgi erişim sisteminin tüm alanları üzerinde arama yapılmıştır.

Formülasyon: "bilgi arama davranışı"

ÜBES formülasyon çıktısı: creator:"bilgi arama davranışı" OR title:"bilgi ara davranış" OR description:"bilgi ara davranış" OR subject:"bilgi ara

davranış"

TBES formülasyon çıktısı: fulltext:"bilgi ara davranış"

KBES formülasyon çıktısı: creator:"bilgi arama davranışı" OR title:"bilgi ara davranış" OR description:"bilgi ara davranış" OR subject:"bilgi ara davranış" OR fulltext:"bilgi ara davranış"

ÜBES, üçüncü soruda dermedeki toplam 9 ilgili dokümandan 4'üne erişebilmiştir. Bunun temel nedeni, derleme makalelerin veya çok fazla konuyu işleyen makalelerin konu başlıklarında ilgili sorgu deyiminin geçmemesidir. Örneğin, kütüphanecilik eğitim programının yeniden yapılandırılmasını işleyen bir dokümanda birçok konuyla birlikte “bilgi arama davranışı” konusunun ne olduğu, hangi ihtiyaçlardan ortaya çıktığı ve ders programındaki kapsamının ne olması gerektiği gibi konuyla ilgili bilgi verilmiştir. Nesnel değerlendirme çerçevesinde söz konusu dokümanın bilgi ihtiyacı ile ilgili olduğuna karar verilmiştir, fakat söz konusu doküman çok sayıda konuyu/dersi işlediği için insana dayalı dizinlemede tüm konularla ilgili konu başlığı verilmemiştir. Bu sebepten ötürü ilgili dokümana ÜBES erişememiştir. Ayrıca, konuyla ilgili bir araştırma makalesinde; yazarın, makalesinin “Anahtar sözcükler” bölümüne söz konusu konu başlığını atamamasından ötürü de erişimin sağlanamadığı tespit edilmiştir.

Bu soruda, ÜBES hiçbir ilgisiz dokümana erişmediği için R_{norm}=1 olmuştur. TBES ve ÜBES ise tüm ilgili dokümanlara erişmiştir. Ancak, TBES 8 adet, ÜBES ise 3 adet ilgisiz dokümana da erişmiştir. Bu sebepten ötürü, TBES için R_{norm}=0,52, KBES için R_{norm}=0,88 olmuştur. Normalize sıralama değerlerinin farklı çıkmasında yine üstveri rol oynamıştır.

5.5.4. 4. Soru: AACR, AACR1, AACR2

Dördüncü soruda, “Anglo-American Cataloguing Rules” basımlarının veya sürümlerinin İngilizce kısaltmalarıyla (AACR, AACR1 ve AACR2) sorgu üretilmiştir. Bu soru, kullanıcıların *Türk Kütüphaneciliği* açık arşivindeki aramalarda sıkça kullanılan özel kısaltmaların araştırmada tasarlanan bilgi erişim sistemlerine yöneltilmesi durumunda bilgi erişim sistemlerinin nasıl bir

performans sergileyeceğini değerlendirmek üzere seçilmiştir. İlgililik değerlendirmesinde, salt AACR hakkında yazılmış, konu hakkında eleştiri yapılan ve AACR özelliklerini işleyen dokümanlar ilgili olarak kabul edilmiştir. Kütüphane tanıtımları, konu hakkında yapılan eğitim haberleri ve editör sunumları ilgisiz kabul edilmiştir.

Formülasyon: AACR OR AACR1 OR AACR2

ÜBES formülasyon çıktısı: creator:"aacr" OR creator:"aacr1" OR creator:"aacr2" OR title:"aacr" OR title:"aacr1" OR title:"aacr2" OR description:"aacr" OR description:"aacr1" OR description:"aacr2" OR subject:"aacr" OR subject:"aacr1" OR subject:"aacr2"

TBES formülasyon çıktısı: fulltext:"aacr" OR fulltext:"aacr1" OR fulltext:"aacr2"

KBES formülasyon çıktısı: creator:"aacr" OR creator:"aacr1" OR creator:"aacr2" OR title:"aacr" OR title:"aacr1" OR title:"aacr2" OR description:"aacr" OR description:"aacr1" OR description:"aacr2" OR subject:"aacr" OR subject:"aacr1" OR subject:"aacr2" OR fulltext:"aacr" OR fulltext:"aacr1" OR fulltext:"aacr2"

Bu soru için dermede 16 adet ilgili doküman bulunmaktadır. ÜBES ilgili dokümanlardan sadece 3'üne, TBES ve KBES 13'üne erişebilmiştir. ÜBES'in az sayıda dokümana erişebilmesinin temel nedeni, üstveri alanlarında kısaltmanın kullanılmamış olmasıdır. Ayrıca, kısaltmalarda denetimsizliklerin/tutarsızlıkların olması da önemli rol oynamıştır. Örneğin, dermedeki birçok dokümanda İngilizce açılımın kısaltması olan “AACR” veya “AACR2” gibi kısaltmalar kullanılırken, bazı dokümanlarda ve konu başlıklarında Türkçe açılımın kısaltması olan “AAKK” kısaltması kullanılmıştır. TBES ve KBES bu sorguda 13 adet ilgili dokümana erişmiş, 3 adet ilgili dokümana erişememiştir. 3 adet ilgili dokümana erişilememesinin sebebi, dokümanlarda “AACR” yerine “AAKK” kısaltmasının veya “AACRII” ve “AACRI” kısaltmalarının kullanılmasıdır. Bu soru için bilgi erişim sistemlerinde eş anlamlılar veya eş anlamlı kısaltmalar sözlüğü ile sorgunun genişletilmesinin gerekli olduğu görülmektedir.

Bu soruda da ÜBES hiçbir ilgisiz dokümana erişmediği için R_{norm}=1 olmuştur. TBES 9 adet ilgisiz dokümana erişerek R_{norm}=0,84 olmuştur. KBES ise 1 ilgisiz dokümana erişerek R_{norm}=0,87 olmuştur. Bu soruda KBES daha az sayıda ilgisiz dokümana erişmesine rağmen normalize sıralama değerlerinde TBES ile aralarında büyük fark oluşmamıştır. Bunun temel nedeni, TBES'de ilgisiz dokümanların büyük ölçüde sonuç kümesinin sonunda toplanmasıdır.

5.5.5. 5. Soru: OPAC, "Çevrimiçi Katalog"

Beşinci soru, dördüncü soruya benzemektedir. Bu soruda, dördüncü sorudan farklı olarak İngilizce kısaltma ile birlikte Türkçe açılım da sorguya eklenmiştir. OPAC'ın tanımını, işleyişini, özelliklerini veren ve OPAC tanıtımı yapılan dokümanlar ilgili kabul edilmiştir. Geri kalan dokümanlar (örneğin, kütüphane tanıtımlarında, söz konusu kütüphanenin OPAC'a sahip olması gibi) ilgisiz kabul edilmiştir.

Formülasyon: OPAC OR "çevrimiçi katalog"

ÜBES formülasyon çıktısı: creator:"opac" OR creator:"çevrimiçi katalog" OR title:"opaç" OR title:"çevrimiç katalogu" OR description:"opaç" OR description:"çevrimiç katalogu" OR subject:"opaç" OR subject:"çevrimiç katalogu"

TBES formülasyon çıktısı: fulltext:"opaç" OR fulltext:"çevrimiç katalogu"

KBES formülasyon çıktısı: creator:"opac" OR creator:"çevrimiçi katalog" OR title:"opaç" OR title:"çevrimiç katalogu" OR description:"opaç" OR description:"çevrimiç katalogu" OR subject:"opaç" OR subject:"çevrimiç katalogu" OR fulltext:"opaç" OR fulltext:"çevrimiç katalogu"

Bu soru için dermede 21 adet ilgili doküman bulunmaktadır. ÜBES ilgili dokümanlardan 10 tanesine, TBES 14 tanesine, KBES ise 21 tanesine erişmiştir. ÜBES'in eriştiği 7 adet dokümana TBES erişememiştir. Bunun temel nedeni, 6 adet dokümanda "OPAC" veya "çevrimiçi katalog" ifadeleri/terimleri yerine "kütüphane otomasyonu" ifadesinin/teriminin kullanılmasıdır. Özellikle, 3

adet dokümanın içeriğini sadece OPAC tanıtımı oluşturmasına rağmen, “OPAC” veya “çevrimiçi katalog” terimleri yerine “kütüphane otomasyonu” teriminin/ifadesinin kullanılmış olması ilginçtir. Ayrıca, eski bir tarihte oluşturulan bir dokümanda OPAC'ın işleyişi anlatılmasına rağmen, “OPAC”, “çevrimiçi katalog” veya “kütüphane otomasyonu” ifadelerinden/terimlerinden hiçbirinin kullanılmaması da dikkat çekicidir. Ancak, insana dayalı dizinlemede söz konusu dokümanların konu alanına “OPAC” veya “çevrimiçi katalog” terimlerinin atanmış olmasından ötürü ÜBES ve KBES bahsi geçen dokümanlara erişim sağlamıştır. Bu soruda da bir önceki soruda olduğu gibi eş anlamlılar sözlüğü kullanılarak sorgunun genişletilmesinin gerekli olduğu görülmektedir.

