T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BAĞLANTILI VERİ KAYNAKLARININ TESPİTİ VE ANALİZİNE İLİŞKİN YENİ BİR YÖNTEM

Semih YUMUŞAK

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ekim-2017
KONYA

TEZ KABUL VE ONAYI

Semih YUMUŞAK tarafından hazırlanan "Bağlantılı Veri Kaynaklarının Tespiti ve Analizine İlişkin Yeni Bir Yöntem" adlı tez çalışması 13/10/2017 tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof.Dr. Ali OKATAN	.....
Danışman Doç.Dr. Halife KODAZ	.....
Üye Prof.Dr. Ahmet ARSLAN	.....
Üye Yrd.Doç.Dr. Nurdan BAYKAN	.....
Üye Yrd.Doç.Dr. Ersin KAYA	.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Mustafa YILMAZ
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Semih YUMUŞAK

Tarih: 13.10.2017

ÖZET

DOKTORA TEZİ

BAĞLANTILI VERİ KAYNAKLARININ TESPİTİ VE ANALİZİNE İLİŞKİN YENİ BİR YÖNTEM

Semih YUMUŞAK

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Halife KODAZ
İkinci Danışman: Prof. Dr. Erdoğan DOĞDU

2017, 92 Sayfa

Jüri
Doç. Dr. Halife KODAZ
Prof. Dr. Ahmet ARSLAN
Prof. Dr. Ali OKATAN
Yrd. Doç. Dr. Nurdan BAYKAN
Yrd. Doç. Dr. Ersin KAYA

Anlamsal ağlar ve bağlantılı veri, çevrim içi veri akışlarını düzenlemek, veri yığınlarını anlamlı hale getirerek bağlantılandırmak ve sonuç olarak kolay sorgulanabilir ve erişilebilir bir düzen içerisine koymak üzere tasarlanmıştır. Bu bağlamda internet sunucularında saklanan veri çöplüklerinin birer dağıtık veri kaynağı haline dönüştürülerek, farklı bölgelerden anlamsal sorgular ile sorgulanabilmesi amaçlanmaktadır. Bu veri kaynaklarının sorgulanması amacıyla geliştirilmiş olan SPARQL sorgulama dili kullanılarak SPARQL uç noktalarına bağlı olan bağlantılı veri kaynakları sorgulanabilmektedir. SPARQL uç noktaları, bağlantılı veri kaynaklarının bağlantı noktaları üzerinden temel HTTP veya SOAP benzeri protokoller ile sorgulanabilmesi amacıyla geliştirilen servislerdir. Çevrim içi bağlantılı veri kaynaklarının sorgulanmasına olanak veren bu uç noktalar, internet üzerinde dağınık olarak bulunmakta ve kullanıcılar tarafından kolaylıkla tespit edilememektedir. Bu veri kaynaklarını ve bağlı olan uç noktaları, kullanıcılar tarafından kolay erişilebilir kılmak amacıyla listeleyen çeşitli çalışmalar olmakla birlikte, bu çalışmaların yetersiz olduğu bu tez çalışmasıyla tespit edilmiştir. Bu yetersizliği giderebilmek amacıyla, SPARQL uç noktalarını otomatik olarak tespit eden, sürekli gözlem ve analizlerini gerçekleştirebilen bir meta-arama ve analiz aracı geliştirilmiştir. Tespit edilen SPARQL uç noktalarının kullanıcılar tarafından kullanılabilmesini sağlamak amacıyla da sınıflandırma, konu önerme, etiketleme gibi işlemler uygulanmıştır. Bu tez çalışmasında, SPARQL uç noktalarının tespit edilmesi aşamasından, içerik analizi yapılarak kullanıcılara sunulabilmesi aşamasına kadar olan tüm süreçler açıklanmaktadır. Tespit edilen tüm uç noktalar ve bunlara bağlı çıkan sonuçlar, mevcut diğer benzer çalışmalarla karşılaştırılmıştır.

Anahtar Kelimeler: Ağ Madenciliği, Anlamsal Ağ, Bağlantılı Veri

ABSTRACT

Ph.D THESIS

A NOVEL METHOD TO DISCOVER AND ANALYZE LINKED DATA SOURCES

Semih YUMUŞAK

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER ENGINEERING

Advisor: Assoc. Prof. Dr. Halife KODAZ
Co-Advisor: Prof. Dr. Erdoğan DOĞDU

2017, 92 Pages

Jury
Assoc. Prof. Dr. Halife KODAZ
Prof. Dr. Ahmet ARSLAN
Prof. Dr. Ali OKATAN
Asst. Prof. Dr. Nurdan BAYKAN
Asst. Prof. Dr. Ersin KAYA

Semantic web and linked data are designed to organize online data flow, link and semantify data stacks, and consequently provide easily queriable and accessible data stores. In this manner, data dumps stored in internet servers are converted into a distributed data source and become available for semantic querying from different locations. In order to provide a querying infrastructure for these data sources, SPARQL querying language was designed. SPARQL query language is used to query SPARQL endpoints, which allows users to query linked data sources through HTTP or SOAP-like protocols. These endpoints are distributed among internet and allows users to query several different data sources. Although there are many endpoints on the web, the discovery of these endpoints is not an easy task for the users. There are studies and repositories to provide link data sources for data consumers; however, the quality and the quantity of these studies are limited. In order to enhance these studies, a discovery and analysis engine is developed to discover and continuously analyze SPARQL endpoints. After the SPARQL endpoint repository creation, classification, topic recommendation, and tagging techniques for SPARQL endpoints are developed. In this thesis, the complete process starting from the discovery to the content analysis and serving of the results are explained. The results coming from the developed engine are compared with other similar studies.

Keywords: Web Mining, Semantic Web, Linked Data

ÖNSÖZ

Doktora tez çalışmam boyunca değerli katkılarını, yönlendirici desteğini ve anlayışını hiçbir zaman esirgemeyen birinci danışmanım Sayın Doç. Dr. Halife KODAZ'a ve ikinci danışmanım Sayın Prof. Dr. Erdoğan DOĞDU'ya, yurtdışı araştırmam sırasında Insight Araştırma Merkezi bünyesinde danışmanlığımı üstlenen Sayın Dr. Pierre-Yves Vandenbussche'ye ve çalışmalarıma değerli yorumlarıyla katkıda bulunan çalışma arkadaşlarım Dr. Anderas KAMILARIS ve Emir Muñoz'a, tezin gelişmesine yönlendirici görüş ve önerileri ile yardımcı olan tez izleme komitesi üyeleri Sayın Prof. Dr. Ahmet ARSLAN'a ve Sayın Yrd. Doç. Dr. Nurdan BAYKAN'a, bu süre boyunca göstermiş oldukları desteklerinden, ilgilerinden ve yardımlarından dolayı Selçuk Üniversitesi Bilgisayar Mühendisliği Bölümü'nün tüm öğretim elemanlarına, özellikle manevi desteğini hiçbir zaman esirgemeyen her zaman ve her konuda hep yanımda olan eşime ve aileme teşekkür ederim.

Semih YUMUŞAK

KONYA-2017

İÇİNDEKİLER

ÖZET ..... **iv**

ABSTRACT ..... **v**

ÖNSÖZ ..... **vi**

ŞEKİLLER LİSTESİ ..... **x**

ÇİZELGELER LİSTESİ ..... **xii**

1. GİRİŞ **1**

1.1. Tezin Amacı ve Literatüre Katkıları 2
1.2. Tezin Organizasyonu 4

2. KAYNAK ARAŞTIRMASI **5**

2.1. Web Tarama ve Arama Motoru Tarama 5
2.2. Bağlantılı veri (Linked Data) 6
2.3. Bağlantılı Meta-veri (Linked Meta data) 10
2.4. Bağlantılı Veri Kaynakları Endeksleri ve Analizi 12
2.4.1. Bağlantılı veri derecelendirme yöntemleri 12
2.4.2. SPARQL uç noktası kaynakları 18
2.4.3. Kategorizasyon ve konu belirleme 18
2.4.4. Wordnet semantik sözlüğü 19

3. MATERYAL VE YÖNTEM **20**

3.1. Diğer Projelerden Veri Toplama ve Analiz 20
3.2. Arama Motorları Sonuçları Üzerinden SPARQL Uç Noktası Tespiti 22
3.2.1. Meta-Tarama için arama kelimelerinin oluşturulması 23
3.2.2. Bağlantı çıkarım kriteri ve filtreleme 24
3.2.3. Alan adı öğrenmesi 26
3.2.4. İstatistiksel analiz yöntemleri 26
3.3. SPARQL Uç Nokta URL'lerinin Sınıflandırılması 27
3.3.1. Yazı içerik toplama 28
3.3.2. Skorlama 28
3.3.3. Bağlantılı veri kaynaklarının sınıflandırılması 30

4. SpEnD META-TARAMA MOTORU UYGULAMASI **31**

4.1. Tarayıcı Grafik Ara Yüzü 33
4.2. Analiz Grafik Ara Yüzü 34

5. ARAŞTIRMA SONUÇLARI VE TARTIŞMA **36**

5.1. Tespit Edilen SPARQL Uç Noktalarının Mevcut Listeler ile Karşılaştırması 37

5.1.1. Karşılaştırmalı istatistiksel sonuçlar 41
5.2. Servis Özellikleri 42
5.3. İnteroperabilite (SPARQL 1.0 ve 1.1 desteği) 43
5.4. Performans Değerlendirmeleri 45
5.4.1. Sonuç akış (streaming) performansı 45
5.4.2. Atomik arama (lookup) ve katılma (join) performansı 47
5.5. SPARQL Uç Noktalarının Değerlendirilmesi 49
5.5.1. İçerik değerlendirme 49
5.5.2. Sözlük ve ontoloji değerlendirmeleri 50
5.6. SPARQL Uç Noktalarının Sınıflandırılma Sonuçları 52
5.6.1.Bağlantılı veri kaynakları için konu tavsiye yöntemi 53
5.6.2. Bağlantılı veri kaynaklarının sınıflandırılması 53
5.6.3. Bağlantılı veri kaynaklarının sınıflandırılma sonuçlarının istatistiksel analizi 62

6. SONUÇLAR VE ÖNERİLER 67

KAYNAKLAR 69

EKLER 77

EK-1 Detay Tablolar 77

ÖZGEÇMİŞ.....91

SİMGELER VE KISALTMALAR

Kısaltmalar

API	:	Application Programming Interface
CKAN	:	The Comprehensive Kerbal Archive Network
HTML	:	Hypertext Markup Language
IoT	:	Internet of Things
JSON	:	JavaScript Object Notation
LOD	:	Linked Open Data
LOD Cloud	:	Linking Open Data Cloud
PLD	:	Pay Level Domain
RDF	:	Resource Description Framework
RDFa	:	The Resource Description Framework in Attributes
SPARQL	:	SPARQL Protocol and RDF Query Language
Turtle	:	Terse RDF Triple Language
URI	:	Unique Resource Identifier
URL	:	Unique Resource Locator
VoID	:	Vocabulary of Interlinked Datasets
XML	:	eXtensible Markup Language

ŞEKİLLER LİSTESİ

Şekil 2.1. Linking Open Data Cloud Diyagramı 2007 7
Şekil 2.2. Linking Open Data Cloud Diyagramı 2014 8
Şekil 2.3. Linking Open Data Cloud Diyagramı 2017 9
Şekil 3.1. Bağlantılı veri kümeleri Venn diyagramı (Yumusak ve ark., 2017) 22
Şekil 3.2. Arama Motoru Objesi XML Şeması 23
Şekil 3.3. Arama Motoru Objesi Örnek XML Bloğu (Yumusak ve ark., 2017) 23
Şekil 4.1. SpEnD sistem diyagramı (Yumusak ve ark., 2017) 31
Şekil 4.2. İş Parçacığı aktivite diyagramı (Yumusak ve ark., 2017) 32
Şekil 4.3. SPECAN v2.0: SpEnD projesi masaüstü yazılımı tarama penceresi 34
Şekil 4.4. SPECAN v2.0: SpEnD projesi masaüstü yazılımı analiz penceresi 35
Şekil 5.1. Uç nokta sayısı ile taranan URL sayılarının karşılaştırılması 37
Şekil 5.2. Tüm projelerde bulunan toplam SPARQL uç noktası sayıları 39
Şekil 5.3. Tüm projelerde bulunan erişilebilir SPARQL uç noktası sayıları 40
Şekil 5.4. Tüm projelerde bulunan erişilebilir SPARQL uç noktalarına ait tekil alan adlarının sayısı 40
Şekil 5.5. Erişilebilirlik aralıklarına göre uç nokta sayıları (Yumusak ve ark., 2017) 41
Şekil 5.6. İstatistiksel karşılaştırmalı analiz sonuçlarına göre uç nokta adetlerinin yüzdesel dağılımı (Yumusak ve ark., 2017) 42
Şekil 5.7. SPARQL sorgu dili v1.0 uyumluluk sonuçları (Yumusak ve ark., 2017) 44
Şekil 5.8. SPARQL sorgu dili v1.1 uyumluluk sonuçları (Yumusak ve ark., 2017) 45
Şekil 5.9. Farklı limit büyüklüklerinin karşılaştırılması (Yumusak ve ark., 2017) 47
Şekil 5.10. ASK sorguları için çalışma zamanı persentil değerleri (Yumusak ve ark., 2017) 48
Şekil 5.11. JOIN sorguları için çalışma zamanı persentil değerleri (Yumusak ve ark., 2017) 49
Şekil 5.12. Kategorilere göre keşfedilen uç noktaları 50
Şekil 5.13. SSN ontolojisi özelliklerinin sayısı 51
Şekil 5.14. Uç noktaların içerildikleri “label” ve “comment” sayılarına göre dağılımı (Yumusak ve ark., 2017) 53
Şekil 5.15. Sınıflandırma yöntemleri ve skorlama yöntemlerinin “label” özelliğine göre doğruluk değerleri (Yumusak ve ark., 2017) 55
Şekil 5.16. Sınıflandırma yöntemleri ve skorlama yöntemlerinin “comment” özelliğine göre doğruluk değerleri (Yumusak ve ark., 2017) 56
Şekil 5.17. Sınıflandırma yöntemleri ve skorlama yöntemlerine göre “comment” özelliklerinin ikinci seviye anlamsal ilişkilerinin doğruluk sonuçları (Yumusak ve ark., 2017) 57
Şekil 5.18. Sınıflandırma yöntemleri ve skorlama yöntemlerine göre “label” özelliklerinin ikinci seviye anlamsal ilişkilerinin doğruluk sonuçları (Yumusak ve ark., 2017) 57
Şekil 5.19. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, “comment” için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017) 58
Şekil 5.20. Naive Bayes sınıflandırıcısına göre F1 değerlerinin, “comment” için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017) 59
Şekil 5.21. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, “label” için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017) 59
Şekil 5.22. Naive Bayes sınıflandırıcısına göre F1 değerlerinin, “label” için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017) 60

