Int. J. Adv. Eng. Pure Sci. 2023, 35(1): 406-417
DOI: 10.7240/jeps.1174193

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Web Tarayıcıları için Etkili Tohum URL Seçimi ve Kapsam Genişletme Algoritması

Effective Seed URL Selection and Scope Extension Algorithm for Web Crawler

Zülfü ALNAOĞLU[1], M.Ali AKCAYOL[2]

[1] *Hatay Mustafa Kemal Üniversitesi, Antakya MYO, Bilişim Teknolojileri Bölümü, Hatay, Türkiye*
[2] *Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Ankara, Türkiye*

Öz
Web, hızla büyüyen ve her türden verilerin bulunduğu devasa bir veri kaynağıdır. Kullanıcılar bu veri kaynağından istedikleri verileri almak için arama motorlarını kullanırlar. Arama motorları bu verileri web tarayıcıları ile elde ederler. Web tarayıcıları web sayfalarındaki tek düzen kaynak bulucuları (URL-Uniform Resource Locator) izleyerek ulaştıkları tüm sayfalardaki verileri alır, ayrıştırır ve indekslerler. Web tarama sürecindeki en önemli konular hangi URL'lerden başlanacağı ve taramanın kapsamıdır. Bu yazıda kapsamı tüm web olan genel bir tarayıcının tohum URL seçim ve kapsam genişletme yöntemleri sunulmuştur. Tohum URL seçiminde 102 farklı ülkede ziyaretçinin günlük harcadığı saat, ziyaretçi başına günlük sayfa görüntüleme sayısı, aramadan gelen trafiğin yüzdesi ve toplam bağlı site sayısı temel alınarak oluşturulmuş üç farklı tohum URL seti oluşturulup detaylı bir şekilde performansları analiz edilmiştir. Ayrıca kapsamı hızlı bir şekilde genişletmek için link skoruna dayalı yeni bir tarama algoritması önerilmiş, tohum URL setleri kullanılarak taramalar yapılmış, karşılaştırılmış ve detaylı analizleri yapılmıştır.
Anahtar Kelimeler: Web Tarayıcıları, Tohum URL Seçimi, Kapsam Genişletme, Link Skoru Hesaplama

Abstract
The web is a huge data source which is rapidly growing and which keeps all kinds of data. Users use search engines to get the data they want from this data source. Search engines obtain these data through web crawlers. Web crawlers retrieve, parse, and index data on all pages they reach by tracking uniform resource locators (URL) on web pages. The most important issues in the web crawling process are which URLs to start from, and the scope of the crawl. In this study, seed URL selection and scope expansion methods of a general web crawler were presented. In the selection of seed URLs, three different seed URL sets were created based on the daily hours spent by the visitors in 102 different countries, the number of daily page views per visitor, the percentage of traffic from the search, and the total number of affiliate sites, and their performance was analyzed thoroughly. Furthermore, a new search algorithm based on link score was proposed to expand the scope quickly, searches were made, compared, and detailed analyzes were performed using seed URL sets.
Keywords: Web Crawler, Seed URL Selection, Scope extension, Link score calculation

I. GİRİŞ

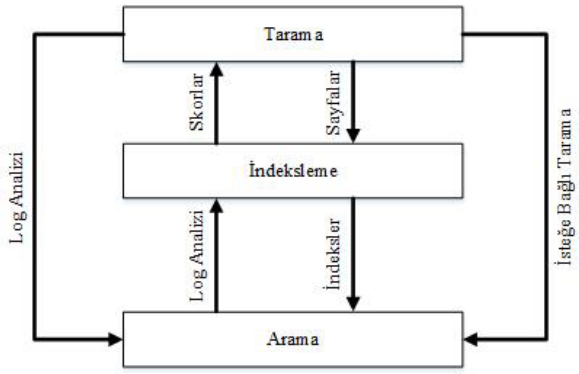
Günümüzde interneti kullanarak Web üzerindeki verilere erişmek hayatımızın önemli bir parçası haline gelmiştir. Şuanda mevcut dünya nüfusu 7.9 milyar olup 5.2 milyar (%66.2) internet kullanıcısı mevcuttur [1]. Bu sayı 2012'de 2.4 milyar [2] iken 2022'de 5.2 milyara yükselmiş, yani yaklaşık %116 artmıştır.

İnternet kullanıcı sayısının artması Web' deki veri miktarının artması anlamına gelmektedir. Web her geçen gün hızla büyüyen ve her türden verilerin bulunduğu devasa bir veri deposudur. Bu veri deposu içinde istenilen verilere doğru bir şekilde ve zamanında ulaşmak, günümüz koşullarında hayati öneme sahip olup her geçen gün zorlaşmaktadır [3]. Bu zorlukları aşabilmek, Web' deki verileri aramak ve istenilen veriye ulaşmak için arama motorları kullanılmaktadır. Arama motorlarını kullanmadan web üzerindeki bilgileri elde etmek için, milyarlarca web sayfasını tek tek ziyaret etmek gerekir. Bu durumda veriye ulaşmak neredeyse imkânsız hale gelmektedir.

Birçok tarayıcı türü olmasına rağmen temelde genel ve odaklı olmak üzere iki ana tarayıcı türü vardır. Odaklı tarayıcılar belirli bir konu ya da alan ile ilgili sayfaları tarama eğilimindedirler. Önceden belirlenmiş bazı verileri kullanarak erişim sayısını sınırlarlar [4]. Genel tarayıcılarda ise böyle bir sınırlama yoktur. Arasu ve arkadaşları [5], yaptıkları çalışmada, odaklı taramanın değişen ön bilgilere göre kullanıcılar arasında değişiklik gösterdiğini vurgulanmış ve gerçek hayatta genel tarayıcıların kullanımının daha önemli olduğunu özellikle belirtmişlerdir.

Sorumlu Yazar: Zülfü ALANOĞLU, **Tel:** +90 505 616 65 60, **e-posta:** zalanoglu@mku.edu.tr
Gönderilme: 13.09.2022**, Düzenleme:** 29.12.2022**, Kabul:** 01.01.2023

Arama motorları temel olarak 3 bölümden oluşurlar. Bu bölümler sırası ile web sayfalarını tarama, verileri indeksleme ve bu veriler içerisinde aramadır[6]. Web sayfalarını tarama işlemini web tarayıcıları (örümcek, tarama botları vb.) gerçekleştirir. Web tarayıcıları tarama işlemine tohum (başlangıç) URL'ler ile başlar. Ziyaret edilen web sayfası içindeki veriler alınıp indekslenir ve sayfa içindeki diğer URL'ler çıkarılarak öncelikli kuyruğa eklenir. Sırası gelen URL taranır ve veriler indekslenir. Taranan URL'ler belirli politikalara göre tekrar taranarak güncellik sağlanır. Şekil 1' de döngüsel yapı kullanarak veri tabanını güncel tutan bir tarayıcı mimarisi gösterilmiştir [6].



Şekil 1. Tarayıcı Mimarisi

Web' in tamamının taranması devasa hacmi nedeniyle uzun bir süreçtir ve neredeyse imkânsızdır. Mevcut arama motorları (Google, Baidu, Yahoo vd.) tüm Web 'in sadece yaklaşık olarak %5 ini tarayabilmektedir [7]. Bu nedenle arama motorlarının en büyük yetersizliklerinin başında kapsam genişliği gelmektedir. Kapsamı genişletmek için özellikle ticari arama motorları birçok algoritma ve özellik (çoğu gizli) kullanmaktadırlar. Kapsamı genişletmenin ilk adımı tohum URL seçimi ve ikinci adımı da iyi bir tarama algoritmasıdır.

II. LİTERATÜR ARAŞTIRMASI

Literatürde yapılan çalışmaları tohum URL seçimi ile ilgili çalışmalar ve kapsam genişletme metotları ile ilgili çalışmalar olmak üzere iki gruba ayırıyoruz.

1.1 Tohum URL Seçimi

Bir web tarayıcısının en önemli sorunlarından biri en uygun sayfaları elde etmek için hangi URL' lerden başlaması gerektiğidir [8]. Bu başlangıç URL seti, tarayıcıların arama işlemlerini başlattığı giriş noktasıdır. Tohum URL'lerin kalitesi tarayıcının performansını ve kapsamını etkileyen en temel özelliklerden biridir. Web' in dinamik yapısı gereğince en iyi tohum URL' lerin de zaman ile değişim göstermeleri kaçınılmazdır.