ÜBES'te bu soruda ilgisiz hiçbir dokümana erişilmemiştir ve R_{norm}=1 olmuştur. TBES'te 14 ilgisiz dokümana erişilerek R_{norm}=0,55 olmuştur. KBES'te de 14 ilgisiz dokümana erişilmiştir. Ancak, üstverinin katkısı ile eriştiği ilgili dokümanların erişim kümesinin üstünde yer almasından ötürü R_{norm}=0,85 olmuştur.

5.5.6. 6. ve 7. Soru: [Engelliler, Özürlüler] ve [Engelli, Özürlü]

Altıncı ve yedinci sorularda engelli veya özürlü kişilerle ilgili literatür taraması yapılmıştır. Ayrıca, soru seçilirken ayrım gücü kazandırabilecek “kişiler” veya “kullanıcılar” gibi terimler soruya alınmamıştır. İlgililik değerlendirmesinde, engelli/özürlü kullanıcıları veya kişileri işleyen tüm dokümanlar ilgili kabul edilmiştir. İnternet/web erişimine yapılan engellemeleri ve sansür konusunu işleyen dokümanlar başta olmak üzere engelli/özürlü kişileri işlemeyen tüm dokümanlar ilgisiz kabul edilmiştir.

Terimlerin ayrım gücüne ek olarak, altıncı ve yedinci soruda sözlük kullanmadan, sadece morfolojik analiz yapan gövdeleme algoritmasının bilgi erişim performansına etkisi değerlendirilmeye alınmıştır. Ayrıca, gövdeleme algoritmasının isim türündeki köke sahip terimlerde kullanılan yapım ekleri ile birlikte çekim eklerini işleyişi (altıncı soru) ve gövdelemeye ihtiyaç duymayan terimleri (yedinci soru) işleyişi gözlenmeye çalışılmıştır.

6.Soru: Engelliler, özürlüler

Formülasyon: engelliler OR özürlüler

ÜBES formülasyon çıktısı: creator:"engelliler" OR creator:"özürlüler" OR title:"engelli" OR title:"özürlü" OR description:"engelli" OR description:"özürlü" OR subject:"engelli" OR subject:"özürlü"

TBES formülasyon çıktısı: fulltext:"engelli" OR fulltext:"özürlü"

KBES formülasyon çıktısı: creator:"engelliler" OR creator:"özürlüler" OR title:"engelli" OR title:"özürlü" OR description:"engelli" OR description:"özürlü" OR subject:"engelli" OR subject:"özürlü" OR fulltext:"engelli" OR fulltext:"özürlü"

7. Soru: Engelli, özürlü

Formülasyon: engelli OR özürlü

ÜBES formülasyon çıktısı: creator:"engelli" OR creator:"özürlü" OR title:"engelli" OR title:"özürlü" OR description:"engelli" OR description:"özürlü" OR subject:"engelli" OR subject:"özürlü"

TBES formülasyon çıktısı: fulltext:"engelli" OR fulltext:"özürlü"

KBES formülasyon çıktısı: creator:"engelli" OR creator:"özürlü" OR title:"engelli" OR title:"özürlü" OR description:"engelli" OR description:"özürlü" OR subject:"engelli" OR subject:"özürlü" OR fulltext:"engelli" OR fulltext:"özürlü"

Altıncı ve yedinci sorular için dermede toplam 8 adet ilgili doküman bulunmaktadır. ÜBES altıncı ve yedinci sorularda ilgili dokümanların tamamına erişebilmiştir. KBES'in de tüm ilgili dokümanlara erişebildiği, fakat TBES'in 1 adet ilgili dokümana erişemediği görülmektedir. Bunun sebebi, dermede yer alan ilgili bir dokümanın “görme engelliler” konusunu işlemesi, fakat doküman içerisinde “engelli” veya “özürlü” terimlerinin geçmemesidir. Söz konusu dokümanda “kör” ve “âmâ” terimleri kullanılmıştır. İnsana dayalı dizinlemede dokümana konu başlığı olarak “engelli” veya “özürlü” terimlerinin atanmış olmasından dolayı ÜBES ve KBES söz konusu dokümana erişebilmiştir. Bu soru için bilgi erişim sistemlerinin denetimli söz dağarcığı kullanmasının ve

“özürlü” veya “engelli” terimlerinin altında yer alacak “kör” veya “sağır” gibi daha dar terimlerle sorgunun genişletilmesinin gerekli olduğu görülmektedir. Ayrıca, bu sorularda gövdeleme algoritmasının sorunsuz işlediği ve bilgi erişim performansına olumlu katkı yaptığı görülmektedir.

ÜBES'te tüm sorular içerisinde sadece bu sorularda birer tane ilgisiz dokümana erişilmiştir. Bunun sebebi, üstveri alanında yer alan “engelli” ifadesinin kişilerle ilgili olmamasından kaynaklanmaktadır. Bu sorular için ÜBES'te R_{norm}=0,75 olmuştur. TBES'te her iki soruda da 29 ilgisiz dokümana erişilerek R_{norm}=0,53 olmuştur. KBES'te ise her iki soruda da 2 ilgisiz dokümana erişilerek R_{norm}=0,77 olmuştur. Ancak, KBES'te erişilen 2 adet ilgisiz dokümanın erişim kümesinin üst sıralarında yer alması nedeniyle TBES'den çok daha az sayıda ilgisiz dokümana erişilmiş olmasına rağmen TBES'e oranla çok yüksek bir normalize sıralama sonuç değeri elde edilememiştir.

5.5.7. 8. ve 9. Soru: [“Kullanıcılara Eğitimler”, “Okuyuculara Eğitimler”, Oryantasyonlar] ve [“Kullanıcı Eğitimi”, “Okuyucu Eğitimi”, Oryantasyon]

Sekizinci ve dokuzuncu sorularda, kütüphane kullanıcılarının eğitimi hakkında bilgi edinmek istenmiştir. Kütüphane ve kütüphane materyalleriyle ilgili kullanıcılara verilen eğitimler, uygulama örnekleri ve konunun önemi hakkında bilgi veren dokümanlar ilgili kabul edilmiştir. Bunlar dışında ve kütüphane konusuyla ilgili olmayan dokümanlar ilgisiz kabul edilmiştir.

Bu soruda da altıncı ve yedinci soruda olduğu gibi gövdeleme algoritmasının bilgi erişim performansına etkisi değerlendirmeye alınmıştır. Altıncı ve yedinci sorulardan farklı olarak, bu soruda iki terimden oluşan deyim sorgu üretilmiştir. Dolayısıyla, gövdeleme algoritmasının iki terimi de sorunsuz gövdelemesi beklenmiştir. Ayrıca, gövdeleme algoritmasının, fiil türündeki köke sahip terimlerde kullanılan çeşitli yapım ekleri (-ci ve -im) ile birlikte çekim eklerinin de nasıl işlendiği gözlenmeye çalışılmıştır.

8. Soru: “Kullanıcılara eğitimler”, “okuyuculara eğitimler”, oryantasyonlar

Formülasyon: "kullanıcılara eğitimler" OR "okuyuculara eğitimler" OR oryantasyonlar

ÜBES formülasyon çıktısı: creator:"kullanıcılara eğitimler" OR creator:"okuyuculara eğitimler" OR creator:"oryantasyonlar" OR title:"kullanıcı eğitim" OR title:"okuyucu eğitim" OR title:"oryantasyon" OR description:"kullanıcı eğitim" OR description:"okuyucu eğitim" OR description:"oryantasyon" OR subject:"kullanıcı eğitim" OR subject:"okuyucu eğitim" OR subject:"oryantasyon"

TBES formülasyon çıktısı: fulltext:"kullanıcı eğitim" OR fulltext:"okuyucu eğitim" OR fulltext:"oryantasyon"

KBES formülasyon çıktısı: creator:"kullanıcılara eğitimler" OR creator:"okuyuculara eğitimler" OR creator:"oryantasyonlar" OR title:"kullanıcı eğitim" OR title:"okuyucu eğitim" OR title:"oryantasyon" OR description:"kullanıcı eğitim" OR description:"okuyucu eğitim" OR description:"oryantasyon" OR subject:"kullanıcı eğitim" OR subject:"okuyucu eğitim" OR subject:"oryantasyon" OR fulltext:"kullanıcı eğitim" OR fulltext:"okuyucu eğitim" OR fulltext:"oryantasyon"

9. Soru: “Kullanıcı eğitimi”, “okuyucu eğitimi”, oryantasyon

Formülasyon: "kullanıcı eğitimi" OR "okuyucu eğitimi" OR oryantasyon.