Şekil 5.23. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, semantik ikinci seviye “label” için skorlanan özellik sayısının artışına göre değişimi 60
Şekil 5.24. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, semantik ikinci seviye “comment” için skorlanan özellik sayısının artışına göre değişimi 61
Şekil 5.25. Naive Bayes sınıflandırıcısına göre F1 skoru değerlerinin, semantik ikinci seviye “label” için skorlanan özellik sayısının artışına göre değişimi 61
Şekil 5.26. Naive Bayes sınıflandırıcısına göre F1 skoru değerlerinin, semantik ikinci seviye “comment” için skorlanan özellik sayısının artışına göre değişimi 62

ÇİZELGELER LİSTESİ

Çizelge 2.1. VoID sözlüğü istatistiksel analiz özellikleri (Alexander ve ark., 2011) 11
Çizelge 2.2. Mevcut Bağlantılı Veri Seti Koleksiyonları 12
Çizelge 2.3. Derecelendirme çalışmaları özeti (Yumusak ve ark., 2014) 18
Çizelge 3.1. Bağlantılı Veri Koleksiyonları Erişim Yöntemleri 21
Çizelge 3.2. Bağlantılı veri koleksiyonlarının içerdiği SPARQL uç noktası sayısı 21
Çizelge 3.3. Arama Sorguları 24
Çizelge 3.4. Arama sonuçlarının alınması algoritması sözde kodu 25
Çizelge 3.5. URL analizi algoritmasının sözde kodu 25
Çizelge 3.6. Basit SPARQL sorgusu 26
Çizelge 3.7. Önceden tespit edilen alan adlarının tekrar aranması sözde kodu 26
Çizelge 3.8. İstatistiksel Sparql Sorguları 27
Çizelge 5.1.Arama sorgularına göre kaydedilen uç nokta sayısı 37
Çizelge 5.2. Keşfedilen SPARQL uç noktalarının erişilebilirlik karşılaştırması 38
Çizelge 5.3. Keşfedilen SPARQL uç noktalarının diğer veri kaynakları ile karşılaştırması 41
Çizelge 5.4. Keşfedilen SPARQL uç noktalarının diğer veri kaynakları ile istatistiksel olarak karşılaştırması 42
Çizelge 5.5. Keşfedilen SPARQL uç noktası ve alan adı sunucu bilgileri karşılaştırması (Yumusak ve ark., 2017) 43
Çizelge 5.6. Sonuç kısıtının sınırları (Yumusak ve ark., 2017) 46
Çizelge 5.7. IoT alanında geliştirilmiş tespit edilen ontolojilerin sayısı 51
Çizelge 5.8. Kruskal-Wallis H test: farklı skorlama tekniklerine göre ortalama doğruluk değerlerinin farklılıkları (Yumusak ve ark., 2017) 63
Çizelge 5.9. Mann-Whitney U test: Farklı skorlama tekniklerinin ortalama doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017) 64
Çizelge 5.10. Mann-Whitney U test: Farklı semantik seviyelerin ortalama doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017) 64
Çizelge 5.11. Kruskal-Wallis H test: farklı skorlama tekniklerine göre maksimum doğruluk değerlerinin farklılıkları (Yumusak ve ark., 2017) 65
Çizelge 5.12. Mann-Whitney U test: Farklı skorlama tekniklerinin maksimum doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017) 65
Çizelge 5.13. Mann-Whitney U test: Farklı semantik seviyelerin maksimum doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017) 65

Çizelge EK- 1.1. Uç noktalar için Stf-Idf skoru en yüksek olan terimlerin detaylı listesi (Yumusak ve ark., 2018) 77
Çizelge EK- 1.2. Keşfedilen SPARQL uç noktalarının alan adları bazında tespit edilen uç nokta sayısı 84
Çizelge EK- 1.3. Keşfedilen SPARQL uç noktalarının üçlü sayıları (100 milyondan fazla üçlü barındıran) 89

1. GİRİŞ

Web teknolojilerinin yaygınlaşması veri üretim-tüketimini hızla artırmış, veri yönetimi ve bilgi çıkarım mekanizmalarının tekrar değerlendirilmesinin zorunluluğunu ortaya çıkarmıştır. Ham verinin bilgiye dönüştürülmesi ve anlam ilişkilerinin oluşturulmasında klasik internetin yetersizliğini öngören Tim Berners-Lee, anlamsal ağlar (semantic web) (Berners-Lee ve ark., 2001) vizyon çalışması ile internet verilerinin anlamlı birer bilgi yumağı halinde sunulabileceği tezini öne sürmüştür. Bu vizyon çalışmasında HTML verisi ve veriler arasındaki ilişkilerin bilgisayarlar tarafından anlaşılabilir hale getirilmesi için altyapının yeniden tasarlanması gerektiği öngörülmüş ve sonrasında anlamsal ağ standartları oluşturulmuştur (Berners-Lee ve ark., 2001). İlerleyen yıllarda anlamsal ağ kavramının kabul görmesi ile birlikte bağlantılı veri (linked data) (Berners-Lee, 2006) standartları oluşturulmuş ve anlamsal veri kümelerinin birbirleri arasında bağlantılandırılması amaçlanmıştır. Bağlantılı veri kaynakları, anlamsal ağ teknolojileri kullanılarak ve özellikle üçlü tabanlı bir çizge yapısı (Bizer ve ark., 2009) kullanılarak oluşturulan yapılandırılmış veri kaynakları ağıdır. Bu veri kaynakları, farklı formatlarda (N-Triples, Turtle, JSON vb.), düz RDF (Kaynak Tanımlama Çerçevesi) veri dosyaları veya RDF veri depoları (Virtuoso, Apache Jena, OntoQuad vb.) gibi birçok farklı şekilde sunulmaktadır. Belirtilen veri kaynakları W3C tarafından standartları tanımlanan SPARQL[1] sorgu dili ile sorgulanabilmektedir. Bu kaynakların internet üzerinden canlı sorgulanabilmesi amacıyla oluşturulan ara yüze de SPARQL uç noktası (SPARQL endpoint) ismi verilmektedir.

Klasik internet sitelerinin içerisinde dağınık olarak bulunmakta olan bağlantılı veri kaynakları çoğunlukla arama motorları tarafından anlamsal olarak endekslenmemek-tedir. Bağlantılı veri içerisinde arama yapmak ve verilerin endekslenmesi amacıyla Swoogle (Finin ve ark., 2004), Falcons (Cheng ve Qu, 2009), Sindice (Campinas ve Ceccarelli, 2011), SWSE (Hogan ve ark., 2011) gibi arama motorları geliştirilmiştir. Bu arama motorları klasik web arama motorlarına benzer şekillerde bağlantılı veri setlerini endeksleyip bilgiye hızlı ve kolay ulaşımı amaçlamaktadır. Klasik arama motoru benzeri yöntemlerin yanı sıra, bağlantılı veri kümelerine canlı erişim imkanı sağlayan federe sorgu sistemleri de (Buil-Aranda, 2012)

[1] https://www.w3.org/TR/rdf-sparql-query/

bilgi erişiminde önemli rol oynamaktadırlar. Federe sorgu sistemleri eş zamanlı farklı bağlantılı SPARQL uç noktalarına (SPARQL endpoints) sorgu göndererek yanıtları yorumlayan sistemlerdir. Federe sorgu sistemleri, sorguları anlık olarak dağıtmak için ihtiyaç duydukları SPARQL uç noktalarını belirli kriterlere göre derecelendirerek sorgularını optimize ederler (Umbrich ve ark., 2014). Örneğin SPLENDID (Grlitz ve Staab, 2011), WoDQA (Akar ve Hala, 2012) gibi federe sorgulama ara yüzleri, VoID (Alexander ve ark., 2011) istatistiksel çıkarım standartlarını kullanarak her bir SPARQL uç noktası hakkında tuttukları meta-veriler üzerinden derecelendirme gerçekleştirmekte ve sorguları bu derecelendirmelere göre farklı sorgu noktalarına dağıtmaktadır. Bu bağlamda, uzak sorgulama sistemlerinin sonuç kalitesi, SPARQL uç noktalarının doğru tespit edilmesi, yönetilmesi ve analiz edilmesi ile doğru orantılıdır. Yeni SPARQL uç noktalarının tespiti, sisteme kazandırılması ve analizi için şu üç ana meta-veri projesi bulunmaktadır:

1- Linking Open Data (LOD Project) (Cyganiak ve Jentzsch, 2017),

2- SPARQL Endpoint Status (Sparqles),

3- LODStats (Ermilov ve ark., 2016)

Bu üç projenin de yeni bağlantılı veri setlerinin tespit edilmesi konusunda el ile ekleme veya topluluk projesi (community project) çalışmaları üzerinden veri ekleme yaparak veri setlerini topladıkları bilinmektedir. Bu projeler ile ilgili tez çalışmasında, LOD Project, Sparqles ve LODStats veri setlerinde barındırılan SPARQL uç noktalarının analizleri gerçekleştirilmiştir. Bu analizlerde üç projenin de sonlandırma noktaları kayıtlarında büyük oranda (>50%) çevrimdışı kayıt tespit edilmiştir. Bu bağlamda SpEnD isimli SPARQL uç noktaları tespit ve analiz motoru geliştirilmiştir. Bu tez kapsamında geliştirilen tüm yazılımlar, kaynak kodları ve verilerle birlikte yazarın açık kaynak hesabında[2] yayınlanmıştır.

Bu tez çalışması TÜBİTAK 2214-A Yurtdışı Doktora Sırası Araştırma Burs Programı tarafından B.14.2.TBT.0.06.01-21514107-020-155998 sayılı karar ile desteklenmiştir.

1.1. Tezin Amacı ve Literatüre Katkıları

SPARQL uç noktaları, bağlantılı veri kaynaklarının canlı sorgulanması için

[2] https://github.com/semihyumusak

tasarlanmıştır. Bununla birlikte, SPARQL uç noktalarının çoğunluğu içerik (bağlı olduğu bilgi tabanı) hakkında herhangi bir bilgi içermemektedir. Bu uç noktaların içeriklerini paylaşmak amacıyla, ilgili bilgileri listeleyen veri depoları (Datahub[3], SPARQLES (Vandenbussche ve ark., 2013), LODStats (Ermilov ve ark., 2016), LOD Cloud (Cyganiak ve Jentzsch, 2017)) bulunmaktadır. Uç noktaların tespit edilmesinin zorluğundan dolayı bu veri depolarında birçok canlı SPARQL uç noktasının dizine eklenmediği ve kategorilere ayrılmadığı görülmektedir. Tez süresince gerçekleştilen bir çalışmada (Yumusak ve ark., 2017) bu eksikliği gidermek amacıyla yeni bir SPARQL uç nokta keşif motoru geliştirilmiştir. Bu keşif motoru, diğer tüm veri depolarından daha büyük ve daha kapsamlı bir SPARQL son nokta kümesini tespit edebilmiştir. Ancak SPARQL uç noktalarında barındırılan içeriğin doğru bir açıklaması olmaksızın, bu uç noktaların veri tüketicileri tarafından etkili bir şekilde kullanılması mümkün olamamaktadır. Bu bağlamda, bağlantılı veri kaynaklarının sınıflandırılması, çeşitli veri erişim senaryolarında bağlantılı veri kaynağı kullanıcıları için önemli bir kılavuz oluşturabilmektedir. Bağlantılı veri kaynağı kullanıcılarına örnek olarak, canlı SPARQL sorgu dili işleme, federasyon sorgulama, doğrudan RDF erişim ve sorgulama, gömülü RDFa (The Resource Description Framework in Attributes) endeksleme ve RDF endeksleme sistemleri örnek gösterilebilir ve bu tez çalışmasında elde edilen sonuçların bu kullanıcılar tarafından kullanılabileceği öngörülmektedir. Özellikle SPLENDID (Grlitz ve Staab, 2011), HiBISCuS (Saleem, 2014), ANAPSID (Acosta ve ark., 2011) gibi SPARQL uç noktalarından eş zamanlı alınan sonuçları birleştirmek için içerik bilgisine ihtiyaç duyan federe sorgu işleme motorları, bu tez çalışmasında elde edilen sınıflandırma verilerine ihtiyaç duymaktadır. Ayrıca, SQUIN (Hartig, 2013) gibi bağlantı dolaşımı (link traversal) temelli sorgu yürütme sistemleri, bağlantılı veri kaynaklarını çapraz sorgulama amaçlı kullanmaktadır ve hangi kaynağın sorgulanacağına karar verilmesi aşamasında kaynak ile ilgili meta-verilere ihtiyaç duymaktadır. Federasyon sorgu motoru optimizasyonu (Saleem, 2014) için sorgunun yapısından ve içeriğinden bağımsız olarak hangi uç noktaların kullanılacağına karar verilmesi aşamasında SPARQL uç noktalarının sınıflandırması gereklidir. SpEnD keşif motoruna ek olarak, bu tez çalışmasında tüm bu gereksinimlere cevap verecek şekilde bir SPARQL uç noktasının anlamsal olarak içeriğini tespit etmek ve sunmak amacıyla, bağlantılı veri kaynakları için bir derecelendirme önerisi ve sınıflandırma metodu

[3] http://datahub.io

geliştirilmiştir.