Daneshpajouh ve arkadaşları [9], farklı topluluklardan tohum URL' leri tanımlayan ve çıkaran ilk tohum çıkarma algoritmasını önermişlerdir. Algoritma, tohum URL'lerin farklı topluluklardan düğümler içermesini garanti etmek için seçilen tohum URL' ler arasındaki mesafeyi ölçmektedir. Kleinberg [10], web sayfalarını merkez ve otorite adını verdiği iki ana grupta toplayan HITS algoritmasını önermiştir. HIST algoritması yalnızca web sayfaları arasından köprüleri dikkate almaktadır. Bundan dolayı en iyi merkez web sayfalarının en iyi tohum URL olduğu söylenebilir. Zheng ve arkadaşları [11], tohum URL seçimi için rastgele, en yüksek PageRank değeri ve en çok alan dışı bağlantıya sahip k sayfaya dayalı tohum seçim stratejilerini kullanan grafik tabanlı bir yaklaşım önermişlerdir. Önerilen yaklaşımın performansını değerlendirmek için her biri en az 100 sayfa içeren ve en az bir harici bağlantıya sahip olan 2000 web sitesinden rastgele örnekler seçerek tohum URL setini oluşturmuşlardır.

Nwala ve arkadaşları [12], web arşivi koleksiyonları için sosyal medya gönderilerinden tohum URL elde etmişlerdir. Bu URL' lere 10 ana boyutta (popülerlik, coğrafik, konu uzmanı, güvenilirlik, itibar vs.) bir kalite puanı atamışlardır. Toplamda referans koleksiyonlarından 1552 ve Twitter Mikro koleksiyonlarından 4.209 tweet' den 2.027 tohum URL elde etmişlerdir.

Tohum URL seçiminde, uzman tarafından manuel seçim, yarı otomatik seçim ve otomatik seçim olmak üzere üç temel seçim metodu vardır. Tarayıcılar için tohum URL çıkarma genellikle manuel yapılır. Manuel seçimde [13-15] bir veya birkaç konu hakkında yapılan taramalarda konunun uzmanları tarafından tohum URL' ler seçilmekte, önceliklendirilmekte ve tarama kuyruğuna eklenmektedir. Yarı otomatik seçimde DMOZ ve curlie.org gibi açık kaynak dizinlerden, belirli özelliklere göre tohum URL' ler seçilmektedir. Chan ve Yamana [16], DMOZ üzerinde bulunan URL' leri belirli alan adları (.com, .net, cn, .tw, .jp, .kr) ve dillere göre (Çince, Japonca ve Korece) ayıklandıktan sonra tohum URL olarak almışlardır. Menczer ve Monge [17], InfoSpider adını verdikleri bir tarayıcı geliştirmişlerdir. InfoSpider, kullanıcı sorgularını genel bir arama motoruna göndermekte ve sonuç olarak dönen URL'leri tohum URL olarak kullanmaktadır. Son olarak otomatik seçimde Twitter gibi sosyal medyadaki kullanıcıların paylaştıkları URL' ler tohum URL olarak alınmaktadır. Priyatam ve arkadaşları [18], Twitter' da paylaşılan URL' lerin her biri köşe noktası olan bir grafik oluşturmuş ve benzer köşeleri birbirine bağlamışlardır. Önerilen tarayıcı taramaya köşe noktalarını oluşturan ve benzersiz olan URL' ler ile başlamıştır. Sanagavarapu ve arkadaşları [19], Wikipedia ve Twitter' ı kullanarak tohum URL' lerin otomatik olarak çıkarılması için puanlama (SeedRel) metriği ve URL' lerin alaka düzeyini belirlemek için çeşitlilik indeksi kullanan bir yaklaşım önermişlerdir. Buna ek olarak Sanagavarapu ve ark. [20], tohum ve alt URL'lerin tanımlanması ve puanlanması için yapay arı kolonisi (Artificial Bee Colony - ABC) algoritmasını önermişlerdir. Önerilen algoritmayı güvenlik alanına

uygulamış ve Wikipedia üzerinden 34.007 tohum URL' i çıkarmışlardır.

1.2 Kapsam Genişletme

Bir arama motorunun performans ölçütlerinden biri de taradığı ve indekslediği web sayfalarının miktarıdır. Arama motorunun, web tarayıcısı ve tarayıcının kullandığı tarama tekniklerine bağlı olarak kapsamı ve etkinliği artar [21]. Web' in grafik yapısı karmaşıktır ve içerik ile köprülere erişmek için verimli bir algoritma gerekir. Kapsam, öncelikle seçilen tohum URL' lere ve ardından taramanın genişleyebilmesi için URL' lerin öncelik durumuna göre sıralanmasına bağlıdır. Web tarayıcıları için web sayfalarının önemleri farklıdır ve hangi sayfaların önce taranması gerektiği ile ilgili çeşitli algoritmalar geliştirilmiştir [22].

Page ve arkadaşları [23], Google arama motorunun temel algoritmalarından biri olan PageRank algoritmasını önermişlerdir. Hangi web sayfalarının öncelikli taranması gerektiğini belirlemek için farklı metrikler mışlardır. Bu metrikler genişlik öncelikli, geri bağlantı sayısı ve PageRank olup tarayıcıyı önemli sayfalara yönlendirme açısından bu üç metrik karşılaştırılmıştır. Sonuç olarak PageRank' in diğerlerine göre önemli sayfaları daha erken taradığı görülmüştür. Prakash ve Kumar [24], çalışmalarında PageRank ve köpekbalığı aramasının (Shark-Search) geliştirilmiş bir versiyonu olan PageRank algoritmasını kullanan köpekbalığı algoritmasını önermişlerdir. Yapılan ön deneylerde orijinal sayfa sıralaması algoritmasına göre önemli gelişmelerin gösterildiği belirtilmiştir. Cao ve arkadaşları [25], çalışmalarında aynı ağda rastgele yürüyüşe izin veren RankCompede adını verdikleri yeni bir model önermişlerdir. Rastgele yürüyüş ile rekabet kavramlarını birleştirerek kümeleme ve sıralama işlemlerini aynı anda yerine getiren bir yöntem geliştirmişlerdir. Geleneksel grafik kümeleme yaklaşımları ile karşılaştırılmış ve yöntemin ağ düğümlerini gruplamada daha hızlı ve sezgisel olduğu belirtilmiştir. Najork ve Wiener [26] önerdikleri yöntemde 328 milyon benzersiz sayfa içeren bir tarama sırasında, taranan sayfaların zaman içerisindeki ortalama sayfa kalitesini incelemişlerdir. Taramaya başlandığında yüksek kalitedeki web sayfalarını seçme eğiliminde olan genişlikte ilk arama yöntemini seçmişlerdir. Sayfaların kalitesini ölçmek ve sıralamak için PageRank algoritmasını kullanmışlardır.

Nisreen ve Elsheh çalışmalarında [22], web sayfalarında benzerliği ve dinamikliği kullanarak web sayfalarına öncelik veren bir yöntem önermişlerdir. Çalışmada dinamik ve statik URL'ler için ayrı iki öncelikli kuyruk kullanılmıştır. Dinamik sayfalar yüksek önceliğe sahip olduğundan dolayı öncelikli taranmaktadır. Elde edilen bulgulara göre web sayfalarının dinamikliğini kullanmanın, URL' lerin taranma sırasının belirlenmesinde etkili bir yol olduğu belirtilmiştir.