ÜBES formülasyon çıktısı: creator:"kullanıcı eğitimi" OR creator:"okuyucu eğitimi" OR creator:"oryantasyon" OR title:"kullanıç eğit" OR title:"okuyuç eğit" OR title:"oryantasyo" OR description:"kullanıç eğit" OR description:"okuyuç eğit" OR description:"oryantasyo" OR subject:"kullanıç eğit" OR subject:"okuyuç eğit" OR subject:"oryantasyo"

TBES formülasyon çıktısı: fulltext:"kullanıç eğit" OR fulltext:"okuyuç eğit" OR fulltext:"oryantasyo"

KBES formülasyon çıktısı: creator:"kullanıcı eğitimi" OR creator:"okuyucu eğitimi" OR creator:"oryantasyon" OR title:"kullanıç

eğit" OR title:"okuyuç eğit" OR title:"oryantasyo" OR description:"kullanıç eğit" OR description:"okuyuç eğit" OR description:"oryantasyo" OR subject:"kullanıç eğit" OR subject:"okuyuç eğit" OR subject:"oryantasyo" OR fulltext:"kullanıç eğit" OR fulltext:"okuyuç eğit" OR fulltext:"oryantasyo"

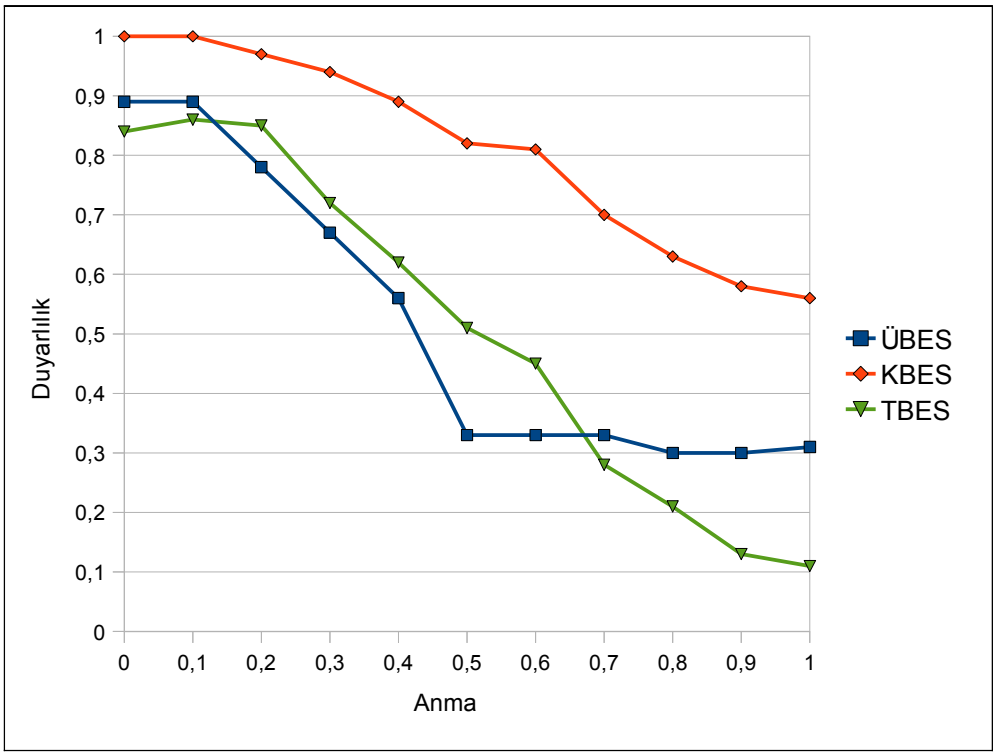
Sekizinci ve dokuzuncu sorular için dermede toplam 14 adet ilgili doküman bulunmaktadır. Sekizinci soruda; ÜBES'te 2 adet, TBES'te ve KBES'te 5 adet ilgili dokümana erişilebilmiştir. Dokuzuncu soruda ise, ÜBES'te 4 adet, TBES'te ve KBES'te 8 adet ilgili dokümana erişilebilmiştir. Bu sorularda gövdeleme algoritması hem dizinleme esnasında hem de sorgularda tutarlılık gösteremediği için çakışma sağlanamamış ve erişim performansına olumlu katkı sağlanamamıştır.

Sekizinci ve dokuzuncu sorularda ÜBES'te yine ilgisiz hiçbir dokümana erişilmediği için Rnorm=1 olmuştur. Sekizinci soruda TBES'te 5 adet ilgisiz dokümana erişilerek Rnorm=0,48 olmuştur. KBES'te de 5 adet ilgisiz dokümana erişilmiştir, fakat Rnorm=0,56 olmuştur. Dokuzuncu soruda ise, TBES'te 16 adet ilgisiz dokümana erişilerek Rnorm=0,67 olmuştur. ÜBES'te de 16 adet ilgisiz doküman erişilerek Rnorm=0,79 olmuştur. Bu sorularda TBES ve KBES eşit sayıda ilgisiz dokümana erişmesine rağmen normalize sıralama değerleri farklı çıkmıştır. KBES'in normalize sıralama değerlerinin yüksek çıkmasının nedeni, yine kısa olan üstveri alanları ile sorguda geçen terimlerin çakışmasıyla ilgili dokümanların erişim kümesinin üst sıralarında yer almasından kaynaklanmaktadır.

5.6. PERFORMANS DEĞERLENDİRME SONUÇLARI

Şekil 24'te bilgi erişim sistemlerinin interpolasyon (interpolate) uygulanmış 11 adet anma basamağındaki ortalama duyarlılık grafiği yer almaktadır. Tablo 23'te ise her bir soru için 11 adet anma basamağındaki tekil duyarlılık değerleriyle, mikro[4] ortalama duyarlılık değerleri yer almaktadır.

4 Mikro ortalamada sayıların aritmetik ortalaması alınmaktadır.



Şekil 24. Bilgi Erişim Sistemlerinin Ortalama Anma ve Duyarlılık Grafiği

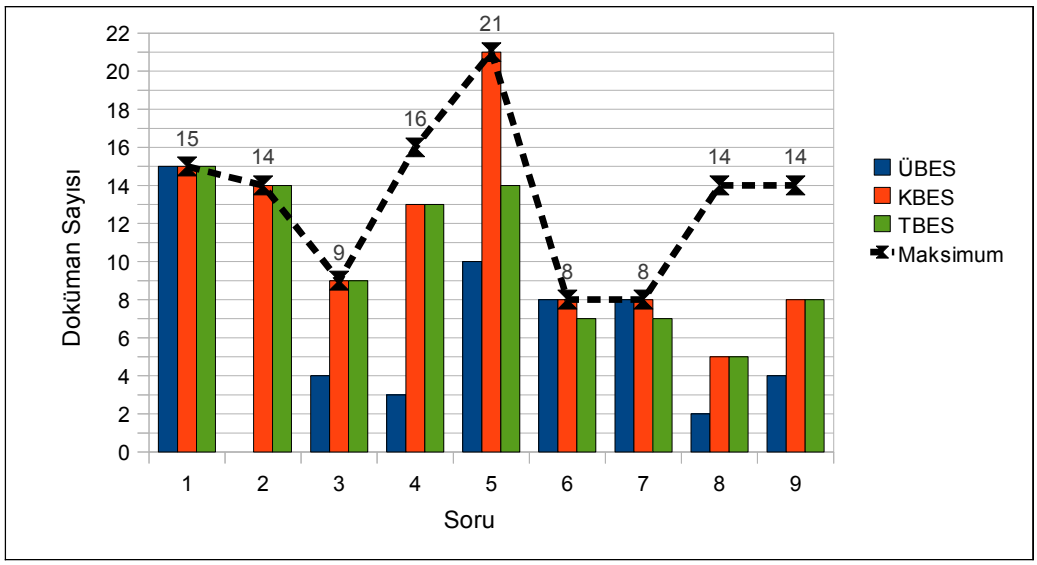
Gerçekleştirilen araştırmanın en önemli amaçlarından biri üç bilgi erişim sisteminin 11 adet anma basamağındaki ortalama duyarlılık değerleri arasında anlamlı bir farkın olup olmadığını saptamaktır. Bu amaçla, bilgi erişim sistemlerinin 11 adet anma basamağındaki ortalama duyarlılık performansları arasındaki farkın istatistiksel açıdan anlamlı olup olmadığını saptamak için Kruskal-Wallis testi uygulanmış ve fark anlamlı bulunmuştur (H = 8,595, *s.d.* = 2, p = .014 < .05). Farkın hangi bilgi erişim sisteminden kaynaklandığını saptamak üzere de Mann-Whitney testi uygulanmış, farkın KBES'ten kaynaklandığı saptanmıştır (TBES ile p = .017 < .05, ÜBES ile p = .008 < .05). ÜBES ve TBES arasındaki duyarlılık performans farkı ise istatistiksel açıdan anlamlı değildir (p = .669 > .05).

Ayrıca, her bir bilgi erişim sistemi için anma ve duyarlılık arasındaki ilişki test edilmiş, güçlü bir negatif ilişki saptanmıştır (Tüm bilgi erişim sistemleri için p = .000, ÜBES için r = -926, TBES için r = -984, KBES için r = -982). Yani, her bir bilgi erişim sistemi için anma değeri arttığında duyarlılık değeri düşmüştür.