1.2. Tezin Organizasyonu

Bu tez çalışması 6 bölümden oluşmaktadır. Bölümler aşağıdaki şekilde düzenlenmiştir.

Birinci bölümde tez çalışması hakkında giriş yapılarak amacı hakkında bilgilendirme yapılmakta ve literatüre katkıları açıklanmaktadır. İkinci bölümde ise tez çalışmasında kullanılan yöntemler ve ilişkili yöntemlerle ilgili literatür bilgilendirmesi yapılmaktadır.

Üçüncü bölümde, tez çalışmasında kullanılan iki ana yöntem olan uç noktaların tespiti ile ilgili yöntem tanımlamaları ve bu uç noktaların sınıflandırılma yöntemleri anlatılmaktadır. Uç noktaların tespitine yönelik geliştirilen yöntemin sürdürülebilir yapısı, diğer benzer çalışmalardan farklı olarak otomatik bilgi çıkarımı yöntemleri ayrıntılı olarak açıklanmaktadır. Devamında, uç noktaları birer doküman olarak değerlendirerek geliştirilen ve anlamsal içeriklerine göre skorlama yaparak sınıflandırma algoritmaları ile daha verimli sınıflandırılmasına olanak tanıyan yöntemler açıklanmaktadır.

Dördüncü bölümde, tez çalışması süresince geliştirilen meta-tarama motoru uygulaması ekranları, çalışma yöntemi ile birlikte açıklanmaktadır. Bu uygulamanın çok iş parçacıklı yapısı ve bu sayede sürekli keşif ve analiz işlemlerini nasıl gerçekleştirdiği açıklanmaktadır.

Beşinci bölümde, kullanılan uygulama ve sonuçları hakkında ayrıntılı analiz çalışmaları anlatılmaktadır. Bu bölümde meta-tarama motoru uygulamasının elde ettiği uç noktaların meta-veri analiz sonuçları, içerik değerlendirmesi ve sınıflandırma işlemlerinin analiz sonuçları verilmektedir. Son olarak altıncı bölümde, tez çalışmasının sonucu özetlenmekte ve kazanımları açıklanmaktadır.

2. KAYNAK ARAŞTIRMASI

Kaynak araştırması dört kategoriye ayrılırmıştır: (1) Web tarama ve arama motoru tarama, (2) Bağlantılı veri, (3) Bağlantılı meta veri ve (4) Sınıflandırma. Birinci bölümde klasik tarama yöntemleri, bağlantılı veri tarama ve meta tarama yöntemleri anlamsal ağ ile bağlantılı olarak incelenmiştir. İkinci bölüm, bağlantılı veri çalışmalarının mevcut durumunu açıklanmaktadır. Üçüncü bölümde ise bağlantılı veri kümelerinin meta-verileri hakkındaki çalışmalar incelenmiş ve son olarak bu veri kümeleri üzerinde yapılmış çeşitli sınıflandırma yöntemleri incelenmiştir.

2.1. Web Tarama ve Arama Motoru Tarama

İnternetin gelişimi ve web sitelerinin (Berners-Lee ve ark., 1992) büyümesiyle bilgi çıkarım ihtiyacı hızla artmış ve bu ihtiyaç karşısında internette birçok veri toplayıcı (Sheldon ve ark., 1995; Knight, 1996; Miller ve Bharat, 1998; Raghavan ve Garcia-Molina, 2000; Shkapenyuk ve Suel, 2002; Boldi ve ark., 2004) geliştirilmiştir. İlerleyen süreçte açık kaynak kodlu çok kanallı tarayıcıların (multithreaded crawlers) (crawler4j[4], websphinx (Miller ve Bharat, 1998)) yanı sıra İnternet'in yönetilemeyen büyüklüğü karşısında dağıtık veri toplama yazılımlarının (Nutch[5], UbiCrawler (Boldi ve ark., 2004)) ve odaklı tarayıcıların (focused crawlers) (Chakrabarti ve ark., 1999; Rungsawang ve Angkawattanawit, 2005; Shi, 2010; Liu ve Du, 2014; Radu ve Rebedea, 2014; Shah ve ark., 2014; Wan ve ark., 2014) geliştirilmesi zorunlu hale gelmiştir.

Anlamsal ağ bağlamında veri toplayıcılar için bir kırılma noktası, temel olarak HTML belgeleri için tasarlanan yazılımların bir anlamsal ağ standardı olan RDF dokümanları ile tanışmaları olmuştur (Berners-lee ve Hendler, 2001). Klasik web tarama ve endeksleme yöntemleri, anlamsal ağ kapsamı içerisinde yetersiz kalmış ve anlamsal olarak işaretlenmiş verileri toplamak için özelleştirilmiş yöntemler (Patel ve ark., 2003; Ding ve ark., 2005; Dodds, 2006; Yang, 2010; Delbru ve ark., 2012) geliştirilmiştir. Temel olarak anlamsal ağ verilerini taramak amacıyla BioCrawler (Batzios ve ark., 2008), MultiCrawler (Harth ve ark., 2006), OntoKhoj (Patel ve ark., 2003), OntoCrawler (Yang, 2010) gibi projeler geliştirilmiştir. Anlamsal ağ tarama projelerinin sonrasında, tespit edilen ve endekslenen bilgiye kolay ulaşım sağlanması amacıyla Semplore (Wang ve ark., 2009), SemSearch (Lei ve ark., 2006), Sindice

[4] https://code.google.com/p/crawler4j/
[5] http://nutch.apache.org/

(Campinas ve Ceccarelli, 2011), Swoogle (Finin ve ark., 2004), SWSE (Hogan ve ark., 2011), Falcons (Cheng ve Qu, 2009) ve Watson (D'Aquin ve ark., 2011) gibi arama motorları geliştirilmiştir.

Anlamsal ağ arama motorları ve tarayıcılarının gelişimine paralel olarak, meta-veri toplama ve arama konusu üzerinde çalışmalar yapılmıştır. Klasik web için arama motorları sonuçları üzerinde kurgulanan "meta arama motoru" kavramı tanımlanmış (Berton ve ark., 2004; Lawrence ve Giles, 2006; Kenneth ve ark., 2012) ve paralelinde birçok meta arama motoru ve tarayıcısı geliştirilmiştir (SavvySearch (Howe ve Dreilinger, 1997) Helios (Gulli ve Signorini, 2005), WebCrawler[6]).

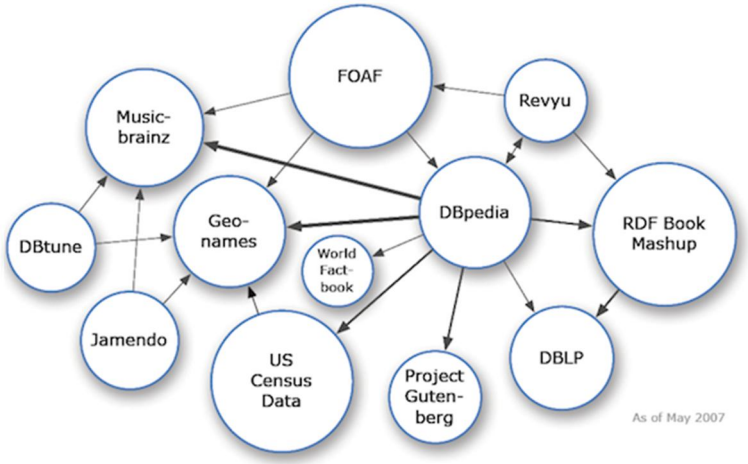
Anlamsal ağ kapsamında meta arama, mevcut arama motorları üzerinden değerlendirilmemekte, ancak veri setleri üzerinde meta veri oluşumu çalışmaları bulunmaktadır (Alexander ve ark., 2011). Federe sorgu sistemleri (Buil-Aranda, 2012) için gerekli görülen veri kümelerini tanımlayıcı meta veri analizleri, doğru bilgiye ulaşım için bir yönlendirme kriteri sunmaktadır. Örneğin, Splendid federe sorgulama projesi VoID meta-tanımlama standartları[7] kullanarak sorgu optimizasyonu ve dağıtımı yapmaktadır (Grlitz ve Staab, 2011). Anlamsal ağlar için meta-veri toplama yöntemleri çoğunlukla iki ana yöntem kullanır: otomatik internet tarama ve topluluk çalışması. Üç temel meta veri analizi projesi olan LOD Cloud (Cyganiak ve Jentzsch, 2017), LodStats (Ermilov ve ark., 2016) ve Sparqles (Vandenbussche ve ark., 2013), veri analizinde kullandıkları bağlantılı veri kümelerini internet toplulukları üzerinden toplamış; LOD (Cyganiak ve Jentzsch, 2017) çalışması buna ek olarak LDSpider (Isele ve ark., 2010) kullanarak otomatik veri toplama yöntemi kullanmıştır.

2.2. Bağlantılı veri (Linked Data)

Bağlantılı veri kavramı (Berners-Lee, 2006) anlamsal ağ standartlarına[8] göre yayınlanmış veri kümelerinin birbirleriyle bağlantılı yeni nesil veri kümeleri haline dönüştürülmesi için oluşturulmuştur. Bağlantılı verinin büyüklüğü ve kapsamının anlaşılması için geliştirilen LOD Cloud projesi (Cyganiak ve Jentzsch, 2017), bağlantılı verinin kuşbakışı bir diyagramını oluşturmayı amaçlamıştır. Çoğunlukla DBpedia veri kümesinin merkezde durduğu bu diyagramda tüm bağlantılı veri kümeleri ve birbirleri arasında olan bağlantıları belirtilmiştir. Şekil 2.1'de görüldüğü şekliyle 2007 yılında 12

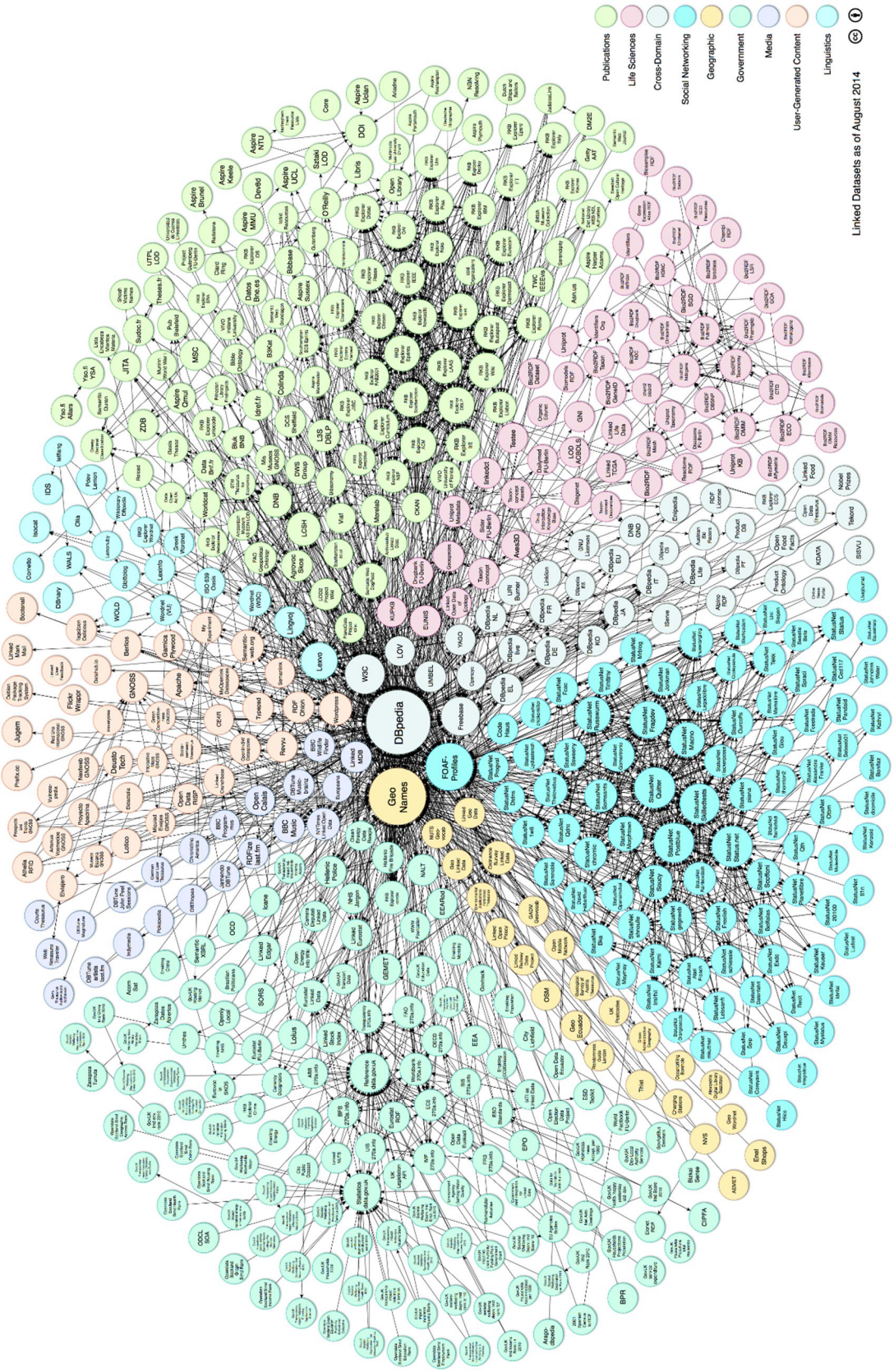
[6] http://www.webcrawler.com/
[7] http://www.w3.org/TR/void/
[8] http://www.w3.org/standards/semanticweb/

olan veri kümesi sayısı, Şekil 2.2'de görüldüğü gibi, tez çalışmasının başladığı 2014 yılı itibariyle 570 olarak belirtilmiştir.