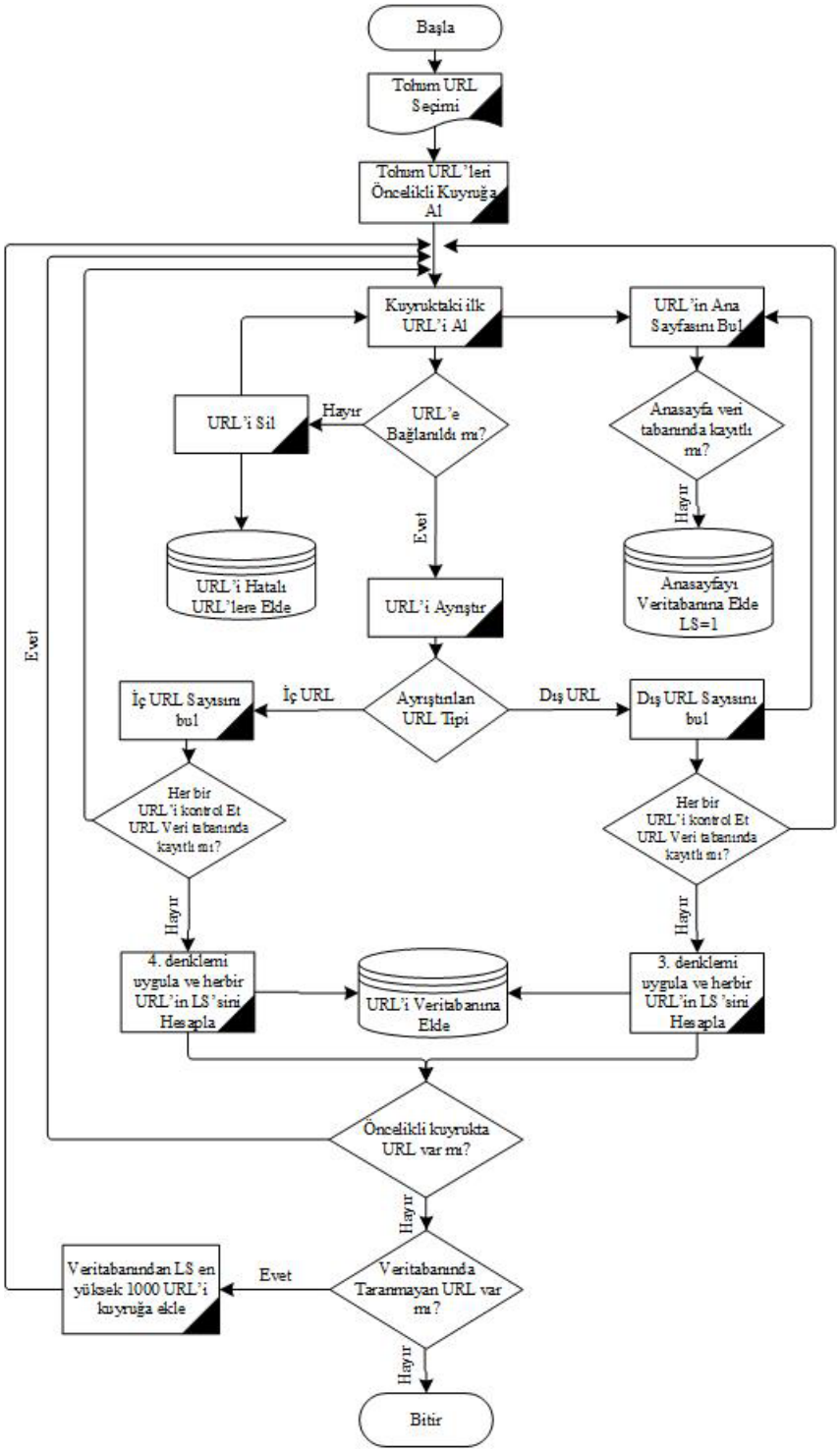
Gupta ve Singh [27] yaptıkları çalışmada kullanıcı tercihi tabanlı sayfa sıralaması adını verdikleri yeni bir sayfa sıralama algoritması önermişlerdir. Önerilen algoritmanın sayfa içerik alaka düzeyini belirlemek için aracılar kullandığı ve sıralamada kullanıcı davranışlarının da dikkate alındığı vurgulanmıştır. Yine kullanıcı davranışı ile ilgili olarak Alhaidari ve arkadaşları [28], kullanıcı davranışı ve tercihine odaklanan sayfa sıralama algoritmaları, PageRank, Ağırlıklı PageRank ve köprü kaynaklı konu arama algoritmalarını tartışmışlardır. Bunlara ek olarak algoritmaların birleşiminden oluşan kullanıcı tercihi tabanlı ağırlıklı sayfa sıralama algoritmasını (User Preference Based Weighted Page Ranking Algorithm - UPWPR) önermişlerdir. UPWPR algoritması, web içerik madenciliği ve web kullanım madenciliğini kullanarak kullanıcı tercihlerine göre arama sonuçlarını sıralamaktadır.

Baker ve Akcayol [29], URL'leri alan içi ve alan dışı olarak sınıflandıran ve bir öncelik sırası kullanan tarayıcı algoritmasını sunmuşlardır. Önerilen algoritma alan dışı bağlantılar için 2/3 ve alan içi bağlantılar için 1/3 değerini vererek alan dışı bağlantılara öncelik vermektedir. Bunun sebebi kapsamı genişletmek ve aynı etki alanı içindeki bağlantı döngülerinden kaçınmaktır. Öncelikli kuyruğa eklenen URL' ler için bir zamanlama mekanizması kullanılmış ve belirli bir süre boyunca taranmayan URL' ler kuyruktan çıkartılmıştır.

III. ÖNERİLEN WEB TARAYICISI

Bu çalışmada önerdiğimiz web tarayıcısı genel bir tarayıcı olduğu için kapsamı tüm webdir. Genel tarayıcı sırası gelen URL' leri ayrıştırarak web üzerindeki tüm URL' lere ulaşmaya çalışır. Belirli dil ve bölgelere bağlı kalmamak için tüm dünya ülkelerinden en popüler ve en çok ziyaret edilen web sayfalarından tohum URL seti oluşturulmuştur. Önerilen algoritmaya göre tohum URL seti öncelikli kuyruğa alınarak her bir URL sırasıyla taranmıştır. Taranan URL' e herhangi bir sebepten dolayı ulaşılamadığında hata kodu, hata mesajı ve hatalı URL veri tabanında ilgili tabloya kaydedilerek URL' lerden silinmiştir. URL' e bağlantı sağlandığında iki işlem gerçekleştirilir. (1) URL'in ana sayfası bulunarak ilgili tabloya kaydedilir ve link skoruna (LS) en yüksek değer olan 1 değeri verilir. (2) Sayfa içerisinde ayrıştırılan URL' ler alan içi (IntraDomain) ve alan dışı (InterDomain) olmak üzere iki gruba ayrılır ve sayıları bulunur. Daha sonra veri tabanından kontrol edilerek daha önce kaydedilmeyen URL' lerin denklem (3) ve (4) ' e göre link skoru hesaplanır. Bu işlemin sonunda URL' ler, link skorları ve tarama zaman damgası veri tabanına kaydedilir. Bu işlemler kuyrukta URL olduğu sürece tekrarlanır. Kuyruktaki son URL tarandıktan sonra veri tabanındaki en yüksek link skoruna sahip bin URL öncelikli

kuyruğa alınarak keşfedilen tüm URL' ler bitene kadar bu işlemler tekrarlanır. Geliştirilen web tarayıcısının akış şeması Şekil 2' de verilmiştir.



Şekil 2. Geliştirilen Web Tarayıcı Algoritmasının Akış Şeması

3.1. Tohum URL Seçim Metodu

Geliştirilen tarayıcıda tohum URL' ler Alexa [30] üzerinden alınmıştır. Alexa bir Amazon şirketidir ve içerik araştırması, rekabet analizi, anahtar kelime araştırması için başvuru kaynağı olarak kullanılmaktadır. Alexa web sitelerini; ziyaretçinin günlük harcadığı saat, ziyaretçi başına günlük sayfa görüntüleme sayısı, aramadan gelen trafiğin yüzdesi ve toplam bağlı site sayısı temel alarak sıralamıştır. Bu çalışmada 102 faklı ülkenin her birinde sıralamaya giren 50 web sayfası alınmıştır. Sayfa üzerinden veri çekme işlemi için Python programlama dili kütüphanesi olan BeautifulSoup kullanılmıştır. Alexa' dan toplam 5079 web sitesi alınmış ve kayıt edilmiştir.