Tablo 23. 11 Adet Anma Basamağında Bilgi Erişim Sistemlerinin Ortalama Duyarlılık Değerleri

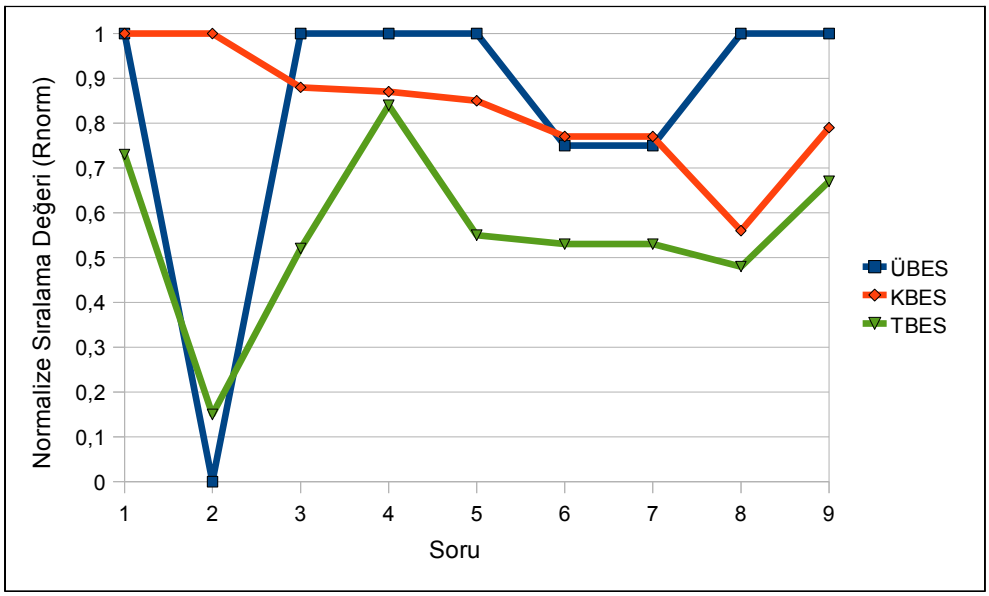
Soru	Anma																														
	0			0,1			0,2			0,3			0,4			0,5			0,6			0,7			0,8			0,9			1		
	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T	Ü	K	T
1	1	1	1	1	1	1	1	1	1	1	1	0,63	1	1	0,67	1	1	0,42	1	1	0,45	1	1	0,45	1	1	0,37	1	1	0,38	1	1	0,27
2	-	1	0,03	-	1	0,06	-	1	0,1	-	1	0,11	-	1	0,13	-	1	0,17	-	1	0,18	-	1	0,2	-	1	0,21	-	1	0,23	-	1	0,24
3	1	1	1	1	1	1	1	1	1	1	1	0,75	1	1	0,8	-	1	0,5	-	1	0,54	-	1	0,58	-	0,88	0,53	-	0,88	0,53	-	0,81	0,52
4	1	1	1	1	1	1	1	1	1	-	1	1	-	1	1	-	1	1	-	1	0,76	-	0,6	0,6	-	0,41	0,41	-	-	-	-	-	-
5	1	1	1	1	1	1	1	1	0,83	1	1	0,53	1	1	0,52	-	1	0,5	-	0,92	0,52	-	0,66	-	-	0,6	-	-	0,62	-	-	0,6	-
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,8	1	1	0,62	1	1	0,33	0,87	0,87	0,19	0,87	0,87	-	0,88	0,8	-
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,8	1	1	0,62	1	1	0,33	0,87	0,87	0,19	0,87	0,87	-	0,88	0,8	-
8	1	1	0,5	1	1	0,66	-	0,75	0,75	-	0,5	0,5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	1	1	1	1	1	1	1	1	1	1	1	1	-	1	0,46	-	0,38	0,38	-	0,33	0,33	-	-	-	-	-	-	-	-	-	-	-	-
Ortalama	0,89	1,00	0,84	0,89	1,00	0,86	0,78	0,97	0,85	0,67	0,94	0,72	0,56	0,89	0,62	0,33	0,82	0,51	0,33	0,81	0,45	0,33	0,70	0,28	0,30	0,63	0,21	0,30	0,58	0,13	0,31	0,56	0,11

Not: T:TBES, Ü: ÜBES ve K: KBES



Şekil 25. 9 Soruya Karşılık Bilgi Erişim Sistemlerinin Erişebildikleri Doküman Sayıları

Şekil 25'te ise 9 soruya karşılık bilgi erişim sistemlerinin erişebildikleri doküman sayıları yer almaktadır. ÜBES, tüm dokümanların sadece %47'sine, TBES %77'sine, KBES ise %85'ine erişebilmiştir. Toplamda; en az ilgisiz dokümana ÜBES, en fazla ilgisiz dokümana TBES, en fazla ilgili dokümana da KBES erişmiştir. Bilgi erişim sistemlerinin erişebildikleri doküman sayıları (veya anma) arasındaki farkın istatistiksel açıdan anlamlı olup olmadığını saptamak için Kruskal-Wallis testi uygulanmış, fakat fark anlamsız bulunmuştur (H = 5,116, $s.d.$ = 2, p =.077 > .05).



Şekil 26. Bilgi Erişim Sistemlerinin Her Bir Sorudaki R_{norm} Değerleri

Şekil 26'da yer alan normalize sıralama değerleri (R_{norm}) incelendiğinde, ÜBES'in toplam 9 sorunun 6'sında (%67) maksimum performans (R_{norm}=1) sergilediği dikkat çekmektedir. ÜBES'in; ikinci, altıncı ve yedinci sorular dışında tüm sorularda TBES ve KBES'ten daha iyi performans sergilediği gözükmektedir. ÜBES yapısı gereği ikinci soruya cevap veremediği için en düşük performansını (R_{norm}=0) ikinci soruda sergilemiştir.

KBES, birinci ve ikinci soruda normalize sıralama performansında maksimuma ulaşmıştır. Özellikle, ikinci soruda (atıf sorusu) TBES'den çok daha iyi performans (KBES için R_{norm}=1, TBES için R_{norm}=0,16) sergilemiştir. Ayrıca, altıncı ve yedinci sorularda da en yüksek performansı KBES sergilemiştir. KBES en düşük performansını (R_{norm}=0,56) sekizinci soruda sergilemiştir.

TBES, normalize sıralama performansında sadece ikinci soruda (atıf sorusu), ÜBES sonuç döndüremediği için ÜBES'ten üstün gelmiştir, fakat üstünlüğü kullanıcı için pek anlamlı olmayabilir (R_{norm}=0,15; ilgisiz dokümanlar çoğunlukla ilgililerin üzerinde listelenmiştir). TBES, hiçbir soruda KBES'e üstünlük sağlayamamış, fakat dördüncü soruda KBES'in performasına çok yaklaşmıştır. Dördüncü soruda, ÜBES'te 3 dokümana, KBES'te ve TBES'te ise 13'er dokümana erişilmiştir. Dolayısıyla, üstverinin sorguyla az çakıştığı dördüncü soruda KBES ile TBES arasındaki normalize erişim performansındaki fark kapanmıştır.

Gerçekleştirilen araştırmanın en önemli amaçlarından bir diğeri de üç bilgi erişim sisteminin normalize sıralama performansları arasında anlamlı bir farkın olup olmadığını saptamaktır. Bilgi erişim sistemlerinin normalize sıralama erişim performansları arasındaki farkın istatistiksel açıdan anlamlı olup olmadığını saptamak amacıyla Kruskal-Wallis testi uygulanmış ve fark anlamlı bulunmuştur (H = 11,744, $s.d.$ = 2, p = .003 < .05). Farkın hangi bilgi erişim sisteminden kaynaklandığını saptamak üzere de Mann-Whitney testi uygulanmış, farkın TBES'ten kaynaklandığı saptanmıştır. Yani, hem ÜBES ile (p = .008 < .05) hem de KBES ile (p = .002 < .05) TBES'in arasındaki normalize sıralama performansı farkı istatistiksel açıdan anlamlıdır. ÜBES ile KBES arasındaki

normalize sıralama performans farkı istatistiksel açıdan anlamlı değildir (p = . 287 > .05).

ÜBES; KBES ve TBES'e oranla çok daha düşük sayıda dokümana erişmiş olsa da normalize sıralama performansında erişilen ilgili dokümanların sayısı değil, ilgili dokümanların ilgisiz dokümanların önünde yer alması önemlidir. Dolayısıyla, normalize performans değerlendirmesi için ÜBES ve KBES arasındaki performans eşitliğinin (istatistiksel olarak anlamsızlığın) kabul edilmesi gerekmektedir.

6. BÖLÜM

SONUÇ VE ÖNERİLER

İnsanlığın ortak ürünü olarak görülen bilimsel bilgiye erişimin kolaylaştırılması amacıyla, 1990'lı yılların sonunda açık erişim modeli ortaya atılmıştır. Açık erişim, kısa bir süre içerisinde akademik çevrelerce benimsenerek desteklenmeye başlanmıştır. Konu hakkında farkındalık yaratma amacıyla birçok toplantı düzenlenmiş ve bu toplantılar sonucunda çok sayıda bildirge yayınlanmıştır. Ayrıca, konunun yasal boyutları da gündeme gelmiş, yayıncılarla yazarlar arasında özel sözleşmeler oluşturularak ve CC lisansları ile açık erişimi destekleyebilecek yasal altyapı oluşturulmaya çalışılmıştır.