Şekil 2.1. Linking Open Data Cloud Diyagramı 2007[9]

[9]http://lod-cloud.net/versions/2007-05-01/lod-cloud.png

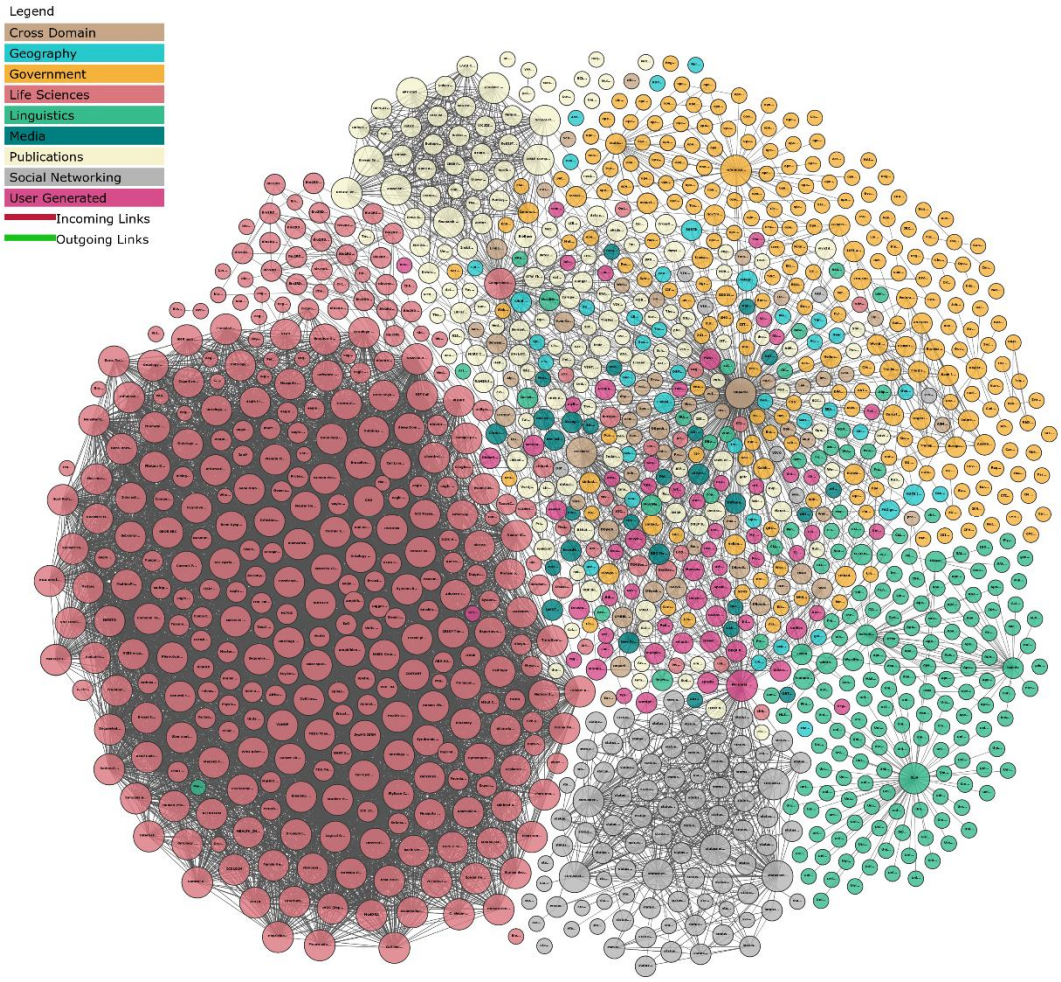


Şekil 2.2. Linking Open Data Cloud Diyagramı 2014[10]

En son 2017 yılında yayınlanan (Şekil 2.3) diyagrama göre 1163 veri kümesi

[10]http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.png

listelenmektedir. Bu gelişim sürecinde gerek devlet kurumları tarafından (data.gov.uk[11], U.S.data.gov[12] vb.), gerekse özel kuruluşlar tarafından (BBC Things [13] Thomson Routers [14] The New York Times[15] vb.) birçok bağlantılı veri kümesi yayınlanmıştır. Yayınlanan bağlantılı veri kümelerinin artışı birçok veri erişim ve arama çalışmasının (örn. Tabulator (Berners-lee ve ark., 2006), Openlink Data Explorer[16], Sig.ma (Tummarello ve ark., 2010)) ve devamında arama motorlarının (örn. Swoogle (Finin ve ark., 2004), Falcons (Cheng ve Qu, 2009), Sindice (Campinas ve Ceccarelli, 2011), SWSE (Hogan ve ark., 2011)) ortaya çıkmasına sebep olmuş ve mevcut veri kümeleri üzerinde meta-veri analizi ihtiyacını artırmıştır.



Şekil 2.3. Linking Open Data Cloud Diyagramı 2017[17]

[11] http://data.gov.uk/
[12] http://www.data.gov/
[13] http://www.bbc.co.uk/things/
[14] http://thomsonreuters.com/site/data-identifiers/
[15] http://developer.nytimes.com/docs/semantic_api
[16] http://ode.openlinksw.com/
[17] http://lod-cloud.net/versions/2017-08-22/lod.png

Bağlantılı veriler, web üzerinde URI'ler ve RDF kullanılarak birbirine bağlı veriler olarak yapılandırılmış anlamsal ağ verilerini ifade etmek için kullanılan bir terimdir. Bizer (Bizer ve ark., 2009) bağlantılı verileri, internetteki veri kaynaklarını birbirine bağlamanın bir yolu olarak açıklamaktadır, ki böylece bu veriler makine tarafından okunabilir, anlamsal olarak açıklama yapılabilir ve diğer veri kaynaklarına bağlanabilir olmaktadır. Bağlantılı veri yayımlama için temel standart (Berners-Lee, 2006), verilerin diğer internet içeriğiyle olduğu gibi URI'leri kullanarak veya RDF modelini kullanarak birbirine bağlanmasını önermektedir. Bağlantılı veri kaynakları ya internette RDF belgeleri veya SPARQL uç noktaları olarak yayınlanmaktadır (Bizer ve ark., 2009). Bağlantılı bir veri kaynağı, internette bağlı veri yayımlama ilkelerini izleyerek yayınlanırsa buna "Bağlantılı Açık Veri" (LOD) adı verilmektedir. Belirli kriterlere uygun olduğu sürece, "Açık Veri Bağlantısı Projesi"'ne (LOD Cloud) dahil edilmesi için bir LOD kaynağı bulunmaktadır. LOD Cloud'da tüm veri kaynakları meta tanımlarıyla sınıflandırılmakta ve tanımlanmaktadır. Bu meta tanımlamaları sağlamak için, VoID sözlüğü tanımlamaları yaygın olarak kullanılmaktadır (Alexander ve Hausenblas, 2009). VoID sözlüğü, bağlantılı veri kaynaklarını tanımlamak için belirli terim ve kalıpları önermektedir. Örneğin, bir veri kümesinin SPARQL uç noktası URL'si, VoID sözlüğündeki *void:sparqlEndpoint* özelliği tarafından ifade edilebilir. Bunun yanında, veri kümeleri ile ilgili istatistiksel veriler de VoID özelliklerini kullanarak ifade edilebilir. Örneğin, üçlü sayı (void:triples), varlıkların sayısı (void:entities), sınıfların sayısı (void:classes) vb.. Bu tez çalışmasında toplanan bağlantılı veri kaynaklarını incelemek ve karşılaştırmak için bu istatistiksel tanımlar kullanılmaktadır.

2.3. Bağlantılı Meta-veri (Linked Meta data)

Bağlantılı veri kaynakları hakkındaki meta veri, bu kaynakları tanımlamak amacıyla kullanılmaktadır. Bu bağlamda farklı kapsamlarda kullanılmak üzere birçok çalışma bulunmaktadır. Bu çalışmalara örnek olarak; "Web Video Text Tracks" (Steiner ve Mhleisen, 2014), "SIOC online community datasets" (Bojars ve ark., 2008) , "Web RDFa statistics" (Pound ve ark., 2010), "research and survey data statistics" (Bosch ve ark., 2013) gösterilebilir. Tüm bu alanların özelleşmiş kendi karakteristikleri, alana özel meta bilgileri bulunmaktadır. Belirtilen çalışmalarda, kendi kapsamları özelinde yaratılmış meta-tanımlamalar bulunmaktadır. Bunun yanında, alan bağımsız

meta-tanımlamalar yapabilmek amacıyla ise VoID sözlüğü (Alexander ve ark., 2011) oluşturulmuştur.

VoID veri sözlüğünü kullanarak her türden bağlantılı veri kümesi hakkında meta veriler anlamsal olarak tanımlanabilmektedir. Çizelge 2.1'de örneklenen VoID sözlüğü özellikleri, istatistiksel analiz sonuçlarının saklanabileceği özelliklerden bir kısmını oluşturmaktadır. VoID veri sözlüğü kullanarak oluşturulmuş çalışmalardan birisi Sparqles projesidir (Vandenbussche, Aranda, Hogan, & Umbrich, 2013). Bu projede veri kümelerinin erişilebilirlik, performans, birlikte çalışabilirlik ve keşfedilebilirlik analizi yapılmakta ve tespit edilen tüm veri kümeleri için meta veriler çıkartılmaktadır (Buil-Aranda ve Hogan, 2013). Bir diğer meta veri çalışması ise, veri setlerini haftalık taramalarla analiz eden ve web üzerinden[18] analiz verilerini yayınlayan Dynamic Linked Data Observatory (Kafer ve ark., 2013) projesidir.

Çizelge 2.1. VoID sözlüğü istatistiksel analiz özellikleri (Alexander ve ark., 2011)

Özellik	Amaç
void:triples	Toplam üçlü (triple)
void:entities	Toplam varlık (entity)
void:classes	Toplam sınıf (class)
void:properties	Toplam özellik (property)
void:distinctSubjects	Toplam tekil özne (subject) sayısı
void:distinctObjects	Toplam tekil nesne (object) sayısı
void:documents	Toplam döküman sayısı

Görsel bir meta veri çalışması olarak LOD Cloud projesi bulut diyagramı, bağlantılı veri setlerinin kuşbakışı bağlantı, büyüklük ve kapsam analizini yapmaktadır. Veri kümeleri arasındaki bağlantıları tespit etmek için ise "rdf:sameAs"[19] özelliklerini kullanmakta ve başka bir veri kümesine referans gösterilen varlıklar tespit edilmektedir. En güncel veri kümesi analizi Nisan 2014 tarihinde gerçekleştirilmiş (Cyganiak ve Jentzsch, 2014) ve istatistiksel sonuçlar rapor halinde yayınlanmıştır[20].

Çizelge 2.2'de tüm bağlantılı veri koleksiyonları ve analiz projeleri listelenmiş ve bu çalışmalarla ilgili veri toplama, veri tanımlama ve veri biçimi detayları gösterilmiştir. Datahub koleksiyonu, Sparqles ve LOD Cloud projeleri için veri deposu olarak kullanılmaktadır. Ancak yapılan ön çalışmalar sonucunda veri deposu olarak

[18] http://swse.deri.org/dyldo/
[19] http://www.w3.org/TR/owl-ref/#sameAs-def
[20] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

kullanan projelerden farklı sonuçlar da içerebildiği görüldüğünden (Şekil 3.1) araştırmaya dahil edilmesine karar verilmiştir.

Çizelge 2.2. Mevcut Bağlantılı Veri Seti Koleksiyonları

	Veri Toplama	**Tanımlama**	**Biçim**
LOD Cloud (Cyganiak & Jentzsch, 2014)	Community, Crawling	VoID	Turtle
Sparqles (Vandenbussche et al., 2013)	Datahub	Web	JSON
LODStats (Auer et al., 2012)	Dbpedia	VoID, Datacube	Html Tablosu
Datahub[21]	Community	CKAN API	CKAN API

2.4. Bağlantılı Veri Kaynakları Endeksleri ve Analizi

Bu bölümde bağlantılı veri kaynaklarını internet üzerinde toplama, endeksleme, analiz etme ve derecelendirme yöntemleriyle alakalı çalışmalar açıklanmaktadır. Bu bağlamda, ilk olarak bağlantılı veri derecelendirme yöntemleri, sonrasında SPARQL uç noktası kaynakları, bağlantılı veri kaynaklarının kategorizasyonu ve bu amaçla kullanılan Wordnet anlamsal sözlüğü açıklanmıştır.

2.4.1. Bağlantılı veri derecelendirme yöntemleri

Bu bölüm, yazarın doktora tez çalışması sırasında anlamsal ağ ve bağlantılı veri konuları ile bağlantılı derecelendirme yöntemlerini incelendiği çalışmadan (Yumusak ve ark., 2014) derlenmiştir. Derecelendirme yöntemleri ilgi alanlarına göre aşağıda listelenen beş farklı kategoride gruplanmıştır. Bu kategoriler; Ontoloji Derecelendirmesi, RDF Döküman derecelendirmesi, Çizge derecelendirmesi, Varlık Derecelendirmesi ve Kaynak derecelendirmesidir.