Google, Youtube, Facebook vb. gibi web siteleri neredeyse tüm ülkelerde ilk sıralarda yer almakta ve veri tabanında tekrar etmektedir. Bu web sayfalarından benzersiz olan 2502 web sayfası tespit edilmiştir. Sadece 1 ülkede listeye giren 1920 web sayfası tohum URL olarak kullanılmıştır.

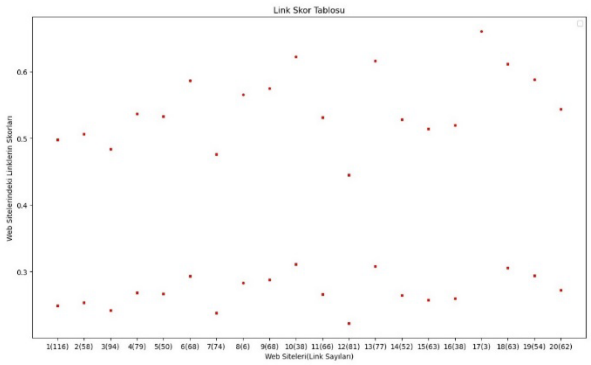
3.2. Geliştirilen Kapsam Genişletme Modülü

Tüm webin taranması olası olmadığından, temel amacımız hem lokasyon olarak hem de taranmamış web sayfalarına öncelik vererek kapsamı genişletmektir. Kapsamı genişletmek için tohum URL seçimi ile birlikte tarama yöntemi hayati derecede önemlidir. Bir web sitesinin ana omurgasının ve zengin URL içeriğinin bulunduğu sayfa ana sayfa (domain adresleri) olduğu görülmüştür. Web sayfası içerisindeki URL ağaç yapısının en üstünde ana sayfa bulunmaktadır. Bunun için ana sayfaların öncelikli olarak taranması gerekmektedir. Bir URL kuyruktan alındığında ve URL' e bağlantı sağlandığında sayfadaki her bir InterDomain için ayrıştırma işlemi gerçekleştirerek URL' in ana sayfasının veri tabanında olup olmadığı sorgulanır. Eğer ana sayfa veri tabanında kayıtlı değil ise öncelikli taranması için link skoruna (LS) en büyük değer olan 1 değeri verilir. Diğer URL' ler için link skoru hesaplanacaktır. Baker ve Akcayol' un [29] geliştirdikleri öncelikli kuyruk yapısında InterDomain linkin skoru (LS_{max}) ve IntraDomain linkin skorunun (LS_{min}) hesaplanması sırasıyla denklem (1) ve (2)' deki gibidir.

$$LS_{max} = \frac{\sum(\propto_{inter}+\propto_{intra}) - \beta_{intra}}{\sum(\propto_{inter}+\propto_{intra})} * 0.66 \quad (1)$$

$$LS_{min} = \frac{\sum(\propto_{inter}+\propto_{intra}) - \beta_{intra}}{\sum(\propto_{inter}+\propto_{intra})} * 0.33 \quad (2)$$

Denklemlerde kullanılan $\propto_{inter}$ InterDomain linklerin toplamını, $\propto_{intra}$ IntraDomain linklerin toplamını ve β_{intra} InterDomain ve IntraDomain arasındaki minimum değeri gösteren değerdir. Bu hesaplama ile bir web sayfasındaki tüm IntraDomainler ile InterDomainler kendi içerisinde aynı değerleri alır. Bunun sonucunda öncelikli taramada artarda tarandıkları için iki temel sorun ile karşılaşılmaktadır. İlk sorun nezaket politikasına aykırı olarak bant genişliği sık kullanmış olmasıdır. İkinci sorun ise LS aynı olduğundan kapsamın hızlı bir şekilde genişlememesidir. Tohum URL' ler arasından InterDomain ve IntraDomain sayıları 20 ile 80 arasında olan rastgele seçilmiş 20 web sitesinin denklem (1) ve (2)' ye göre dağılımı Şekil 3' de gösterilmiştir.



Şekil 3. Baker ve Akcayol 'a göre LS dağılımı

Bu sorunları çözmek için hesaplanan değeri sabit bir sayı ile çarpmak yerine her bir URL için rastgele bir sayı ile çarpmanın kapsamı ve dallanmayı arttırdığını söyleyebiliriz. Bunun için geliştirilen denklemlerimiz sırasıyla denklem (3) ve (4) ' deki gibidir

$$LS_{min} = \frac{\sum(\propto_{inter}+\propto_{intra}) - \beta_{intra}}{\sum(\propto_{inter}+\propto_{intra})} * rand(x_{min}, x_{max}) \quad (3)$$

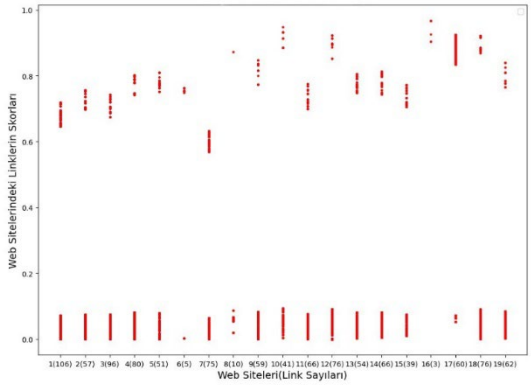
$$LS_{max} = \frac{\sum(\propto_{inter}+\propto_{intra}) - \beta_{intra}}{\sum(\propto_{inter}+\propto_{intra})} * rand(y_{min}, y_{max}) \quad (4)$$

Denklem (3) ve (4)' de görüldüğü üzere her bir URL için LS hesaplanırken, denklem sabit bir sayı yerine LS_{min} x_{min}, x_{max} aralığındaki rastgele bir sayı ile LS_{max}' da y_{min} ve y_{max} aralığındaki rastgele bir sayı ile çarpılmıştır. En uygun x_{min}, x_{max} ve y_{min}, y_{max} değerlerini tespit etmek için Tablo 1' deki 5 farklı değer kullanılarak link skorları ayrı ayrı hesaplanmıştır.

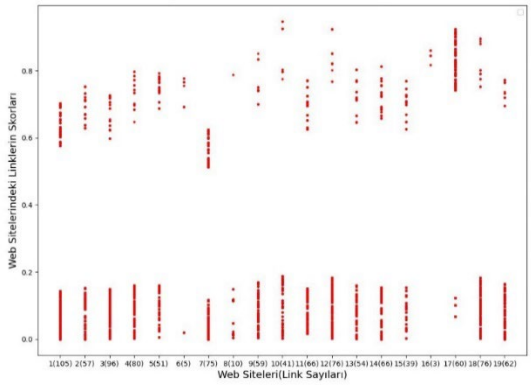
Tablo 1. Kullanılan x_{min}, x_{max} *ve* y_{min}, y_{max} değerleri

Aralık Setleri	x_{min}	x_{max}	y_{min}	y_{max}
1	0	0.1	0.9	1
2	0	0.2	0.8	1
3	0	0.3	0.7	1
4	0	0.4	0.6	1
5	0	0.5	0.5	1

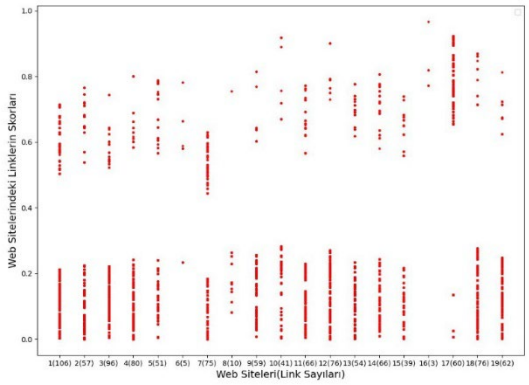
Şekil 3' de kullanılan 20 web sitesi üzerinde Tablo 1' deki 5 faklı değer aralık seti kullanılarak LS_{min} ve LS_{max} dağılımı test edilmiştir. Hesaplanan LS dağılımları Şekil 4,5,6,7 ve 8' de gösterilmiştir.