Açık erişimin önemli bir konusu da bilginin erişilebilmesi için nasıl organize edileceğidir. Bilginin organizasyonu ve erişimi hakkında standartların belirlenmesi amacıyla da OAI kurulup, önemli kuruluşlarca desteklenmiştir. OAI, bilginin organizasyonu konusunda 2001 yılında OAI-PMH 1. sürümünü yayınlamış, ancak 2002 yılında OAIS'in kesin referans modelini ortaya koymasıyla OAI-PMH 2. sürümünü yayınlanmıştır. OAI-PMH 2 ile de DC kullanımı zorunlu tutulmuştur.

OAI-PMH 2 ayrıntılı olarak incelendiğinde, gerçekleştirimi kolay bir protokol olması için özel bir çaba sarf edildiği rahatlıkla görülebilmektedir. Ancak, bu durum birçok özelliğin de kısıtlı olmasına neden olmuştur. Ayrıca, OAI-PMH sürümlerinin ikisinde de “ham bilgi kaynağına doğrudan erişim” belirtecinin sağlanmasının zorunlu tutulmamış olması atlanmış olan en kritik özelliktir. Bu sebepten ötürü açık arşivler için geliştirilmiş yazılımların neredeyse tamamı, DC elementlerinden “identifier” elementine bilgi kaynağının gösterim URL'ini atamaktadır. Bunun sonucunda, söz konusu bilgi kaynaklarının kendilerine/ham hallerine erişim güçleşmektedir. Bu durum, bilgi kaynaklarının sadece üstveri ile erişilebilir olmasına ve koruma amaçlı depolanamamasına neden olmuştur. Bu sorunların üstesinden gelmek üzere birçok kurum kendi ihtiyaçları doğrultusunda uygulama geliştirmiş, 2008 yılının sonununda da OAI konuya

ilişkin OAI-ORE standardını yayınlamıştır. Ancak, OAI-ORE halen birçok açık arşiv yazılımınca desteklenen bir standart olamamıştır. Destekleyen yazılımlar da varsayılan ayar olarak OAI-ORE özelliğini devre dışı bırakmaktadır. Bu durum, hem açık erişim hem de açık erişim için önerilen CC lisanslarının temel amaçlarına ters düşmektedir.

Bu çalışmada açık erişim standartlarını değerlendirme ve açık arşivlerde bilginin organizasyonu ve erişimi konusunda gerçekleştirilebilecek erişim tekniklerini sınamak amacıyla "KBES'in duyarlılık performansı ÜBES'ten ve TİBES'ten yüksektir", "TİBES'in normalize sıralama performansı ÜBES'ten ve KBES'ten düşüktür" ve "Üç bilgi erişim sisteminin seçilen sorulara karşı eriştikleri doküman sayısı birbirinden farklıdır" şeklinde hipotezler oluşturulmuştur. Oluşturulan hipotezlerin sınanabilmesi için *Türk Kütüphaneciliği* dergisinde yayınlanmış yaklaşık 2000 adet doküman OCR aracılığı ile metin tabanlı PDF formatına dönüştürülmüş, 2215 adet doküman DC setine bağlı kalınarak dizinlenmiş ve OJS 2.2.4 sürümü kullanılarak açık arşiv yaratılmıştır. Çalışma kapsamında, açık arşivlerin özellikleri göz önünde bulundurularak üç farklı bilgi erişim sistemi (ÜDES, TBES ve KBES) tasarlanmıştır. Araştırmada betimleme yöntemi kullanılarak hipotezler sınanmış; anma-duyarlık ve normalize sıralama perfomanslarının verileri ilgili literatüre göre değerlendirilmiştir.

Araştırmadan elde edilen bulgular değerlendirilerek aşağıda sıralanan sonuçlara ulaşabilmek mümkündür.

- OAI-PMH sürümleri tek başına açık erişimin amacını destekler nitelikte değildir. OAI-PMH'nin OAI-ORE ile birlikte kullanımı, üstveri ile birlikte dokümanların tam-metnine doğrudan erişimi sağlayabilmektedir. Bunun sonucunda, açık arşivler için daha yüksek performanslı bilgi erişim sistemleri tasarlanabileceği gibi, açık arşivlerde yer alan dokümanların uzun süreli korunabilmesine de zemin hazırlanacaktır.

- Bilgi erişim sistemleri arasında en az ilgisiz dokümana ÜBES'te erişilmiştir. Bunun sebebi, ÜBES'te az sayıda (ancak, erişim açışından özgül veya yoğun) ayrık terimin dizinlenmesine dayanmaktadır. Ayrıca,

dermede yer alan dokümanların %85'inin öz alanına sahip olmaması, ÜBES'te dizinlenen ayrık terim sayısının az olmasının sebebi olarak görülmelidir. Bu durum, az sayıda ilgisiz dokümana erişilmesine neden olmuştur. Ancak, ayrık terim sayısının az olması doküman temsilini olumsuz yönde etkilemiş ve ÜBES'te en az sayıda ilgili dokümana erişilmeye de neden olmuştur. Sonuç olarak, ÜBES hem en az sayıda ilgili (tüm dokümanların %45'i) hem de ilgisiz (2 adet) dokümanın erişildiği bilgi erişim sistemi olmuştur. Bu durum duyarlılık ve normalize sıralama değerlerini de etkilemiştir. ÜBES'in en düşük duyarlılık değeri 0,3 olmuştur. Normalize sıralama değerlerinde de soruların %65'inde maksimum (1) performans sergilenmiştir.

- Sadece tam-metne/doğal dile dayalı otomatik dizinleme yapan TBES en fazla ilgisiz dokümana erişilen bilgi erişim sistemi olmuştur. Bunun sebebi, ayrık terim sayısının fazla olmasına dayanmaktadır. Fazla sayıda ortak terim birçok dokümanda geçtiği için erişilen ilgisiz doküman sayısı artmıştır. Ancak, ayrık terim sayısının fazla olması doküman temsilini artırmış, TBES'in ikinci sıradaki en fazla ilgili dokümana erişilebilen bilgi erişim sistemi olmasına neden olmuştur. Sonuç olarak, TBES'te en fazla ilgisiz dokümana (195) ve ikinci sırada en fazla ilgili dokümana (tüm dokümanların %77'si) erişilmiştir. Erişilen ilgisiz doküman sayısı TBES'te duyarlılık performansına yansımış, bilgi erişim sistemleri arasında en düşük duyarlılık değeri olan 0,11'e inmiştir. Erişilen en fazla sayıda ilgisiz dokümandan ötürü normalize sıralama performansı bakımından da en kötü performansı sergileyen bilgi erişim sistemi TBES olmuştur. Sonuç olarak, araştırma hipotezlerinden “TBES'in normalize sıralama performansı ÜBES'ten ve KBES'ten düşüktür” hipotezi doğrulanmıştır.

- Bilgi erişim sistemleri arasında en fazla ilgili dokümana KBES'te erişilmiştir. KBES'te hem insana dayalı olarak üretilmiş üstverinin hem de dokümanın tam-metninin dizinlenmesi sorgu ve dokümanları temsil eden terimlerin çakışmasına neden olmuştur. Böylece, ÜBES'teki az ama bilgi erişim açısından yoğunluğa sahip terimlerle ve TBES'teki ayrık terimlerin

fazla olma avantajı kullanılarak ilgili dokümanların %88'ine erişilmiştir. KBES'in eriştiği ilgili doküman sayısı diğer bilgi erişim sistemlerinden fazla olmasına rağmen, eriştiği ilgisiz doküman sayısı (43) TBES'ten çok daha azdır. Bunun sebebi, sorgu terimlerinin dokümanı temsil eden terimlerle kısa olan üstveri alanları üzerinde çakışmasıdır. Kısa alanlar, doküman uzunluk normalizasyonunun kısa alanlara yüksek skorlar ataması sebebiyle ilgili dokümanları erişim kümesinin üst sırasına taşımaktadır. İlgili dokümanların erişim kümesinin üst sıralarında toplanması da hem duyarlılık hem de normalize sıralama performansını olumlu yönde etkilemektedir. Sonuç olarak, KBES hem en fazla ilgili dokümana erişebilen hem de en yüksek duyarlılığa sahip bilgi erişim sistemi olarak, araştırma hipotezlerinden “KBES'in duyarlılık performansı ÜBES'ten ve TBES'ten yüksektir” hipotezi doğrulanmıştır. KBES, normalize sıralama performansı bakımından da ÜBES'e en yakın bilgi erişim sistemi olmuştur.

- Araştırmada, ÜBES'in erişim noktalarının veya ayrık terim sayısının az olması sebebiyle erişilen doküman sayısı bakımında TBES'ten ve KBES'ten farklı olması sonucu beklenmiştir. Ancak, araştırma sonucunda üç bilgi erişim sisteminin döndürdüğü doküman sayıları aralarında istatistiksel açıdan anlamlı bir farka ulaşılamamış ve araştırma hipotezlerinden “Üç bilgi erişim sisteminin seçilen sorulara karşı eriştikleri doküman sayısı birbirinden farklıdır” hipotezi doğrulanamamıştır.