2.4.1.1. Ontoloji derecelendirmesi

Ontoloji, veriyi anlamsal ağ biçimlerinde saklarken kullanılan kavramsal tanımlamadır ve tanım dosyası biçiminde oluşturulmaktadır. Ontoloji derecelendirmesi yöntemleri iki şekilde incelenebilir; ontolojiler kullanarak derecelendirme veya ontolojilerin derecelendirilmesi. Ontolojileri kullanarak derecelendirme yapan ilk çalışma (Skoutas, Simitsis, & Sellis, 2007) web servisler üzerinde uygulanmıştır. İlgili çalışmada servis talepleri doğrultusunda anlamsal olarak reklam yönlendirilmesi

[21] http://datahub.io

amaçlanmaktadır. Derecelendirme, bir kapsam ontolojisi kullanılarak web servis parametreleri ile servis reklamlarının arasındaki anlamsal yakınlığın hesaplanması şeklinde gerçekleştirilmiştir. Bir diğer çalışmada (Stojanovic ve ark., 2003) ise sorgu sonuçlarının ontoloji temelli olarak çıkarımlama yöntemiyle derecelendirilmesi gerçekleştirilmiş ve bu yönteme "ağırlık eşleştirmesi" ismi verilmiştir. Benzer bir çalışmada (Rocha ve ark., 2004), bir ontolojide bulunan her ilişki örneğine bir sayısal ağırlık verilmesi önerilmiş ve bu yönteme ontoloji tabanlı derecelendirme ismi verilmiştir.

Bir diğer ontoloji çalışması olan OntoKhoj (Patel ve ark., 2003) projesinde, anlamsal ilişkilerin önem seviyesine göre önceliklendirilerek ağırlık değeri verilmesi önerilmiştir.

Swoogle (Finin ve ark., 2004) arama motoru ise ilk ontoloji arama motoru olarak devreye alınmıştır ve çalışmada ontoloji derecelendirmesi yöntemi ile arama sonuçlarının sıralanması gerçekleştirilmiştir. Çalışmada, bir ontolojinin derecesinin kullanım sayısı oranında artırılması önerilmiştir. Swoogle çalışmasında rastlantısal tarama modeli (Chebolu ve Melsted, 2008) yerine, rasyonel tarama modeli (Ding et al., 2005) kullanıldığı belirtilmiştir.

AKTiveRank (Alani ve ark., 2006) isimli bir çalışmada, ontolojilerin yapısal bazı metriklere göre derecelendirilmesi önerilmiştir. Kullanıcı kontrollü çok boyutlu ontoloji derecelendirme yöntemi olarak tasarlanan bu çalışmada, sınıf eşleşme ölçüsü, yoğunluk ölçüsü, anlamsal benzerlik ölçüsü ve aralık ölçüsü gibi değerleme ölçüleri kullanılmıştır. Watson (D'Aquin ve ark., 2011) isimli başka bir çalışmada, basit yapısal ve konu ilişkili kalite ölçüleri kullanarak veri üzerinde analiz gerçekleştirilerek hesaplanan değerlerin ontolojide saklanması önerilmiştir. Ortaya çıkan bu skorların, veri sorgulanırken bir sıralama parametresi olarak kullanılması önerilmektedir.

2.4.1.2. RDF belge derecelendirmesi

RDF terimi, anlamsal ağlarda veri gösterimi ve değiş tokuşunda kullanılmak üzere tasarlanmış bir belge biçimi standartıdır[22]. RDF belge derecelendirmesi yöntemleri üç çeşittir: Belge kaynağının bir bütün olarak derecelendirilmesi, RDF içeriklerinin derecelendirilmesi ve kaynak açıklamalarının derecelendirilmesi.

[22]http://www.w3.org/RDF/

RDF kaynakları hakkındaki ilk çalışmalardan birisi olan QuizRDF (Davies ve Weeks, 2004), RDF arama motoru olarak tasarlanmıştır. RDF kaynaklarının derecelendirmesi Tf-Idf skorlama ile gerçekleştirilmesi önerilmiştir.

Bir diğer çalışmada ise RDF belge içeriklerinin kapsam bağımsız bir şekilde derecelendirilmesi önerilmiştir (Bai ve ark., 2008). RDF ifadelerinin derecelendirilmesi, başka bir çalışmada konu-ilişkili cümlelerin ve sorgu-ilişkili cümlelerin derecelendirmesi olarak iki şekilde ele alınmıştır (Bai et al., 2008). Sig.ma (Tummarello ve ark., 2010), anlamsal ağlar için bir veri toplama ve görselleştirme aracı olarak geliştirilmiştir. Sig.ma içerisinde birçok veri kümesi (örn. Sindice (Tummarello ve ark., 2007), OKKAM[23], YBoss[24]) toplanmıştır. Sig.ma, veri elde etme sürecinde iki tip (kaynak tanımlama ve özellik) derecelendirme yöntemi kullanmaktadır. Kaynak tanımlama derecelendirmesinde RDF parçacıklarının içerisinde bulunan URI'lerin içerisinde geçen anahtar kelimeler derecelendirilmektedir (Tummarello ve ark., 2010). Aynı çalışmada bahsedilen özellik derecelendirmesi ise varlık derecelendirmesi olarak değerlendirildiğinden, ilgili bölümde açıklanmıştır.

SWSE (Harth ve ark., 2007) bir anlamsal arama motoru olarak PageRank (Brin ve Page, 1998) temelli bir derecelendirme algoritması kullanmaktadır. Bu algoritma ReConRank (Hogan ve ark., 2006) ismiyle anılmakta olup bağlam ve kaynak bazında derecelendirme yapmaktadır. Bu değerler RDF belgelerinin listelenmesi sırasında sıralama değeri olarak kullanılmaktadır.

2.4.1.3. Çizge derecelendirme

Bağlantılı veri çizge derecelendirme yöntemleri düğümler arası ilişkilerin derecelendirme yöntemlerine göre iki şekilde incelenebilir: Anlamsal derecelendirme ve istatistiksel derecelendirme. Anlamsal derecelendirme yöntemini kullanan Touchgraph (Aleman-Meza ve ark., 2005), anlamsal derecelendirme metrikleri (bağlam, kapsama ve güven) kullanılmasını önermektedir. İstatistiksel derecelendirme yöntemleri de "enderlik, popülerlik ve ilişki uzunluğu" (Aleman-Meza et al., 2005) gibi metrikler içermektedir. Bir diğer derecelendirme yöntemi olan SemRank (Anyanwu, Maduko, & Sheth, 2005), ilişkilerin anlamsal olarak derecelendirilmesi esasına göre skorlanmasını önermektedir.

[23] http://api.okkam.org/
[24] http://boss.yahoo.com/

RDF çizge derecelendirme amacıyla yapılan bir diğer çalışmada, "dil modelleme yaklaşımı" (Elbassuoni ve ark., 2009) adı altında derecelendirme yapılmakta ve "tanık üçlülerin sayısı (count of witness triples)" (Elbassuoni ve ark., 2009)'nı derecelendirme özelliği olarak kullanmaktadır.

ObjectRank (Balmin ve ark., 2004) isimli yöntemde, çizge içerisinde bulunan varlık düğümlerini derecelendirmek için PageRank (Brin ve Page, 1998) ve benzeri yöntemlerle evrensel bir değer hesaplanır. Hermes projesinde (Tran ve ark., 2009) EF-IDF isminde bir skorlama yöntemi kullanılmaktadır. Bu yöntem ile popülariteye ek olarak ayırt edicilik özelliği de hesaplanır. EF-IDF, temel olarak bir elemanın bir veri kümesinde bulunma sayısı üzerinden popülaritesini hesaplamaktadır (Tran ve ark., 2011). Semplore isimli çalışma, "ilişki temelli derecelendirme" (Wang ve ark., 2009) konusuna odaklanmış ve bir düğümün başka düğümlerle olan ilişkisine göre TF-IDF skorlama ve arka plan skorlamalarını birleştirerek bir derecelendirme yapmaktadır (Wang ve ark., 2009). Semplore projesinin ölçeklenebilirliği ve performansı bakımından kısıtlı olduğu belirtilmiştir (Delbru ve ark., 2012).

DBpedia[25] üzerinde popülarite temelli yeni bir derecelendirme yöntemi geliştirilmiştir (Mirizzi ve ark., 2010). Diğer PageRank (Brin ve Page, 1998) temelli çalışmaların aksine, "bağıl derecelendirme" (Mirizzi ve ark., 2010) isimli bir yöntem geliştirilmiş ve bu yöntemle bir düğümün derecesinin ilgili sorgu ve düğümlerle değişebilen bir değer olarak hesaplanması önerilmiştir. Bağlantılı veri ve veri çizgeleri üzerinde dağıtık endeksleme yöntemi olarak geliştirilen bir çalışmada, "top-N ranking and skylines" (Karnstedt ve ark., 2012) teknikleri kullanılmıştır. Triplerank (Franz ve ark., 2009) yönteminde 3 boyutlu tensör kullanılarak RDF çizgesinin özne, nesne ve nitelik özelliklerinin ayrı ayrı puanlaması önerilmiştir. Anlamsal ağ veri modelinin Triplerank yöntemi için önemi, düzensiz anlamsal bağlar için katı olmayan bir gösterim yöntemi olarak tanımlanmasıdır (Franz ve ark., 2009). RDFXpress (Elbassuoni ve ark., 2012) projesinde, RDF verilerinin aranması amaçlanmaktadır. Bu proje bünyesinde kullanılan derecelendirme yöntemi, yazarın RDF çizge araması üzerine yapmış olduğu eski bir çalışmasından alınmıştır (Elbassuoni ve ark., 2010). Başka benzer çalışmalarda ise (Kasneci ve ark., 2008; Elbassuoni ve ark., 2009; Elbassuoni ve Blanco, 2011), alt çizgelerin istatistiksel dil modellemesi yöntemleri (Ponte ve Croft, 1998) ile nasıl derecelendirildiği belirtilmiştir.

[25]http://dbpedia.org/

2.4.1.4. Varlık derecelendirmesi

Varlık derecelendirmesi yöntemleri iki şekilde incelenebilir: tekil bir varlığın derecelendirmesi veya varlık tipinin özelliğinin derecelendirilmesi. Bir anlamsal ağ arama motoru olarak SemSearch, bünyesinde varlık derecelendirmesi yöntemleri barındırmaktadır ve veri kaynağından bağımsız bir şekilde sorgunun ilgili varlıklarla hangi oranda bağlantılı olduğuna dair bir değer ataması yapmaktadır (Lei ve ark., 2006).

Falcons arama motoru da nesnelerin derecelendirilmesi olarak adlandırılan bir yöntemle, varlıkların popülariteleri ve sorgu ile ilgisine göre derecelendirme yapmaktadır (Cheng ve Qu, 2009). Bir diğer anlamsal arama motoru olan NAGA, yapısal verilerin derecelendirmesini "çıkarım güven" ve "sorgu uzunluğu" gibi etkenler kullanarak yapmaktadır (Kasneci ve ark., 2008).

Sig.ma projesinde, varlıkların özellikleri derecelendirilir. Derecelendirme metriği olarak bir özelliğin tip popülaritesi kullanılır ve bir özelliğin kaç adet kaynakta değerlerinin olduğu sayılarak hesaplanır (Tummarello ve ark., 2010). EntityAuthority (Stoyanovich ve ark., 2007) çalışması web sayfalarında gömülü olarak bulunan varlıkların tespit edilerek derecelendirilmesini önermektedir. Belirtilen yöntem HITS (Kleinberg, 1999) ve ObjectRank (Balmin ve ark., 2004) yöntemlerine benzer olarak tasarlanmış ve değişik tipte düğümler ve matematiksel tanımlamaları için daha zengin bir yaklaşım sunmaktadır. WebOWL (Batzios ve Mitkas, 2012), bir anlamsal ağ arama motoru olarak OWL nesnelerini derecelendirmektedir. Derecelendirme, PageRank (Brin ve Page, 1998) ve bazı sezgisel yöntemlerden esinlenerek uygulanmıştır. TRank (Tonon ve Catasta, 2013) çalışmasında, tekil varlıkların derecelendirmesi yöntemi kullanılmamakta, bunun yerine varlık tiplerinin derecelendirilmesi yöntemi kullanılmaktadır. Bu amaçla, verilen bir bağlamda varlık tiplerinin derecelendirilmesi için yeni bir teknik geliştirilmiştir. ECSSE (Cyganiak ve ark., 2009) isimli çalışmada, kaynak tanımlaması ve özellik derecelendirmesi yöntemleri kullanılmıştır. Kaynak tanımlaması derecelendirmesi basit anahtar kelime eşlemesi tekniğini kullanmakta olup, özellik derecelendirmesi ise basit bir sayma temelli (özelliği kullanan kaynak sayısı) bir metrik kullanmıştır (Cyganiak ve ark., 2009). Wikipedia[26]'nın kategori yapısı kullanılarak varlıkları derecelendiren başka bir çalışmada (Kaptein ve Kamps, 2013), Wikipedia'nın insan destekli kategori yapısı derecelendirme için bir girdi olarak kullanılmıştır.

[26] http://www.wikipedia.com

Veri internetinde arama konulu bir literatür taramasında (Melo ve ark., 2013), varlık derecelendirme yöntemleri "INEX Entity Ranking Track" (Demartini ve ark., 2010) ve "TREC Entity Search Tracks" (Balog ve ark., 2010) başlıkları altında incelenmiştir. "The INEX Entity Ranking Track" (Demartini ve ark., 2010), "TREC Entity Search Tracks" (Craswell ve Soboroff, 2005; Balog ve ark., 2010) ve "Semantic Search Challenge"[27], veri interneti üzerinde kullanılan üç farklı değerlendirme platformu sunmaktadır.