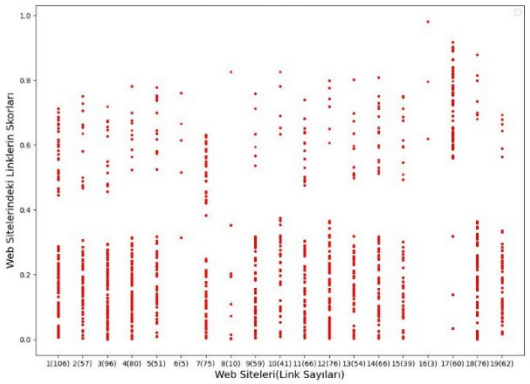
Şekil 4. 1 Numaralı Aralık Seti için LS Dağılımı



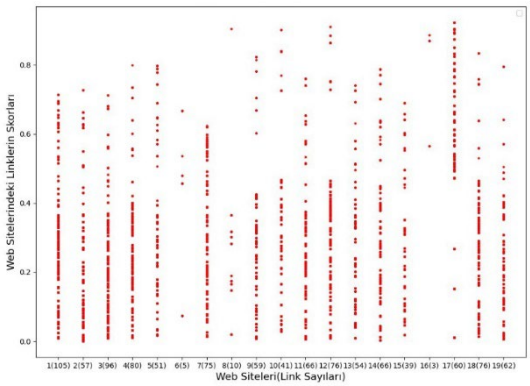
Şekil 5. 2 Numaralı Aralık Seti için LS Dağılımı



Şekil 6. 3 Numaralı Aralık Seti için LS Dağılımı

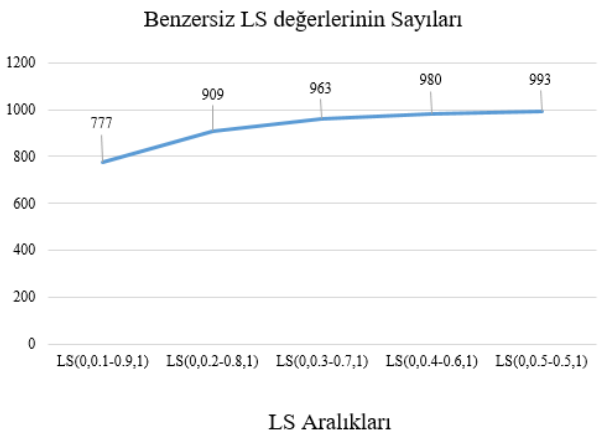


Şekil 7. 4 Numaralı Aralık Seti için LS Dağılımı



Şekil 8. 5 Numaralı Aralık Seti için LS Dağılımı

Şekil 4, 5, 6, 7 ve 8' de görüldüğü gibi LS değeri IntraDomainler ile InterDomainler hesaplandığında sadece 2 farklı değer değil, 0 ile 1 arasında dağılmış değerlerden oluşacaktır. Sistemin test edildiği 5 faklı aralık incelendiğinde aralık değeri arttırıldıkça link skorları da buna paralel olarak artmaktadır. 20 web sayfasından alınan 1082 farklı URL' in ayrı ayrı link skorları hesaplanmıştır. Alınan 5 farklı aralıkta URL' lerin link skorlarının benzersiz olanlarının sayısı Şekil 9' da gösterilmiştir.



Şekil 9. LS aralıklarına göre URL' lerin benzersiz LS dağılımları.

Grafikte görüldüğü üzere LS hesaplanırken aralık değeri arttırıldığında URL' lerin LS değerlerinin dağılımı homojen olmaktadır. LS_{min} hesaplanırken aralık 0 ile 0.5, LS_{max} hesaplanırken de aralığın 0.5 ile 1 arasında alınması en uygun sonucu vermektedir. LS' ye göre öncelikli kuyruk oluşturulduğunda aynı domainde tarama olasılığı azalırken farklı domain tarama olasılığı artmaktadır. Bunun sonucunda hem farklı sayfalarda işlem yapıp bant genişliğini ihlal etmeyecek hem de hızlı bir şekilde keşfedilmemiş yeni web sayfalarını keşfedecektir.

IV. DENEYSEL SONUÇLAR

Deneysel çalışmada 3 farklı tohum URL seti kullanılmıştır. Bunlar ; (1) farklı ülkeler ile kesişimi olmayan tüm URL' ler, (2) her bir ülkede toplam bağlı site sayısı en fazla olan URL' ler ve (3) her bir ülkede InterDomain sayısı en fazla olan 5 URL. Burada (1) numaralı tohum URL seti oluşturulurken neredeyse bütün ülkelerde popüler olan web sitelerine (Google, Twetter, Instagram vs.) diğer çoğu URL' lerden köprü bulunduğu için, bu siteler tohum setine eklenmemiştir. Benzersiz olan 2502 URL' den tüm ülkelerde popüler olan sayfalar çıkartıldığında 1920 tohum URL (1) numaralı tohum URL setine eklenmiştir. (2) numaralı tohum seti oluşturulurken hem sayı fazla tutulmamış hem de her bir ülkede HITS [10] algoritmasına göre merkez sayılan web siteleri alınmıştır. Bu özelliklere sahip 102 faklı ülkeden 102 farklı URL alınarak (2) numaralı tohum URL seti oluşturulmuştur. Son olarak, (3) numaralı tohum URL setinde ise diğer web sayfalarına bağlantı URL' i sayısına göre sıralama yapılmış ve her bir ülkede eşit sayı olacak şekilde en çok bağlantıya sahip olan 5 URL alınmıştır. Diğer web sitelerine bağlantı URL sayıları fazla olan URL' leri almadaki temel amaç kapsam genişliğini arttırmak ve mümkün olduğu kadar az tohum URL kullanarak elde edilebilecek en iyi kapsama ulaşmaktır. Bu nedenle 101 farklı ülkeden bu özelliklere sahip 5 URL alınarak 505 URL' e sahip (3) numaralı tohum URL seti oluşturulmuştur.

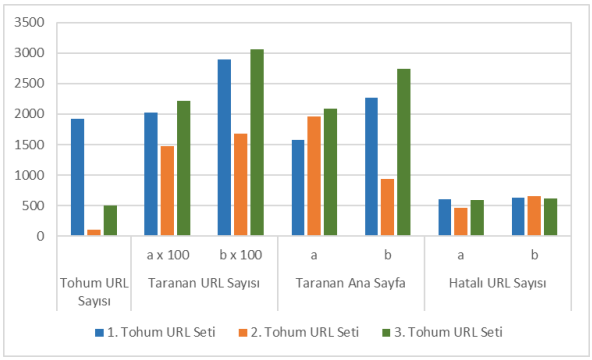
Tablo 2' de 3 farklı tohum URL setinde 3 saatlik zaman periyoduna göre yapılan taramalarda elde edilen sonuçlar listelenmiştir. Burada (a) [29]' da ki çalışmada Baker ve Akcayol' un önerdiği yöntem ile yapılan tarama sonuçlarını, (b) ise çalışmamızda önerdiğimiz yöntem ile yapılan tarama sonuçlarını göstermektedir. Deneyler Intel Xeon CPU E5-2650 2.00 GHz işlemci, 8 GB RAM ve Windows Server 2019 işletim sistemine sahip bilgisayar üzerinde yapılmıştır. Geliştirilen tarayıcı Python programlama dili ve kütüphaneleri kullanılarak MySQL veri tabanı üzerinde uygulanmıştır.