- Tam-metne dayalı otomatik dizinleme ile üstverinin birlikte kullanımı, sadece doğal dile dayalı otomatik dizinlemeye ve sadece üstveriye dayalı dizinlemeye göre daha yüksek duyarlılık performansı sağlamaktadır. Ayrıca, alanlara ayrılmış dizin ile oluşturulan bilgi erişim sistemi sadece doğal dile dayalı tam-metin alanına sahip bilgi erişim sistemine göre daha iyi bir performans sergilemektedir. Bu sonuç, dijital kütüphaneler üzerinde yapılan önemli bir çalışmanın (Gonçalves, Fox, Krowne, Calado, Laender, Silva ve diğerleri, 2004) sonuçlarıyla da örtüşmektedir.

- Tasarlanan üç bilgi erişim sisteminde de anma ve duyarlılık arasında güçlü bir negatif ilişki saptanmıştır. Yani, anma arttığında duyarlılık değeri düşmüştür. Bu sonuç, genel olarak TREC araştırmalarında elde edilen sonuçlarla örtüşmektedir.

- Geçmişte doğal dil ile ilgili yapılan tüm çalışmalarda karşılaşılan doküman ve sorgu temsili sorunlarıyla, yapılan çalışmada da karşılaşılmıştır. Çalışmada, kısaltmalarda ve açık terimlerde kullanıcının bilgi erişim sistemine yönelttiği terimlerle, bilgi erişim sisteminde dizinlenmiş olan eş anlamlı terimlerin çakışmaması erişim performansını olumsuz yönde etkilemiştir. Ayrıca, geniş terimlerle yapılan sorgunun ilgili ama dar terimlerle çakışmaması da bilgi erişim performansını olumsuz yönde etkilemiştir.

- Snowball gövdeleme algoritmasının performansı seçilen sorulardan sadece dört tanesinde erişim performansını etkilemiştir. Ancak, seçilen dokuz sorunun sorgu terimlerinin çıktıları göz önünde bulundurulduğunda gövdeleme algoritmasının ciddi sorunlarının olduğunu görülmektedir. Bunun sebebi, algoritmanın kök veya gövde sözlüğü kullanmamasından kaynaklanmaktadır.

Elde edilen bu sonuçlarla bağlantılı olarak şu öneriler sıralanabilir:

- Açık arşivler için geliştirilecek olan bilgi erişim sistemlerinde hem insana dayalı dizinlemenin hem de makineye dayalı dizinlemenin avantajlı yönlerinden faydalanılarak karma dizinleme yapılmalıdır.

- Bilgi ihtiyaçlarının ve dokümanların temsilinde kullanılan terimlerin daha sağlıklı çakışabilmesi için denetimli dil, eş anlamlılar sözlüğü, kısaltmalar sözlüğü ve eş anlamlı kısaltmalar sözlüğünün sorgu genişletmede kullanılması veya sorgu genişletme-daraltma seçeneklerinin kullanıcıya sunulması bilgi erişim performansının iyileşmesine neden olabilir.

- Kosinüs uzunluk normalizasyonunu kullanan alana dayalı bilgi erişim

sistemlerinde herhangi bir ağırlıklandırma şeması geliştirmeye gerek yoktur. Uygulama geliştiriciler, doküman uzunluk normalizasyonun kısa dokümanlara yüksek skor atamasından faydalanarak alan uzunluklarına göre dinamik bir ağırlıklandırma elde edebilir.

- Uygulama geliştiriciler, varsayılan arama alanı olarak tam-metin alanını kullanmamalıdırlar. Kullanıcı sorguları alındıktan sonra OR işleciyle bağlanarak tüm alanlar üzerinden aranmalıdır.

- Uygulama geliştiricilerin, sadece morfolojik analiz yapan gövdeleme algoritmaları yerine sözlükle birlikte morfolojik analiz yapan gövdeleme algoritmalarına yönelmeleri yararlı olabilir.

Ayrıca, Türkçe açık arşivlerde veya sayısal kütüphanelerde bilgi erişim konusuyla ilgili gelecekte yapılması gereken bazı çalışmalar önerilebilir. Bunlar şu şekilde sıralanabilir:

- *Türk Kütüphaneciliği* açık arşivinde yer alan PDF dokümanlarının analizinde makine öğrenim tekniklerinden faydalanılarak üstveri otomatik olarak çıkartılıp, insana dayalı olarak oluşturulan üstveri ile kalite açısından karşılaştırılabilir.

- Büyük bir derme üzerinde, Zemberek ve Snowball gibi kolay ulaşılabilir ve ticari kullanıma izin veren gövdeleme algoritmalarının bilgi erişim performansına etkisi karşılaştırılabilir.

- Uzun ve kısa doküman bakımından homojen bir derme üzerinde oluşturulmuş alanlara dayalı dizinlemede, alanlar üzerinde çeşitli ağırlıklandırma şemalarının kullanımı ile kosinüs uzunluk normalizasyonunun alanlar üzerinde oluşturduğu dinamik ağırlıklandırma performansı karşılaştırılabilir.

- Bu çalışmada kullanılan dermeyi meydana getiren dokümanların %85'inde öz bölümü bulunmamaktadır. Dolayısıyla, ayrık terim bakımından fakir bir üstveri kümesi ile performans değerlendirmesi

yapılmıştır. Bu çalışma, öz alanına sahip dokümanlardan oluşan bir derme üzerinde tekrarlanıp performans değerlendirmesi yeniden yapılabilir.

KAYNAKÇA

About Nutch: Overview. (2010). 20 Ekim 2010 tarinide http://nutch.apache.org/about.html adresinden erişildi.

Afzali, M. (2009). *Türkiye'de açık erişim, kurumsal arşivler ve akademik kütüphaneler.* Yayınlanmamış Doktora Tezi, H.Ü. Sosyal Bilimler Enstitüsü, Ankara.

Akın, A. A. ve Akın, M. D. (2010). *Zemberek, an open source NLP framework for Turkic languages.* 10 Ekim 2010 tarihinde http://zemberek.googlecode.com/files/zemberek_makale.pdf adresinden erişildi.

ANKOS. (2010a). *Açık erişim sözlük.* 9 Ekim 2010 tarihinde, http://acikerisim.ankos.gen.tr/sozluk.html adresinden erişildi.

ANKOS. (2010b). *Açık Erişim ve Kurumsal Arşivler Çalışma Grubu.* 9 Ekim 2010 tarihinde http://acikerisim.ankos.gen.tr/hakkimizda.html adresinden erişildi.

ARL. (2009). *ARL statistics 2007-08.* 5 Ekim 2010 tarihinde http://www.arl.org/bm~doc/arlstat08.pdf adresinden erişildi.

Arslan, A. ve Yilmazel, O. (2008). A comparison of relational databases and information retrieval libraries on Turkish text retrieval. *Natural Language Processing and Knowledge Engineering, 2008 (NLP-KE '08)* içinde (ss. 1-8). 10 Ekim 2010 tarihinde http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4906748 adresinden erişildi.

Atılgan, D., ve Bulut, B. (2008). Açık erişim olgusu ve Ankara Üniversitesi. *Balkan Ülkeleri Kütüphaneler Arası Bilgi ve Belge Yönetimi ve İşbirliği* içinde (ss. 92-100). Edirne: Trakya Üniversitesi Rektörlüğü Yayınları. 10 Ekim 2010 tarihinde http://eprints.rclis.org/bitstream/10760/12203/1/Ankara_Universitesi_Bildiri

_d%c3%bczeltilmis.pdf adresinden erişildi.

Atılgan, D. ve Yalçın, Y. (2009). Elektronik Kaynakların Seçimi ve Değerlendirilmesi. *Türk Kütüphaneciliği, 23*(4), 769-802. 5 Ekim 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2096/4146 adresinden erişildi.

Baeza-Yates, R. ve Ribeiro-Neto, B. A. N. (1999). *Modern information retrieval.* New York: ACM Press.

Baydur, G. (2010). Değişim ve bibliyografik denetim. *Türk Kütüphaneciliği, 24*(3), 526-532. 12 Kasım 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2225/4273 adresinden erişildi.

Bhattacharya, S. (2006). Metadata harvesting. 11 Ekim 2010 tarihinde http://ir.inflibnet.ac.in/dxml/bitstream/handle/1944/533/6%28cal %2006%29.pdf?sequence=1 adresinde erişildi.

BOAI (2002). *Budapest Open Access Initiative.* 9 Ekim 2010 tarihinde http://www.soros.org/openaccess/read.shtml adresinden erişildi.

Bollmann, P. (1983). The normalized recall and related measures. *Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '83)* içinde (ss. 122-128). New York, NY, USA: ACM.

Brown, E. W., Callan, J. P., ve Croft, W. B. (1994). Fast Incremental Indexing for Full-Text Information Retrieval. Jorge B. Bocca, Matthias Jarke, Carlo Zaniolo (Yay. Haz.). *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)* içinde (ss. 192-202). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Bush, V. (1945). *As we may think.* 10 Ekim 2010 tarihinde http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/ adresinden erişildi.

Byrne, D. J. (1998). *MARC manual: understanding and using MARC records.* Englewood, Colo: Libraries Unlimited.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Öcalan, H. Ç. ve Vursavas, O. M. (2008). Information retrieval on Turkish texts. *J. Am. Soc. Inf. Sci. Technol., 59*(3), 407-421.