2.4.1.5. Belge/Kaynak derecelendirmesi

Belge/Kaynak derecelendirme yöntemleri kapsam bağımsız bir şekilde belge veya belge kaynaklarını derecelendirmek için kullanılır. Belge derecelendirmesi çevrim içi veya çevrim dışı veri işleme uygulamalarında karşımıza çıkmaktadır.

Belge derecelendirmesinin temel kavramları; frekans derecelendirmesi, görünüş derecelendirmesi ve kosinüs ilgi derecelendirmesi olarak tanımlanmaktadır (Materne ve Sleightholme, 2013). Çevrimiçi bir uygulama örneği olan bir web tarayıcısında (Du ve Hai, 2013), web sayfalarının herhangi bir kavram ile olan ilişkilerinin yakınlığı üzerinden derecelendirme yapılmıştır. Yapısal veri alanlarının belge derecelendirmesinde nasıl kullanılabileceği BM25F (Robertson ve ark., 2004) ve PRM-S (Kim ve Croft, 2012) üzerinden bir çalışmada incelenmiştir (Melo ve ark., 2013). Bu çalışmada, bu yöntemlerin kullanıcıların derecelendirme yaparken aynı zamanda yazı ve meta-veriyi entegre etmelerine de olanak sağladığı belirtilmektedir.

Bir soru cevaplama aracı olan PowerAqua, belge derecelendirme yöntemleri kullanmaktadır. Bu araçta, belgeler sorgu ile anlamsal ilişkilerine göre değerlendirilip derecelendirilir (Fernandez ve ark., 2008). Daha önce de bahsedilen ve bir anlamsal ağ arama motoru olan Sindice, belgelerin bazı özel amaçlı kurallar belirlenerek bulundukları kapsama ve alan adlarına göre öncelik tanınarak derecelendirilmesini önermektedir (Tummarello ve ark., 2007).

"Naming authority matrix" (Harth ve ark., 2009) olarak nitelendirilen bir yöntemle kaynak derecelendirmesini öneren bir diğer çalışmada, "naming authority" çizgesi baz alınarak PageRank temelli bir derecelendirme önerilmiştir (Harth ve ark., 2009). Belge derecelendirmesi yöntemlerinin en önde gelen çalışması olan PageRank yöntemi Google (Brin ve Page, 1998) tarafından geliştirilmiştir. Web teknolojilerine en

[27] http://challenge.semanticweb.org

fazla etki eden bu çalışmadan esinlenerek geliştirilmiş ve bu bölümde bahsedilen çalışmalar Çizelge 2.3'de listelenmiştir.

Çizelge 2.3. Derecelendirme çalışmaları özeti (Yumusak ve ark., 2014)

Proje/Çalışma İsmi	Derecelendirme Kategorisi	İlgili Yayın
Swoogle	Ontoloji	(Finin et al., 2004)
OntoKhoj	Ontoloji	(Patel et al., 2003b)
ObjectRank	Çizge	(Balmin et al., 2004)
ReConRank	Çizge	(Hogan et al., 2006)
Hermes	Çizge	(Tran et al., 2009)
Using naming authority..	Çizge	(Harth et al., 2009)
Dirichlet PageRank.	Çizge	(Chung, Tsiatas, & Xu, 2011)
Triplerank	Çizge	(Franz et al., 2009)
Semplore	Çizge	(H. Wang et al., 2009)
SWSE	Çizge	(Hogan et al., 2011)
OWLRank	Varlık	(Batzios & Mitkas, 2012)
Ranking complex relations.	Varlık	(Aleman-Meza et al., 2005)
Global PageRank of web communities	Belge/Kaynak	(Davis & Dhillon, 2006)

2.4.2. SPARQL uç noktası kaynakları

Büyük bağlantılı veri kaynakları ile ilgili meta veriler CKAN[28] açık kaynak kodlu veri portalları üzerinde saklanabilmektedir. LOD projesi (Cyganiak ve Jentzsch, 2017) ve SPARQLES (Vandenbussche ve ark., 2017) projeleri, Datahub web projesinde veri kümeleri depolamak için CKAN kullanmaktadır. LODStats (Ermilov ve ark., 2016) projesi, web sitesinde bulunan farklı kaynaklardan toplanan bağlantılı veri kaynakları üzerinde istatistiksel bir analiz sunmaktadır. Bu tez çalışmasının sonucu ortaya çıkan SpEnD (Yumusak ve ark., 2017), meta arama teknikleri kullanarak Web'in genelinde yeni SPARQL uç noktalarını keşfetmeye odaklanmaktadır. SpEnD keşif motoru, diğer tüm veri setlerini de biriktirerek bağlantılı veri tüketicileri için SPARQL uç noktası kaynakları sunmaktadır.

2.4.3. Kategorizasyon ve konu belirleme

Belge tabanlı sistemlerdeki konu modelleme yaklaşımları (Tuarob ve ark., 2015) bir belgenin konusunu tanımlamak için metin tabanlı belge analizi için çeşitli imkanlar

[28]https://ckan.org/

sunmaktadır. Bu yaklaşımlar, web sayfalarında (Scaiella ve ark., 2012) ve bağlantılı veri kaynakları üzerinde (Roder ve ark., 2015) uygulanmaktadır. LOD veri kümelerinin otomatik sınıflandırılması (Meusel ve ark., 2015) üzerine çalışmalar bulunmasına rağmen, veri kümesi konusunun tanımlanması ve sınıflandırılması esas olarak elle seçme ve kategorilendirme ile yapılmaktadır (Cyganiak ve Jentzsch, 2017). Bu açıdan, LOD bulut diyagramı, CKAN veri yayıncıları tarafından manuel olarak girilen konu adlarına dayanan, bağlantılı veri kaynakları için kategori etiketleri içermektedir. LOD bulut kategorileri; yayınlar, yaşam bilimleri, alanlar arası, sosyal ağlar, coğrafi, hükümet, medya, kullanıcı tarafından üretilen içerik ve dilbilim (Cyganiak ve Jentzsch, 2017) olmak üzere dokuz kategori içermektedir. (Meusel & Sarca, 2015) 'ın belirttiği gibi, veri kümelerinin LOD bulut diyagramında olduğu gibi elle sınıflandırılması, %81,62 doğrulukla sonuçlanan istatistiksel bir sınıflamayla karşılaştırıldığında yanlış etiketlemelere yol açabilmektedir. Daha önceki bir yaklaşım (Ferrara ve ark., 2013) tarafından özelliklere dayalı bağlantılı veri sınıflandırması da geliştirilmiştir. Lalithsena ve ark.(Lalithsena ve ark., 2013) Freebase'i bir konu keşif aracı olarak kullanarak benzer bir yaklaşım geliştirmiştir ve bu yaklaşımı LOD veri kümelerine uygulamıştır. (Meusel ve ark., 2015) ve (Lalithsena ve ark., 2013) esas olarak LOD Bulutu üzerinde yoğunlaşırken, (Ferrara ve ark., 2013) kümeleme algoritmasını örnek olarak kümeleme yaklaşımı ile eşleşen genel bir özellik ile örneklendirmiştir.

2.4.4. Wordnet semantik sözlüğü

WordNet (Miller, 1995), İngilizce kelimeler için bir sözlük veritabanıdır ve kelimeler arasındaki semantik ilişkileri içermektedir. Semantik ilişkiler; eşanlamlılık, hipernimi, antonimi, hiponimi, meronimi, toponimi ve şartlı bağlanma (entailment) olarak tanımlanır (Miller, 1995). Hipernimi, temelde bir tür ilişkisi olarak tanımlanır ve bir kelimenin türünü tanımlar. Hipernimi'nin tersi olan Hiponimi ise tersine tür ilişkisi olarak tanımlanır. Örneğin, A, B'nin bir hipernimi olarak tanımlanırsa, her zaman B'nin bir tür A (Miller, 1995) kategorisine ait olduğu söylenebilir. Wordnet kitaplığını kullanarak, bir sözcüğün konu semantiği ilişkilerini istemek suretiyle bir sözcüğün olası başlığı da çıkarılabilir. Bir belgedeki her kelimenin eksiksiz bir analizi, belgenin konusunu öngörmemize veya belgeye ilişkin içerikle ilgili etiketler oluşturmamıza yardımcı olabilir.

3. MATERYAL VE YÖNTEM

Bu bölümde bağlantılı veri kaynaklarının toplanması ve analizi, SPARQL uç noktalarının tespiti ve sınıflandırması ile ilgili kullanılan araçlar ve yöntemler açıklanmaktadır.

3.1. Diğer Projelerden Veri Toplama ve Analiz

Bu tez çalışmasında üç farklı bağlantılı veri analizi projesinin veri setleri incelenmiş ve karşılaştırmalı analizi gerçekleştirilmiştir. Bahsi geçen projelerin analizi ve edinilmesi için kullanılan yöntem, tanım ve bilgiler aşağıda listelenmiştir.

- LOD Bulutu: Bu çalışma VoID sözlüğü kullanarak meta veri tasniflemesi yapmakta ve yayınlamaktadır (Cyganiak ve Jentzsch, 2017). Mevcut meta veri kümeleri Turtle[29] dosya formatında sunulmaktadır.
- Lodstats (Ermilov ve ark., 2016): Bu çalışmada veri kümesini almak için herhangi bir yöntem bulunamamıştır. Veri kümesinin sorgulanabilmesi için SPARQL uç noktası bağlantısı olmasına rağmen çalışmamaktadır. Bu çalışma ile ilgili veri setine ulaşabilmek için HTML tarama yöntemleri kullanılarak web sitesinden ham veri olarak bilgilere erişilmesi amaçlanmaktadır.
- Sparqles (Vandenbussche ve ark., 2013): Mevcut veri kümesi, projenin web sitesi üzerinde tanımlanan sparqles api[30] ara yüzü kullanılarak JSON formatında sunulmaktadır.
- Datahub[31]: Bu veri kümesi esasında bir analiz çalışması olmayıp Sparqles ve LOD Cloud çalışmalarında kullanıldığı belirtilen veri kümelerini içermektedir. Yapılan ön çalışmalar göstermiştir ki bu veri seti bir kaynak veri çalışması olmasının yanında en güncel verileri de içerisinde barındırması açısından analiz edilmeye değer bir konumdadır. Bu çalışmada CKAN API[32] ara yüzü ile sorgulama sağlanabilmekte ve sorgular JSON formatında dönüş almaktadır.

Bahsi geçen dört veri kümesi hakkındaki veri tanımlama ve yayınlama formatları Çizelge 3.1'de listelenmiştir.

[29]https://www.w3.org/TeamSubmission/turtle/
[30] http://sparqles.okfn.org/api
[31] http://datahub.io
[32] http://docs.ckan.org/en/latest/api/index.html

Çizelge 3.1. Bağlantılı Veri Koleksiyonları Erişim Yöntemleri

Veri Kümesi	Tanımlama	Yayınlama
LOD Cloud	VoID	Turtle[33]
Sparqles	Custom	JSON[34]
LodStats	VoID, Datacube	Html Tablosu [35]
Datahub	CKAN API	JSON[36]

Çizelge 3.1'de listelenen veri kümeleri için ön çalışmalar yapılmış ve kaç adet SPARQL uç noktası içerdikleri tez çalışması sırasında sorgulanmıştır. Bu sorgulamalar sonucunda ortaya çıkan sonuçlar

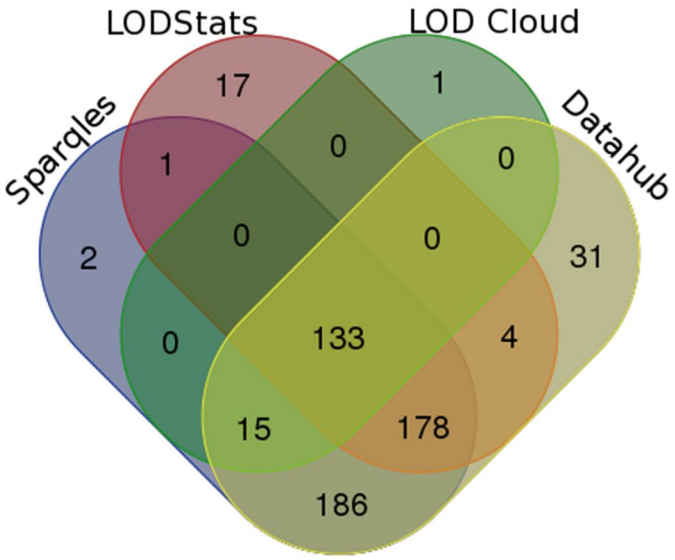
Çizelge 3.2.'de özetlenmiştir.

Çizelge 3.2. Bağlantılı veri koleksiyonlarının içerdiği SPARQL uç noktası sayısı

Kaynak	Toplam	Aktif	Pasif
LOD Cloud	149	75	74
Sparqles	524	256	268
LodStats	339	139	200
Datahub	556	261	295

Bu bağlamda elde edilen sonuçlar, tüm veri kümelerinde çevrimdışı bulunan birçok kayıt olduğunu ve bunların araştırılması gerekliliğini bir kez daha doğrulamıştır. Tüm veri kümelerinin birbirleri arasında ne kadar farklılık gösterdiklerini analiz edebilmek amacıyla Şekil 3.1'de gösterilen şemanın tez çalışmasının sonuçlarıyla karşılaştırması yapılmış olup, sonuçlar "5.1. Tespit Edilen SPARQL Uç Noktalarının Mevcut Listeler ile Karşılaştırması" bölümünde gösterilmektedir.