Tablo 2. 3 farklı tohum URL setinde 3 saatlik zaman periyoduna göre yapılan tarama sonuçları

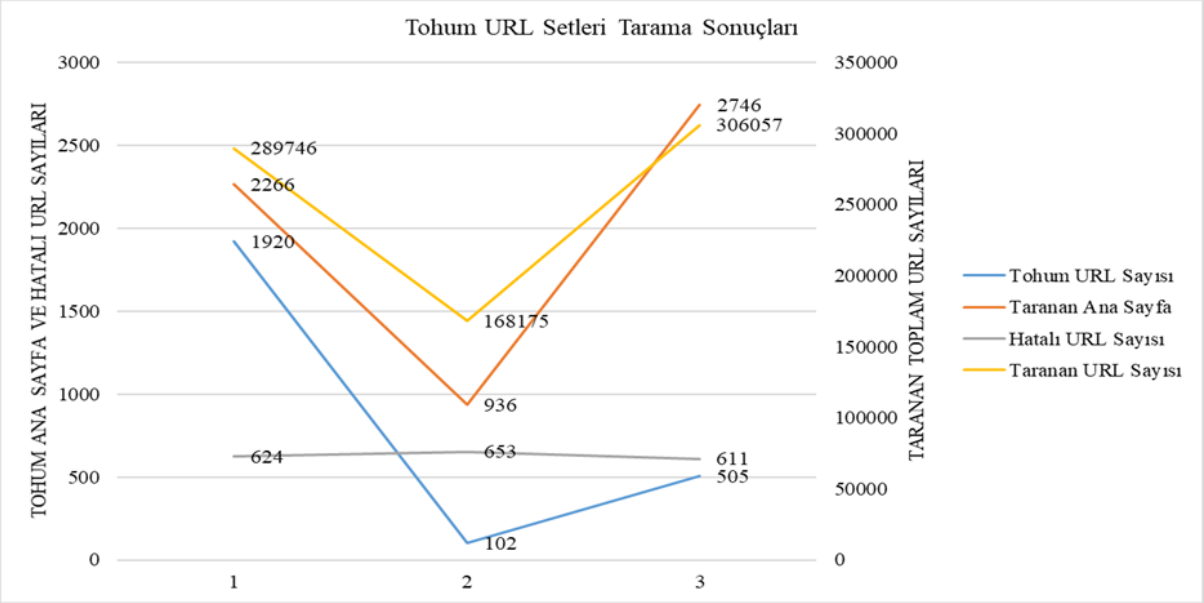
Tohum URL Seti	Tohum URL Sayısı	Taranan URL Sayısı		Taranan Ana Sayfa		Hatalı URL Sayısı		Hatalı URL Oranı (%)	
		a	b	a	b	a	b	a	b
1	1920	202588	289746	1579	2266	601	624	0,30	0,22
2	102	147380	168175	1956	936	452	653	0,31	0,39
3	505	221071	306057	2089	2746	589	611	0,27	0,20

Tablo 2' de görüldüğü gibi önerilen yöntemde, taranan URL sayıları ve taranan ana sayfa sayıları daha başarılı sonuç vermektedir. Hatalı URL oranı, hatalı URL sayısının taranan toplam URL sayısına oranını göstermektedir. Hatalı URL oranları incelendiğinde, çalışmamızda önerilen yöntem 1 ve 3 numaralı tohum URL setlerinde daha başarılı sonuç verirken 2 numaralı tohum URL setinde daha kötü sonuç vermiştir. Tablo 2' nin verilerinin görselleştirilmiş hali Şekil 10' da gösterilmiştir Önerilen yöntem ile yapılan taramalar sonucunda elde edilen sayısal verilerden tohum URL sayıları, taranan ana sayfa sayıları ile hatalı URL sayıları sol eksende ve taranan URL sayıları sağ eksende olmak üzere Şekil 11' de gösterilmiştir.

.



Şekil 10. 3 farklı tohum URL setinde 3 saatlik zaman periyoduna göre yapılan tarama sonuçları



Şekil 11. 3 farklı tohum URL setinde 3 saatlik zaman periyoduna göre yapılan tarama sonuçları

Tarama sonucundan da görüldüğü gibi tohum URL sayısının taranan URL, ana sayfa ve hatalı URL sayısı üzerinde etkisi yoktur. Performansı etkileyen en önemli gösterge kullanılan tohum URL setinin kalitesidir. Tarama sonuçlarına göre en çok tohum URL kullanan (1) ve en az tohum URL kullanan (2) numaralı tohum URL setlerine göre (3) numaralı tohum URL setinin tüm sonuçlarda daha iyi performans gösterdiği görülmüştür. Özellikle (3) numaralı tohum URL seti ile tarama başlatıldığında diğer setlere göre hem hatalı sayfa sayısı ve oranı daha düşük, hem de taranan ana sayfa ve toplam URL sayıları daha fazla çıkmıştır.

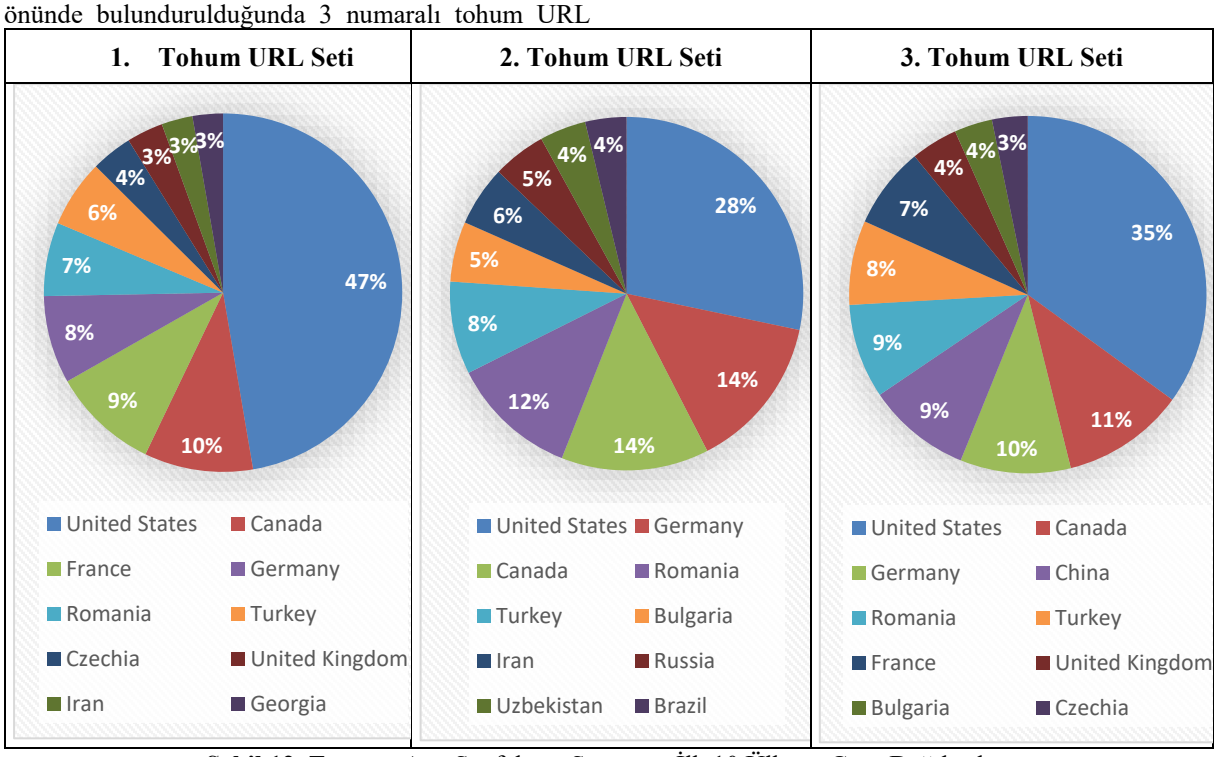
Şekil 11' de görüldüğü gibi en kötü performansı (2) numaralı tohum URL setinin gösterdiği görülmektedir. Bunun nedenlerinden biri tohum URL setinin az sayıda URL den oluşmasıdır. Bir değer nedeni ise toplam bağlı site sayısının fazla olması, o sitenin [10]' da tanımlanan otorite web sayfaları kategorisinde olup bilgilendirici bir içeriğe sahip olduğunu ve InterDomain sayısının az olmasından dolayı kapsamın hızlı bir şekilde genişlememesine sebep olduğunu göstermektedir. Birbirine yakın performans gösteren (1) ve (3) numaralı tohum URL setlerinin incelendiğinde (1) numaralı setin tohum URL sayısının (3) numaralı setten yaklaşık 4 kat daha fazla olduğu görülmektedir. Buna rağmen (1) numaralı setin (3) numaralı setten daha kötü performans gösterdiği görülmektedir. Bunun temel sebebinin (1) numaralı setin içerisindeki URL'lerin hem merkez hem otorite sayfalarından oluşmasından, (3) numaralı setin ise ağırlıklı olarak merkez sayfalardan oluşmasından kaynaklandığı söylenebilir.