Cen, R., Lui, Y., Zhang, M., Ru, L. ve Ma, S. (2009). Automatic search engine performance evaluation with the wisdom of crowds. *Information retrieval technology: 5th Asia Information Retrieval Symposium (AIRS 2009)* içinde (ss. 351-362). Berlin: Springer.

Cohen D., Amitay E. ve Carmel D. (2007). Lucene and Juru at Trec 2007: 1Million Queries Track. 20 Ekim 2010 tarihinde http://trec.nist.gov/pubs/trec16/papers/ibm-haifa.mq.final.pdf adresinden erişildi.

Cooper, W. S. (1988). Getting beyond Boole. *Information Processing and Management. 24*(3), 243-48.

Creative Commons nedir? (2010). 10 Ekim 2010 tarihinde http://tr.creativecommons.org/cc-hakkinda/ adresinden erişildi.

Çelik, A. (1987). Enformasyon teknolojisi ve kütüphanecilik. *Türk Kütüphaneciliği, 1*(3), 125-131. 10 Ekim 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1035/2069 adresinden erişildi.

Çilden, E. (2006). *Stemming Turkish words using Snowball.* 20 Ekim 2010 tarihinde http://snowball.tartarus.org/algorithms/turkish/ accompanying_paper.doc adresinden erişildi.

DC (2010). *Dublin Core metadata element set, version 1.1.* 12 Ekim 2010 tarihinde http://dublincore.org/documents/dces/#DCTERMS adresinden erişildi.

Dominich, S. (2001). *Mathematical foundations of information retrieval.* Dordrecht: Kluwer Academic Publishers.

Dominich, S. (2008). *The modern algebra of information retrieval.* Berlin: Springer.

Duran, G. (1997). *Gövdebul : Türkçe gövdeleme algoritması.* Yayımlanmamış Yüksek Mühendislik Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Eroğlu, M. (2000). *Gövdelemenin ve gömünün Türkçe bir bilgi erişim sistemi üzerindeki etkisinin araştırılması.* Yayımlanmamış Yüksek Mühendislik Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Ertürk, K. L. (2008). *Türkiye'de bilimsel iletişim: Bir açık erişim modeli önerisi.* Yayınlanmamış Doktora Tezi, H.Ü. Sosyal Bilimler Enstitüsü, Ankara.

Ertürk, K. L., ve Küçük, M. E. (2010). Bilimsel bilginin görünürlüğü: Hacettepe Üniversitesi'nde açık erişim farkındalığı. *Türk Kütüphaneciliği, 24*(1), 63-93. 10 Aralık 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2182 adresinden erişildi.

Europeana: About us. (2010). 11 Ekim 2010 tarihinde http://www.europeana.eu/portal/aboutus.html adresinden erişildi.

Gillies, D. F. (2010). *Lecture 1: An introduction to Boolean algebra.* 10 Ağustos 2010 tarihinde http://www.doc.ic.ac.uk/~dfg/hardware/HardwareLecture01.pdf adresinden erişildi.

Gonçalves, M. A., Fox, E. A., Krowne, A., Calado, P., Laender, A. H. F., Silva, A. S. ve diğerleri. (2004). The effectiveness of automatically structured queries in digital libraries. *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries (JCDL '04)* içinde (ss. 98-107). New York, NY, USA: ACM.

Gospodnetić, O., ve Hatcher, E. (2005). *Lucene in action.* Greenwich, CT: Manning Publications.

Göker, A. ve Davies, J. eds.(2008). *Information retrieval: Searching in the 21st century.* Chichester: Wiley.

Grossman, D. A. ve Frieder, O. (2004). *Information retrieval: Algorithms and heuristics.* Dordrecht: Springer.

Gürdal, O. ve Ertürk, K. L. (2002). Serüvende sayısal adımlar, "kısa bir öykü". *Türk Kütüphaneciliği, 16*(3), 329-344. 24 Kasım 2010 tarihinde, http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1784/3567 adresinden erişildi.

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez ve Rapor Yazım Yönergesi. (2004). 10 Ekim 2010 tarihinde http://www.sosyalbilimler.hacettepe.edu.tr/belgeler/Tez_ve_Rapor_Yazim_Yonergesi.pdf adresinden erişildi.

Hacettepe Üniversitesi Bilimsel Yayınlarında Kaynak Gösterme İlkeleri. (2006). 10 Ekim 2010 tarihinde http://www.sosyalbilimler.hacettepe.edu.tr/belgeler/bilimsel_yayinlarda_kaynak_gosterme_ilkeleri.pdf adresinden erişildi.

Hyde, R. (2005). *Write great code: Thinking low-level, writing high-level.* San Francisco, Calif: No Starch Press.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri.* Ankara: Tekışık Ofset.

Karasözen, B. (1996). Bilimsel bilgiye erişimde yeni paradigmalar ve internetin rolü. *Türk Kütüphaneciliği, 10*(3), 231-243. 5 Ekim 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1432/2864 adresinden erişildi.

Karasözen, B., Zan, B., ve Atılgan, D. (2010). Türkiye'de açık erişim ve bazı

ülkelerle karşılaştırma. *Türk Kütüphaneciliği, 24*(2), 235-257. 10 Aralık 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2200 adresinden erişildi.

Kayaoğlu, H. (2006). Açık erişim kavramı ve gelişmekte olan bir ülke olarak Türkiye için anlamı. *Türk Kütüphaneciliği, 20*(1), 29-60. 9 Ekim 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1995/3969 adresinden erişildi.

Khan, M. A. (1997). *Cataloguing in library science.* New Delhi: Sarup & Sons.

Koch, T. W. (1913). *The library assistants manual.* Michigan: Y.Y.

Kowalski, G. ve Maybury, T. (1998). *Information retrieval systems: Theory and implementation.* Boston, Mass: Kluwer Academic.

Kurbanoğlu, S. (2004). *Kaynak gösterme el kitabı.* Ankara: ÜNAK.

Küçük, M. E. ve , Al, U. (2003) Üst veri standartları ve uygulamaları. *Hacettepe Üniversitesi Edebiyat Fakültesi Dergisi, 20*(1), 167-185.

Küçük, M. E., Al, U. ve Olcay, N. E. (2008). Türkiye'de Bilimsel Elektronik Dergiler. *Türk Kütüphaneciliği, 22*(3), 308-319. 4 Ekim 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2064/4091 adresinden erişildi.

Lancaster, F. W. (1995). The evolution of electronic publishing. *Library Trends*, *43*(4), 518-27.

Lucid Imagination. (2010). *Library/Catalog case study: Europeana - bringing European culture online.* 20 Ekim 2010 tarihinde http://www.lucidimagination.com/Community/Marketplace/Business-Use-Case-Studies/Europeana adresinden erişildi.

Luhn, H. P. (1957). A statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development, 1,* 309-317.

Manning, C.D., Raghavan, P. ve Schütze, H. (2008). Evaluation of ranked retrieval results. 20 aralık 2010 tarihinde http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html adresinden erişildi.

Marcus, R. (1991). Computer and human understanding in intelligent retrieval assistance. *Proceedings of the 54th American Society for Information Science meeting* içinde (ss. 49-59), Washington: Medford.

McCansless, M., Gospodnetić, O. ve Hatcher, E. (2010). *Manning: Lucene in action.* Greenwich, CT: Manning Publications.

Meadow, C. T., Boyce, B. R. ve Kraft, D. H. (2007). *Text information retrieval systems.* Amsterdam: Academic.

NISO (2004). *Understanding Metadata.* 11 Ekim tarihinde http://www.niso.org/publications/press/UnderstandingMetadata.pdf adresinden erişildi.

OAI (2008). *ORE User Guide – Primer.* 20 Ekim 2010 tarihinde http://www.openarchives.org/ore/1.0/primer adresinden erişildi.

OAIS (2002). *Reference model for an open archival information system.* 10 Ekim 2010 tarihinde http://public.ccsds.org/publications/archive/650x0b1.pdf adresinden erişildi.

Özer, O. (1998). Soyut Matematik. 10 Aralık 2009 tarihinde http://www.aof.anadolu.edu.tr/kitap/IOLTP/2287/unite02.pdf adresinden erişildi.

Parks, R. (2002). The faustian grip of academic publishing. *Journal of Economic Methodology, 9*(3), 317-335.

Polat, C. (2006). Bilimsel Bilgiye Açık Erişim ve Kurumsal Açık Erişim Arşivleri. *A.Ü. Fen Edebiyat Fakültesi Sosyal Bilimler Dergisi, 6* (37), 53-80.

Rainardi, V. (2008). *Building a data warehouse: With examples in SQL Server.* Berkeley, CA: Apress.

Rob, P. ve Coronel, C. (2009). *Database systems: Design, implementation, and management.* Boston, Mass: Course Technology.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. ve Gatford, M. (1995). Okapi at TREC–3. D. K. Harman (Yay. Haz.). *Proceedings of the Third Text REtrieval Conference (TREC–3)* içinde (ss. 109-126). Gaithersburg, MD: NIST.

Salton, G. (1984). The use of extended Boolean logic in information retrieval. *Proceedings of the 1984 ACM SIGMOD international conference on Management of data (SIGMOD '84)* içinde (ss. 277-285). New York, NY, USA: ACM.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Commun. ACM. 29*(7), 648-656.