[33]http://lod-cloud.net/data/void.ttl
[34]http://sparqles.okfn.org/api/endpoint/list
[35]http://stats.lod2.eu/rdfdocs
[36]http://datahub.io/api/3/action/resource_search?query=format:sparql



Şekil 3.1. Bağlantılı veri kümeleri Venn diyagramı (Yumusak ve ark., 2017)

3.2. Arama Motorları Sonuçları Üzerinden SPARQL Uç Noktası Tespiti

Arama motorları kullanarak birçok bilgiye ulaşabildiğimiz gibi, SPARQL uç noktalarına da erişebileceğimiz bağlantıları tespit edebileceğimiz öngörülmüştür. Bu amaçla arama motorları için ortak özellikler belirlenerek Şekil 3.2'de görülen XML şeması oluşturulmuş ve buna bağlı olarak Şekil 3.3'de görülen XML örneği yaratılmıştır. Belirtilen parametrelerle arama motorlarına sorgu göndermek amacıyla crawler4j[37] ve websphinx (Miller ve Bharat, 1998) gibi web tarama yazılımları denenmiş ve arama motorlarının engelleriyle karşılaşılmıştır. Örneğin, Google arama motoruna yönlendirilen arama sorguları, klasik web tarama tekniklerinde "Server returned HTTP response code: 403 for URL"[38] hata kodunu vermekte ve sorgulamalara izin vermemekte olduğu tespit edilmiştir. Bu kısıtlamayı aşabilmek ve farklı internet tarayıcıları üzerinden arama simülasyonu yapabilmek amacıyla HtmlUnit[39] web tarayıcı kütüphanesi belirlenmiş ve örnek uygulamalar yapılmıştır.

[37]https://code.google.com/p/crawler4j/
[38]http://www.w3.org/Protocols/HTTP/HTRESP.html
[39]http://htmlunit.sourceforge.net/

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <xs:element name="SearchEngine">
    <xs:complexType>
        <xs:sequence>
            <xs:element name="name" type="xs:string"/>
            <xs:element name="excludedKeywords" type="xs:string"/>
            <xs:element name="baseUrl" type="xs:string"/>
            <xs:element name="queryTextBoxName" type="xs:string"/>
            <xs:element name="submitButtonId" type="xs:string"/>
            <xs:element name="submitButtonName" type="xs:string"/>
            <xs:element name="defaultBrowser" type="xs:string"/>
            <xs:element name="nextButtonIdentifier" type="xs:string"/>
            <xs:element name="useUrlRedirection" type="xs:boolean"/>
            <xs:element name="waitIntervalMs" type="xs:int"/>
        </xs:sequence>
    </xs:complexType>
    </xs:element>
</xs:schema>
```

Şekil 3.2. Arama Motoru Objesi XML Şeması

```
<searchEngine>
<name>yahoo</name>
<excludedWords>yahoo|bing|yimg|zenfs</excludedWords>
<baseUrl>https://www.yahoo.com</baseUrl>
<queryTextBoxName>p</queryTextBoxName>
<submitButtonId>search-submit</submitButtonId>
<submitButtonName></submitButtonName>
<defaultBrowser></defaultBrowser>
<nextButtonIdentifier>Next</nextButtonIdentifier>
<useUrlRedirection>false</useUrlRedirection>
<waitIntervalMs>1000</waitIntervalMs>
</searchEngine>
```

Şekil 3.3. Arama Motoru Objesi Örnek XML Bloğu (Yumusak ve ark., 2017)

3.2.1. Meta-Tarama için arama kelimelerinin oluşturulması

Arama motorları üzerinden sorgulanacak sözcük gruplarının belirlenmesi amacıyla mevcut SPARQL uç noktaları üzerinde kelime analizleri yapılarak Çizelge 3.3'de örneklenen arama sorguları oluşturulmuştur.

Çizelge 3.3. Arama Sorguları

Arama Sorgusu	Açıklaması
sparql	sparql kelimesini içeren
sparql -language	sparql kelimesini içeren ve language kelimesini içermeyen
"sparql endpoint"	tam ifade
allintitle: sparql data	sayfa başlığında sparql ve/veya data kelimesi olan
allinurl: sparql data	URL adresinde sparql ve/veya data kelimesi olan
intitle:sparql	başlığında sparql olan
inurl:sparql	URL adresinde sparql olan
"Virtuoso SPARQL Query Editor"	tam ifade
inurl:PoolParty inurl:sparql	URL adresinde PoolParty ve sparql olan
"sparql endpoint" site:edu	alan adı edu olan ve içerisinde "sparql endpoint" tam ifadesi olan
"sparql endpoint" site:gov	alan adı gov olan ve "sparql endpoint" tam ifadesi olan

Bu çizelgenin oluşturulabilmesi için bağlantılı veri havuzları (LOD Cloud, SPARQLES, LODStats ve DataHub) taranarak SPARQL uç noktası olarak yayında olan HTML sayfaları toplanmıştır. Bu SPARQL uç noktaları meta tarama anahtar kelimelerini oluşturmak için kullanılmıştır. En sık kullanılan anahtar kelimeler şunlardır: Sparql, query, rdf, virtuoso, openlink, inference, endpoint. Tek kelimelerin yanı sıra, yukarıda belirtilen anahtar kelimelerle birlikte kullanılan yaygın HTML etiketleri şunlardır: label, a, span, header, meta, h1, h2, h3, li, dt, p ve option. Bu kelimeler birleştirilerek, meta tarama arama anahtar kelimeleri ve belirli arama yönergelerinden oluşan bir liste hazırlanmıştır.

3.2.2. Bağlantı çıkarım kriteri ve filtreleme

Arama motorları sonuçlarından SPARQL uç noktası bağlantılarının çıkarımı ve filtrelenmesi için işlemler gerekmektedir. Çizelge 3.4'de, bağlantıların çıkarımı ve filtrelenmesi için takip edilecek prosedürler listelenmektedir. İlk aşamada, tanımlanmış XML parametreleri (Şekil 3.3) üzerinden arama motoru objesi yaratılmaktadır. Arama motoru üzerinden yapılan otomatik arama sonuçları taranarak bağlantılar çıkarılmaktadır. Bağlantılar çıkartılırken ilgisiz dosya tipleri (pdf, gif, jpeg vb.) ve bağlantı adresinde dışlanan kelimeleri içeren bağlantılar kapsam dışı bırakılır. Her bir sayfa için aynı işlem yapılarak bir sonraki sayfaya geçilir.

Çizelge 3.4. Arama sonuçlarının alınması algoritması sözde kodu

```
Algoritma 1: Arama Motoru Sonuçlarının Toplanması (string S, SearchEngine SE, int n)
 1: /* SE objesi içerisinde S arama sorgusu çalıştırılır ve ilk n sayfada bulunan bağlantılar çıkarılır */
 2: parametreDizisi = XMLDosyasindanParametreleriAl (SE)
 3: mevcutSayfa = ilkSayfayiAl (parametreDizisi, s, SE)
 4:
 5: for i← 1, n do
 6:         URLListesi = HTMLKaynagindanURLCikar (mevcutSayfa)
 7:         URLListesi.AlakasizDosyaTipleriniCikar ()
 8:         URLListesi.Dislanan KelimesiOlanURLCikar ()
 9:         URLListesi.kaydet ()
10:         mevcutSayfa = SonrakiSayfayiAl (mevcutSayfa)
11: end for
```

SPARQL uç noktası olmaya aday bağlantıların Çizelge 3.4'de belirtilen şekilde çıkarılmasından sonra, Çizelge 3.5'de bu bağlantıların SPARQL uç noktası olup olmadığının tespiti yapılır.

Çizelge 3.5. URL analizi algoritmasının sözde kodu