Kapsam genişliği bakımından bir diğer performans ölçütü taranan URL' lerin bulundukları ülkeye göre dağılımlarıdır. Tablo 3' de 3 farklı tohum URL seti ile yapılan taramada elde edilen ana sayfaların sayısının ülkelere göre dağılımları gösterilmektedir.

Tablo 3. 3 farklı tohum URL seti ile yapılan taramada elde edilen ana sayfaların sayısının ülkelere göre dağılımları

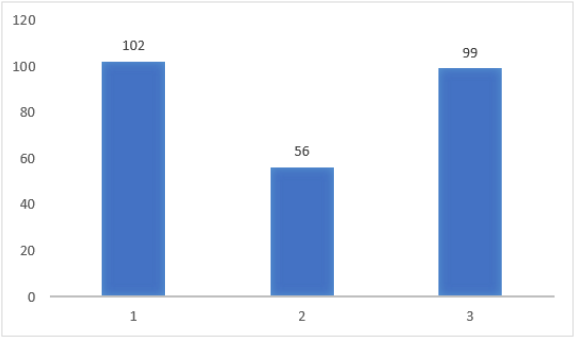
1. Tohum URL Seti		**2. Tohum URL Seti**		**3. Tohum URL Seti**	
Web Sayfası Sayısı	Ülke	Web Sayfası Sayısı	Ülke	Web Sayfası Sayısı	Ülke
776	United States	180	United States	651	United States
161	Canada	90	Germany	207	Canada
158	France	86	Canada	186	Germany
130	Germany	74	Romania	175	China
109	Romania	54	Turkey	159	Romania
100	Turkey	35	Bulgaria	142	Turkey
61	Czechia	35	Iran	137	France
54	U. Kingdom	31	Russia	79	U. Kingdom
46	Iran	27	Uzbekistan	64	Bulgaria
45	Georgia	24	Brazil	60	Czechia
626	Diğer	300	Diğer	886	Diğer

Taranan ana sayfaların ülkelere göre dağılımları incelendiğinde 2 ve 3 numaralı tohum URL setlerinin 1 numaralı tohum URL setine göre daha homojen dağıldığı görülmüştür. Toplam taranan URL sayısı göz önünde bulundurulduğunda 3 numaralı tohum URL setinin 2 numaralı tohum URL setine göre yüzdesel olarak daha başarılı sonuç verdiği tespit edilmiştir. Şekil 12‘ de bu dağılımların yüzdesel olarak grafiği gösterilmiştir.



Şekil 12. Taranan Ana Sayfaların Sayısının İlk 10 Ülkeye Göre Dağılımları

Şekil 13‘ de 3 tohum URL seti ile yapılan taramalar sonucunda ulaşılan ülke sayıları gösterilmiştir.



Şekil 11. 3 Tohum URL Seti ile Yapılan Taramalar Sonucunda Ulaşılan Ülke Sayıları

Şekil 13’de özellikle 2 numaralı tohum URL setinin tarama sonucunda daha az ülkeye ulaşma sebebi her ülkeden en çok ziyaret edilen sadece bir URL alınması ve alınan bu URL ‘in de başka bir ülkeye ait anabilgisayarlarda barındırıldığından kaynaklandığı görülmüştür. 1 ve 3 numaralı tohum URL setleri ile yapılan taramada ise yaklaşık olarak ülkelerin tamamı taranmaktadır. Bölgesel olarak taramanın dağılımı, kapsamı ve performansı olumlu etkilediği söylenebilir.

Tablo 2’de görüldüğü gibi her ülkede IntraDomain sayısı en fazla olan 5 web sitesi (toplan 505) tohum URL olarak alındığında hem keşfedilen URL sayısı ve ana sayfa sayısı en fazla hem de oransal olarak hatalı sayfa sayısı daha azdır. 3 tohum URL setinin taranması sonucunda taranan hatalı sayfa sayıları birbirine çok yakındır. Tablo 4’ de bu hatalar ve hataların tespit edildiği URL sayıları gösterilmiştir.

Tablo 4. 3 tohum URL seti ile yapılan taramada elde edilen hatalar ve sayıları

Hata Kodu	Hata	Hatalı URL Sayısı		
		1	2	3
403	İstemci Hatası: Yasaklanmış URL	177	211	188
11001	Bilinmeyen isim ya da servis	84	74	44
10060	Bağlantı zaman aşımı	73	77	101
	SSL: Başarısız Sertifika Doğrulama	53	66	47
	30 Saniyelik zaman aşımı	41	34	26

406	İstemci Hatası: Kabul Edilmeyen URL	34	27	26
	Maksimum yeniden deneme	33	6	19
10054	Bağlantı sıfırlama hatası	24	16	26
404	İstmeci Hatası: URL bulunamadı	19	22	11
503	Sunucu Hatası: URL için sunucu bulunamadı	17	34	28
500	Sunucu Hatası: URL için Dahili Sunucu Hatası	10	18	24
405	İstemci Hatası: İzin verilmeyen URL	4	12	21
	Diğerleri	42	56	63

Tablo 4 incelendiğinde, 403 hata kodlu yasaklanmış bağlantı sorununun en çok karşılaşılan hata olduğu görülmektedir. Bunun temel nedeninin Tablo 5' de görüldüğü gibi sunucu bilgisayarın kullandığı internet bağlantısının üniversite içinde kurumsal bir bağlantı olması söylenebilir. İkinci en yüksek sayıda karşılaşılan hatanın sebebi de web sitelerinin otomatik tarama yapan tarayıcıları engellemesi ve bu engelin tarayıcı tarafından aşılamamasından kaynaklanmaktadır. Diğer hataların oranlarının ise taranan toplam URL sayısına göre oransal olarak makul seviyede olduğu görülmüştür.

Tablo 5. 3 numaralı tohum URL seti ile üniversite içi ve dışında yapılan taramada elde edilen hatalar ve sayıları

Hata Kodu	Hata	Hatalı URL Sayısı	
		Üniversite Dışı	Üniversite İçi
403	İstemci Hatası: Yasaklanmış URL	151	188
11001	Bilinmeyen isim ya da servis	28	44
10060	Bağlantı zaman aşımı	29	101
	SSL: Başarısız Sertifika Doğrulama	33	47
	30 Saniyelik zaman aşımı	23	26
406	İstemci Hatası: Kabul Edilmeyen URL	7	26
	Maksimum yeniden deneme	7	19
10054	Bağlantı sıfırlama hatası	8	26
404	İstemci Hatası: URL bulunamadı	4	11
503	Sunucu Hatası: URL için sunucu bulunamadı	10	28
500	Sunucu Hatası: URL için dâhili Sunucu Hatası	5	24
405	İstemci Hatası: İzin verilmeyen URL	4	21
	Diğerleri	40	63

Tarama işleminin yapıldığı sunucu bilgisayar üniversite içerisinde barındırılmakta ve bazı web sitelerine (illegal, bahis vb.) giriş izni bulunmamaktadır. Bundan dolayı en başarılı sonucu veren (3) numaralı tohum URL seti üniversite dışında özel bir hat ile bağlantı sağlanarak tarama gerçekleştirilmiştir. Bağlantı özellikleri ve hızının farklılığından dolayı farklı sürelerde aynı sayıda URL elde edilene kadar tarama gerçekleştirilmiştir. Tablo 5' de de görüldüğü gibi tüm hataların üniversite dışında daha az olduğu görülmüştür.

VI. SONUÇ VE TARTIŞMA

Bu çalışmada web sayfalarını taramak için tohum URL seçim metodu ve tarama algoritması tanıtılmıştır. Tohum URL' ler, 102 farklı ülkede ziyaretçinin günlük harcadığı saat, ziyaretçi başına günlük sayfa görüntüleme sayısı, aramadan gelen trafiğin yüzdesi ve toplam bağlı site sayısı temel alarak sıralamış toplam 5079 web site URL'lerinden seçilmiştir. Bu URL' ler içerisinden bazı özelliklere göre seçilmiş ve sırası ile 1920, 102, 505 URL' den oluşan 3 tohum URL seti kullanılarak sistem performansı test edilmiştir.