Salton, G. ve Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage., 24*(5), 513-523.

Salton, G., Fox, E. A. ve Wu, H. (1982). *Extended Boolean information retrieval.* 1 Kasım 2010 tarinide http://ecommons.library.cornell.edu/handle/1813/6351 adresinden erişildi.

Salton, G., Fox, E. A. ve Wu, H. (1983). Extended boolean information retrieval. *Commun. ACM, 26*(11), 1022-1036.

Salton, G., Wong, A. ve Yang, C. S. (1975). A Vector Space Model for information retrieval. *Journal of the American Society for Information Science, 18*(11): 613-620.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. Edward A. Fox, Peter Ingwersen, Raya Fidel (Yay. Haz.). *Proceedings of the 18th annual international ACM SIGIR conference on Research and*

development in information retrieval (SIGIR '95) içinde (ss. 138-146). New York, NY, USA: ACM.

Sezer, E. (1999). *Smart Bilgi Erişim Sistem'nin Türkçe yerelleştirilmesi ve otomatik gömü üretimi.* Yayımlanmamış Yüksek Mühendislik Tezi, H.Ü. Fen Bilimleri Enstitüsü, Ankara.

Shannon, C. E. (1959). *Prediction and entropy of printed English. Bell Sys. Tech. J. Ocak,* 50-64. 12 Kasım 2010 tarihinde http://hum.uchicago.edu/~jagoldsm/CompLing/shannon-1951.pdf adresinden erişildi.

Shields, G. (2005). *What are the main differences between human indexing and automatic indexing?.* 12 Ekim 2010 tarihinde http://www.shieldsnetwork.com/LI842_Shields_Automatic_Indexing.pdf adresinden erişildi.

Singhal , A., Buckley, C. ve Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96)* içinde (ss. 21-29). New York, NY, USA: ACM.

Singhal, A., Salton, G., Mitra, M. ve Buckley, C. (1995). *Document Length Normalization.* 20 Ekim 2010 tarihinde http://ecommons.library.cornell.edu/handle/1813/7186 adresinden erişildi.

Smiley, D. ve Pugh, E. (2009). *Solr 1.4 enterprise search server.* Birmingham: Packt Publ.

Solrmarc: Indroduction. (2010). 20 Ekim 2010 tarihinde http://code.google.com/p/solrmarc/wiki/SolrMarc adresinden erişildi.

Soydal, İ. (2000). *Web arama motorlarında performans değerlendirmesi.* Yayımlanmamış Yüksek Lisans Tezi, H.Ü. Sosyal Bilimler Enstitüsü, Ankara.

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. J*ournal of Documentation, 28*(1), 11-21.

Spoerri, A. (1995). *INFOCRYSTAL: A visual tool for information retrieval.* Yayınlanmamış Doktora Tezi, Massachusetts Institute of Technology. 13.01.2010 tarihinde http://hdl.handle.net/1721.1/36946 adresinden erişildi.

Steele, C. (2005). *The Library's perspective on scholarly publishing in the twenty-first century.* 9 Ekim 2010 tarihinde http://dspace-prod1.anu.edu.au/handle/1885/42610 adresinden erişildi.

Tonta, Y. (1997). Elektronik yayıncılık, bilimsel iletişim ve kütüphaneler. *Türk Kütüphaneciliği, 11*(4), 305-314. 4 Ekim 2010 tarihinde, http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1474/2948 adresinden erişildi.

Tonta, Y. (2006). Açık erişim: Bilimsel iletişim ve sosyal bilimlerde süreli yayıncılık üzerine etkileri. Kasım Karakütük (Yay. Haz.). *Sosyal Bilimlerde Yayıncılık 1. Ulusal Kurultay Bildirileri* içinde (ss. 23-32). Ankara: TÜBİTAK ULAKBİM.

Tonta, Y. (2007). Kütüphaneler sanal güzergahlara mı dönüşüyor? Ayşe Üstün ve Ümit Konya (Yay. Haz.). *I. Uluslararası Bilgi Hizmetleri Sempozyumu: İletişim, 25-26 Mayıs 2006, İstanbul* içinde (ss. 353-366). İstanbul: Türk Kütüphaneciler Derneği İstanbul Şubesi. 10 Aralık 2010 tarihinde http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-istanbul-mayis-2006-bildiri.pdf adresinden erişildi.

Tonta, Y. (2010a). Açık erişim ve tıpta bilimsel iletişimin geleceği. Hamdi Akan (Yay. Haz.). *Bilimsel yayınlar kitabı* içinde (ss. 225-235). Ankara: Bilimsel Tıp Yayınevi. 6 Ekim 2010 tarihinde http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-tipta-acik-erisim.pdf adresinden erişildi.

Tonta, Y. (2010b). Elektronik yayıncılık ve elektronik bilgi kaynakları. 5 Ekim 2010 tarihinde http://yunus.hacettepe.edu.tr/~tonta/courses/fall2002/kut655/02-e-yayincilik-e-bilgi-kaynaklari.pdf adresinden erişildi.

Tonta, Y., Bitirim, Y. ve Sever, H. (2002). *Türkçe Arama Motorlarında Performans Değerlendirme.* Ankara: Total Bilişim Ltd. Şti.

Tonta, Y., Küçük M. E., Al, U., Alır, G., Ertürk, K. L., Olcay, N. E., ve diğerleri. (2006). *Hacettepe Üniversitesi Elektronik Tez Projesi: Yüksek lisans, doktora ve sanatta yeterlik tezlerinin dijitalleştirilmesi ve tam metinlerinin internet aracılığıyla erişime açılması.* Ankara: Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü. 5 Ekim 2010 tarihinde http://www.bby.hacettepe.edu.tr/02G064.pdf adresinden erişildi.

Tonta, Y., Ünal, Y. ve Al, U. (2007). The research impact of open access journal articles. Leslie Chan, Bob Martens (Yay. Haz.). *ELPUB2007. Openness in Digital Publishing: Awareness, Discovery and Access - Proceedings of the 11th International Conference on Electronic Publishing held in Vienna, Austria 13-15 June 2007* içinde (ss. 321-330). 6 Ekim 2010 tarihinde http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-unal-al-elpub2007.pdf adresinden erişildi.

Tonta, Y. ve Küçük, M. E. (2005). Sanayi toplumundan bilgi toplumuna geçiş sürecinde temel dinamikler. *Türk Kütüphaneciliği, 19*(4), 449-464. 4 Ekim 2010 tarihinde, http://tk.kutuphaneci.org.tr/index.php/tk/article/view/1920/3835 adresinden erişildi.

Toplu Katalog – Hakkında. (2010). 20 Ekim 2010 tarihinde http://www.toplukatalog.gov.tr/index.php?cwid=8 adresinden erişildi.

Townsend, J. J., Riz, D. ve Schaffer, D. (2004). *Building portals, intranets, and corporate web sites using Microsoft servers.* Boston, Mass.: Addison-Wesley.

Turtle, H. R. ve Croft, W. B. (1997). Uncertainty in information retrieval systems. A. Motro, P. Smets (Yay. Haz.). *Uncertainty management in information systems: From needs to solutions* içinde (ss. 189-224). Boston: Kluwer Academic.

Turtle, H. ve Croft, W. B. (1989). Inference networks for document retrieval. Jean-Luc Vidick (Yay. Haz.). *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '90)* içinde (ss. 1-24). New York, NY, USA: ACM.

Uçak, N. Ö. (2000). Bilgi üzerine kuramsal bir yaklaşım. *Bilgi Dünyası, 1*(1), 143-159.

Uçak, N. Ö. (2010). Bilgi: Çok Yüzlü Bir Kavram. *Türk Kütüphaneciliği, 24*(4), 705-722. 30 Aralık 2010 tarihinde http://tk.kutuphaneci.org.tr/index.php/tk/article/view/2252/4293 adresinden erişildi.

Van de Sompel, H. ve Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine, 6*(2). 10 Ekim 2010 tarihinde http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html adresinden erişildi.

Van Rijsbergen, C. J. (1979). Information retrieaval: Introduction. 20 Ekim 2010 tarihinde http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html adresinden erişildi.

w3schools (2010). *Introduction to XML.* 10 Ekim 2010 tarihinde, http://www.w3schools.com/xml/xml_whatis.asp adresinden erişildi.

Wedgeworth, R. (1993). *World encyclopedia of library and information services.* Chicago: American Library Association.

Weibel, S., Godby, J., Miller, E. ve Daniel, R. (1995). *OCLC/NCSA metadata workshop report.* 11 Ekim 2010 tarihinde http://dublincore.org/workshops/dc1/report.shtml adresinden erişildi.

White, T. (2009). *Hadoop: The definitive guide.* CA: O'Reilly.

Yamaç, K. (2009). *Bilgi toplumu ve üniversiteler.* Ankara: Eflatun Yayınevi.

Yao, Y. Y. (1995). Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science, 46*(2), 133-145.

Yao, Y. Y. (2004). Granular computing for the design of information support systems. W. Wu, H. Xiong, S. Shekhar (Yay. Haz.). *Clustering and Information Retrieval* içinde (ss. 299-329). Dordrecht: Kluwer Academic Publishers.