Önerilen algoritma kapsamı hızlı bir şekilde genişletmek için alan içi, alanlar arası ve ilk rastlanan ana sayfalara link skoru belirleyerek önceliklendirmektedir. Ana sayfaların önem derecesi en yüksek olduğu için link skorlarına 1 değeri atanmaktadır. Bunun dışında link skoru hesaplanmasında farklı değer aralıkları ile testler yapılmıştır. En uygun değer aralığının alan içi URL' lerde (LSmin) 0 - 0.5 ve alanlar arası URL' lerde (LSmax) 0.5 - 1 aralığında olduğu görülmüştür. Algoritmanın 3 farklı tohum URL seti kullanılarak 3 saatlik taramalar sonucunda elde ettiği taranan URL sayısı, taranan ana sayfa sayısı ve hatalı URL sayıları incelenmiştir. Önerilen algoritma, daha önce Baker ve Akcayol tarafından yapılan çalışmada kullanılan algoritma ile karşılaştırılmıştır. Önerilen algoritma hem yeni sayfaları ve bu sayfalardaki URL' leri elde edip link skorları atayarak öncelikli kuyruğa eklemekte, hem de alan içi döngülerden etkili bir şekilde kaçınmaktadır. Yapılan deneysel çalışmalara göre önerilen algoritma kapsamı genişletme, yinelenen URL' ler ve alan içi döngülerden kaçınma konusunda daha iyi bir performans göstermektedir.

Ayrıca kapsamın dünya genelindeki bölgesel dağılımları incelenmiş ve taramalar sonucunda elde edilen ana sayfaların ülkelere göre dağılımları karşılaştırılmıştır. Elde edilen URL' lerin ülkelere göre dağılımları incelendiğinde, genellikle gelişmiş ülkelerde yoğunlaştığı görülmüştür. Bunlara ek olarak 3 tohum URL seti için, taramalar sırasında meydana gelen hatalar, hata sayıları ve hata içerikleri incelenmiş ve analiz edilmiştir.

TEŞEKKÜR

Bu çalışma, TÜBİTAK tarafından BİDEB-2244 Sanayi Doktora Programı kapsamında 118C127 numara ile desteklenen "İnternette Heterojen Veri Kaynaklarından Veri Toplanması, Doğrulanması ve Sorgulanması" başlıklı projenin bir parçasıdır. Sağladığı destek için TÜBİTAK'a teşekkür ederiz.

KAYNAKLAR

[1] "Internet Users Distribution in the World." https://www.internetworldstats.com/stats.htm (accessed 30/03/2022.

[2] M. Abu Kausar, V. Dhaka, and S. Singh, "Web Crawler: A Review," *International Journal of Computer Applications,* vol. 63, pp. 31-36, 02/01 2013, doi: 10.5120/10440-5125.

[3] S. M. Pavalam, S. V. K. Raja, F. K. Akorli, and M. Jawahar, "A survey of web crawler algorithms," *International Journal of Computer Science Issues (IJCSI),* vol. 8, no. 6, p. 309, 2011.

[4] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz, "Evaluating topic-driven Web crawlers," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 241-249.

[5] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the web," *ACM Transactions on Internet Technology (TOIT),* vol. 1, no. 1, pp. 2-43, 2001.

[6] C. Castillo, "Effective web crawling," *SIGIR Forum,* vol. 39, no. 1, pp. 55–56, 2005, doi: 10.1145/1067268.1067287.

[7] X. Zhang and K. P. Chow, "A Framework for Dark Web Threat Intelligence Analysis," *International Journal of Digital Crime and Forensics (IJDCF),* vol. 10, no. 4, pp. 108-117, 2018, doi: 10.4018/IJDCF.2018100108.

[8] M. R. Henzinger, "Algorithmic challenges in web search engines," *Internet Mathematics,* vol. 1, no. 1, pp. 115-123, 2004.

[9] S. Daneshpajouh, M. M. Nasiri, and M. Ghodsi, "A Fast Community Based Algorithm for Generating Web Crawler Seeds Set," in *WEBIST (2)*, 2008, pp. 98-105.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," 1998, vol. 98: Citeseer, pp. 668-677.

[11] S. Zheng, P. Dmitriev, and C. L. Giles, "Graph-based seed selection for web-scale crawlers," presented at the Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, 2009. [Online]. Available: https://doi.org/10.1145/1645953.1646277.

[12] A. C. Nwala, M. C. Weigle, and M. L. Nelson, "Garbage, Glitter, or Gold: Assigning Multi-dimensional Quality Scores to Social Media Seeds for Web Archive Collections," ed. Ithaca: Cornell University Library, arXiv.org, 2021.

[13] B. Ganguly and R. Sheikh, "A review of focused web crawling strategies," *International Journal of Advanced Computer Research,* vol. 2, no. 4, p. 261, 2012.

[14] F. J. M. Shamrat, Z. Tasnim, A. S. Rahman, N. I. Nobel, and S. A. Hossain, "An effective implementation of web crawling technology to retrieve data from the world wide web (www)," *International Journal of Scientific & Technology Research,* vol. 9, no. 01, pp. 1252-1256, 2020.

[15] L. Jiang and H. Zhang, "Multi-agent based individual web spider system," in *2010 World Automation Congress*, 2010: IEEE, pp. 177-181.

[**16**] S.-B. Chan and H. Yamana, "The method of improving the specific language focused crawler," in *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010.

[17] F. Menczer and A. E. Monge, "Scalable web search by adaptive online agents: An infospiders case study," in *Intelligent Information Agents*: Springer, 1999, pp. 323-347.

[18] P. N. Priyatam, A. Dubey, K. Perumal, S. Praneeth, D. Kakadia, and V. Varma, "Seed selection for domain-specific search," presented at the Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 2014. [Online]. Available: https://doi.org/10.1145/2567948.2579216.

[19] L. Sanagavarapu, S. Sarangi, R. Reddy, and V. Varma, *Fine Grained Approach for Domain Specific Seed URL Extraction*. 2018.

[20] L. M. Sanagavarapu, S. Sarangi, and Y. R. Reddy, "ABC Algorithm for URL Extraction," in *ICWE Workshops*, 2017.

[21] S. Pavalam, S. K. Raja, F. K. Akorli, and M. Jawahar, "A survey of web crawler algorithms," *International Journal of Computer Science Issues (IJCSI),* vol. 8, no. 6, p. 309, 2011.

[22] N. Alderratia and M. Elsheh, "Using Web Pages Dynamicity to Prioritise Web Crawling," in *Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence*, 2019, pp. 40-44.

[23] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," *Computer networks and ISDN systems,* vol. 30, no. 1-7, pp. 161-172, 1998.

[24] J. Prakash and R. Kumar, "Web Crawling through Shark-Search using PageRank," *Procedia Computer Science,* vol. 48, pp. 210-216, 2015/01/01/ 2015, doi: https://doi.org/10.1016/j.procs.2015.04.172.

[25] L. Cao *et al.*, "Rankcompete: Simultaneous ranking and clustering of information networks," *Neurocomputing,* vol. 95, pp. 98-104, 2012.

[26] M. Najork and J. L. Wiener, "Breadth-first crawling yields high-quality pages," presented at the Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong, 2001. [Online]. Available: https://doi.org/10.1145/371920.371965.

[27] D. Gupta and D. Singh, "User preference based page ranking algorithm," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 29-30 April 2016 2016, pp. 166-171, doi: 10.1109/CCAA.2016.7813711.

[28] F. Alhaidari, S. Alwarthan, and A. Alamoudi, "User Preference Based Weighted Page Ranking Algorithm," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 19-21 March 2020 2020, pp. 1-6, doi: 10.1109/ICCAIS48893.2020.9096823.

[29] M. Baker and M. Akcayol, "Priority Queue Based Estimation of Importance of Web Pages for Web Crawlers," *International Journal of Computer Electrical Engineering,* vol. 9, pp. 330-342, 07/27 2017, doi: 10.17706/ijcee.2017.9.1.330-342.

[30] Alexa. "The top 500 sites on the web." Amazon. https://www.alexa.com/topsites/countries (accessed 9:12:2021, 2021).