```
Algoritma 2: URL Analizi (string URL)
 1: /* URL parametresinin SPARQL uç noktası olup olmadığının kontrolü*/
 2:
 3: while URLVarMi () do
 4:         url =yeniUrlAl()
 5:         if eskiURLListesindeVarMi (url)
 6:                 UrlTipi = eskiURLTipi
 7:         else if SparqlSonlandirmaNoktasiMi ()
 8:                 SparqlSonlandirmaNoktasiOlarakIsaretle (true)
 9:         else
10:                 SparqlSonlandirmaNoktasiOlarakIsaretle (false)
11: end while
```

Çizelge 3.5, çıkarımı yapılan tüm bağlantıları SPARQL uç noktası olup olmadığı şeklinde test eder. Daha önce negatif veya pozitif tespiti yapılan bağlantıların tekrar test edilmesini önlemek amacıyla geçmiş bağlantılar kontrol edilir. Geçmiş kayıtlarda bulunmadığı tespit edilen tüm bağlantılara Çizelge 3.6'da gösterilen basit SPARQL sorgusu gönderilerek yanıt beklenir. Bağlantı adresinden SPARQL sorgu cevabı alınırsa sonlandırma noktası olarak işaretlenir.

Çizelge 3.6. Basit SPARQL sorgusu

SPARQL Sorgusu
1: SELECT DISTINCT ?Concept WHERE ([] a ?Concept)LIMIT 100

3.2.3. Alan adı öğrenmesi

Tüm SPARQL uç noktası keşif aramalarından sonra, bulunan alan adları için derin aramalar yapılabilmesi amacıyla yeni arama görevleri yaratılmaktadır. Yapılan ön çalışmalar sonucunda arama motorlarının sonuç listelerinde çıkan web sitelerinde SPARQL uç noktası bulunmasına rağmen SPARQL uç noktası olmayan sayfalarının listelenebildiği, bu gibi sitelerdeki sonlandırma noktalarının da ancak web sitesi içerisinde tekrar arama yapma yöntemiyle çıkarılabileceği öngörülmüştür. Bu amaçla yapılacak aramalar "site" anahtar sözcüğü kullanılarak "sparql site:alanAdi.com" şeklinde tasarlanmıştır. Çizelge 3.7'de önceden tespit edilen alan adlarının tekrar aranması için tasarlanan yöntemin sözde kodu bulunmaktadır.

Çizelge 3.7. Önceden tespit edilen alan adlarının tekrar aranması sözde kodu

Algoritma 3: Eski Alan Adlarından Yeni Arama Oluşturma (List AlanAdlari)
1: /* Aramalardan tespit edilen tüm tekil alan adlarından yeni arama görevi yaratır
2: Yaratılan arama görevi "sparql site:AlanAdi.uzanti" şeklinde düzenlenir.
3: */
4:
5: **for each** tekilAlanAdi: AlanAdlari **do**
6: yeniAramaKuyruguOgesiYarat ("sparql site:" + tekilAlanAdi)
7: **end for each**

3.2.4. İstatistiksel analiz yöntemleri

Bağlantılı veri kaynakları hakkında istatistiksel meta analiz sonuçları üretebilmek amacıyla VoID (Alexander et al., 2011) sözlüğü incelenmiştir. VoID (Alexander et al., 2011) sözlüğünün arka plan çalışmaları sırasında kullanılan ve Çizelge 3.8'de listelenen istatistiksel SPARQL sorgu cümlecikleri, yapılacak olan istatistiksel analizler için temel teşkil etmektedir. Bu amaçla tespit edilen tüm SPARQL uç noktalarında belirtilen sorgu cümlecikleri çalıştırılmıştır.

Çizelge 3.8. İstatistiksel Sparql Sorguları[40]

ID	Sparql Sorgusu	Tanım
1	SELECT COUNT(*) ? s ? p ? o	üçlü (triples)
2	SELECT COUNT(distinct ? s) ? s a []	varlık(entities)
3	SELECT COUNT(DISTINCT ? s) ? s ? p ? o UNION ? o ? p ? s FILTER(! isBlank(? s) && ! isLiteral(? s))	farklı kaynak bağlantıları (distinct resource URIs)
4	SELECT COUNT(distinct ? o) ? s rdf:type ? o	farklı sınıflar (distinct classes)
5	SELECT count(distinct ? p) ? s ? p ? o	farklı yüklemler (distinct predicates)
6	SELECT COUNT(DISTINCT ? s) ? s ? p ? o	farklı özne düğümleri (distinct subject nodes)
7	SELECT COUNT(DISTINCT ? o) ? s ? p ? o filter(! isLiteral(? o))	farklı nesne düğümleri (distinct object nodes)

Çizelge 3.8'de listelenen sorguların uzak SPARQL uç noktalarında çalıştırılmasıyla, tüm veri kümelerinin üçlü (triple), varlık (entity), tekil kaynak bağlantısı (distinct resource urls), tekil sınıf, tekil yüklem (distinct predicates), tekil özne düğümü (distinct predicates) ve tekil obje düğümü sayılarının çıkarımı gerçekleştirilmektedir. Bu bilgiler bağlantılı veri kümelerinin her biri için büyüklük ve kapsam analizi sunmaktadır.

İstatistiksel analiz ön çalışmasında Çizelge 3.2'de adetleri listelenen veri kümelerinde bulunan SPARQL uç noktalarına (Toplam 731 sonlandırma noktası) Çizelge 3.8'de bulunan sorgular Apache Jena[41] kütüphanesi kullanılarak gönderilmiştir.

Bu bağlamda mevcut veri kümelerinde kayıtlı olan üçlü (triple), varlık (entity), tekil kaynak bağlantısı (distinct resource urls), tekil sınıf, tekil yüklem (distinct predicates), tekil özne düğümü (distinct predicates) ve tekil obje düğümü sayıları çıkarılmıştır.

3.3. SPARQL Uç Nokta URL'lerinin Sınıflandırılması

Bu bölümde SPARQL uç noktaların sınıflandırılması için kullanılan yöntemler

[40] https://code.google.com/p/void-impl/wiki/SPARQLQueriesForStatistics
[41] https://jena.apache.org/

açıklanmaktadır. Açıklanan yöntemler, tez süresince toplanan SPARQL uç nokta URL'leri üzerinde uygulanabildiği gibi, diğer tüm veri kümeleri üzerinde de uygulanabilmektedir. Bu bağlamda dört farklı veri koleksiyonu da (LOD Projesi, SPARQLES, LODStats ve Datahub) analiz edilebilmesi amacıyla listeye eklenmiş ve analiz için Şekil 5.2'de dağılımları gösterilen 1068 adet benzersiz SPARQL uç noktası URL'lerini içeren bir liste hazırlanmıştır.

3.3.1. Yazı içerik toplama

Apache Jena[42] kütüphanesi kullanılarak, SPARQL sorgulama ile bağlantılı veri kümelerindeki metin açıklamalarının toplanması işlemi gerçekleştirilmektedir. Bu şekilde sorgulanarak tüm SPARQL uç noktalarında metin içeriğinden oluşan rdfs: comment ve rdfs: label özellik değerleri elde edilebilmektedir. Uç noktalarda bulunan metin içeriğinin toplanması amacıyla aşağıdaki SPARQL sorguları kullanılmıştır:

SELECT DISTINCT ?o WHERE ?s rdfs:comment ?o
SELECT DISTINCT ?o WHERE ?s rdfs:label ?o

Toplanan ham metin verileri (yorumlar ve etiketler), tanınmayan karakterlerden temizlenmekte, ayrıştırılmakta ve kelime analizi için sözcüklere bölünmektedir. Toplanan veriler bir veritabanında saklanmaktadır.

3.3.2. Skorlama

Tf-Idf skorlaması (Salton, Wong, & Yang, 1975), bir belgede üzerinde bir kelimenin belgeler kümesindeki alaka düzeyini hesaplamak için kullanılan bir yöntemdir. Belge sınıflamasında (Sebastiani, 2002) frekans skorları, özellik seçim süreçlerini hassaslaştırmak için ek olarak kullanılabilmektedir (Jain & Zongker, 1997). Tez çalışmasında, Tf-Idf skorlaması SPARQL uç nokta sınıflandırma yöntemi için temel olarak kullanılmış ve Wordnet hipernim ve konu bilgileri kullanılarak yeni bir skorlama yöntemi geliştirilmiştir. Literatürde, Wordnet, doküman sınıflandırma görevlerinde sınıflandırma doğruluğunu iyileştirmek için kullanılmıştır (Du ve Hai, 2013). Bu bağlamda, her bir SPARQL uç noktasını bir belge olarak düşünerek, SPARQL uç noktalarını sınıflandırmak için anlamsal olarak önerilen yeni Tf-Idf

[42]https://jena.apache.org/

skorlaması geliştirilmiştir. Klasik Tf-Idf skorlamasına yapılan bu iyileştirmenin bir sonucu olarak anlamsal skorlama adı verilen Stf-Idf skorlaması geliştirilmiştir.

Daha detaylı bir anlatımla, klasik Tf-Idf skorlaması (Sparck Jones, 1972) Denklem 3.1'deki şekilde açıklanır:

$$Tf - Idf(t, d, D) = tf\ (t, d).idf(t, D) \tag{3.1}$$

Kısaltmalar şu şekilde açıklanmaktadır:

- D: belgeler kümesi
- d: tekil belge
- t: tekil terim

Bu ifadeden yola çıkarak önce "t" terimi "s" semantik terimle değiştirilmiş ve tüm terimler anlamsal olarak ilişkili terimlerle Denklem 3.2'de gösterilen şekilde değiştirilmiştir:

$$Stf - Idf\ (s, ds, Ds) = tf\ (s, ds).idf(s, Ds) \tag{3.2}$$

Kısaltmalar şu şekildedir:

- Ds: anlamsal olarak dönüştürülmüş belgeler
- ds: tekil anlamsal olarak dönüştürülmüş belge
- s: WordNet'ten elde edilmiş anlamsal terim

Bu bağlamda Ds, tüm kelimeleri WordNet hipernimleri veya konuları ile değiştirerek orijinal belgelerden oluşturulan belge kümesini temsil etmektedir. Örneğin, *compilers* kelimesi hipernimi olan *computer program'a* dönüştürülür veya *infection* kelimesi konusu olan *medicine'a* dönüştürülmektedir ve bu dönüşüm dökümandaki ilgili kelimeler ile değiştirilmektedir.

Ayrıca, bu skorlamanın klasik Tf-Idf skorlamasına olan etkisini ölçmek için, Stf-idf skorlamasının birleştirilmiş hali de denenmiş ve Ctf-Idf ismi verilerek Denklem 3.3'deki şekliyle hesaplanmaktadır:

$$Ctf - Idf(t, s, dc, Dc) = Tf - Idf(t, d, D).Stf - Idf\ (s, ds, Ds) \tag{3.3}$$

Kısaltmalar şu şekildedir:

- Dc: Orijinal kelimelerle birlikte hipernim ve konu etiketlerini içeren dökümanlar
- dc: Anlamsal etiketler içeren tekil doküman

Ctf-Idf skorlaması sınıflandırma algoritmalarına entegre edilirken, WordNet terimleri ile kaynak kelimelerin birleştirildiği dökümanların (Dc) Tf-Idf skorlamasına tabi tutulması yoluyla özellik vektörüne çevrimi şeklinde uygulanmıştır.

3.3.3. Bağlantılı veri kaynaklarının sınıflandırılması

Önerilen Tf-Idf skorlama yöntemlerinin (Stf-Idf ve Ctf-Idf) etkisini anlamak için, farklı sınıflandırma yöntemlerini uygulamadan önce özellik vektörlerini oluşturmak için skorlama fonksiyonları kullanılmaktadır. Bu bağlamda, daha önce de belirtildiği gibi SPARQL uç noktaları, sınıflandırılacak bağlantılı veri kaynakları olarak değerlendirilmiştir. Bağlantılı veri kaynakları, her veri kaynağına ait bir belge vektörü oluşturmak için kullanılmıştır. Daha sonra belge vektörleri eğitim kümesi olarak kullanılmıştır. Sınıf etiketleri (yayınlar, yaşam bilimleri, alanlar arası, sosyal ağlar, coğrafi, hükümet, medya, kullanıcı tarafından oluşturulan içerik ve dilbilim) olarak LOD Bulutu (Cyganiak & Jentzsch, 2014) kategorileri kullanılmıştır.

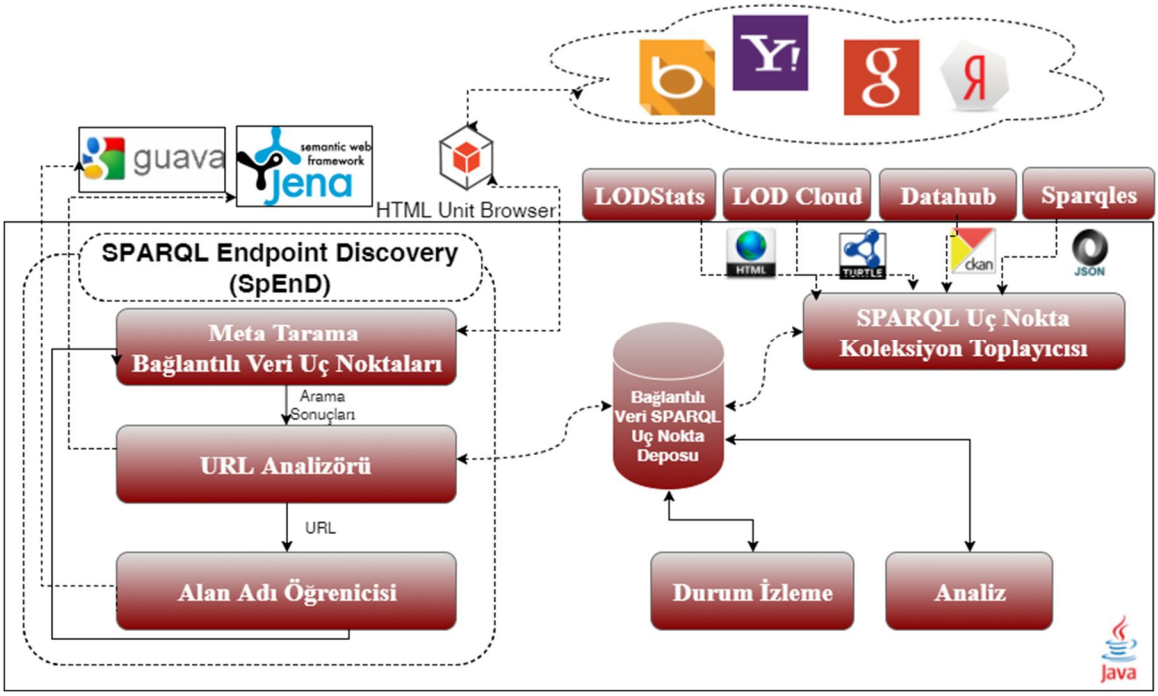
SPARQL uç noktalarının sayısı çok sınırlı olması sebebiyle, en doğru sınıflandırma sonuçlarını hesaplamak için "Leave-One-Out cross validation tekniği" (Pedregosa et al., 2011) kullanılmaktadır. Sınıflandırma algoritmaları için giriş parametreleri bu özel durum için ayarlanmış ve sonuçlar artırımsal özellik seçimi (H. Liu & Setiono, 1998) yöntemi kullanılarak hesaplanmıştır. Böylece, skorlama yönteminin etkisi, birçok farklı özellik ile denenebilmiştir.

Bu çalışmada yedi farklı sınıflandırma algoritması ile Python dilinde deneyler gerçekleştirilmiş ve sonuçları karşılaştırılmıştır. Bu sınıflandırma algoritmaları: Ada Boost, Decision Tree, Linear SVM, Naive Bayes, Nearest Neighbors, Random Forest, RBF SVM'dir.

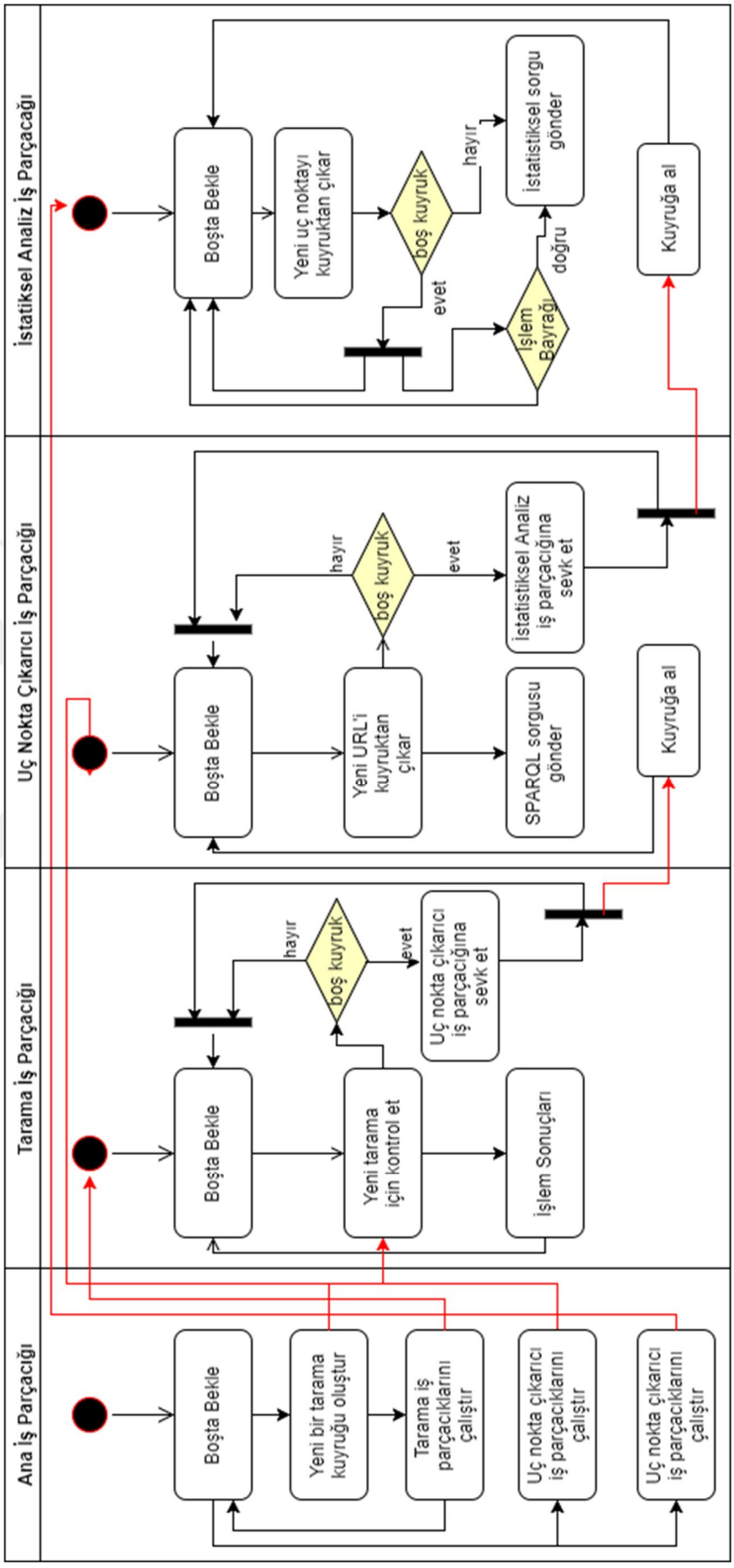
4. SpEnD META-TARAMA MOTORU UYGULAMASI

SpEnD uç nokta tespit sistemi, SPARQL uç noktaları için bir meta tarama, analiz ve yayınlama servisidir. Uygulamanın çıktıları canlı olarak paylaşılacak şekilde saklanabilmekte ve süreğen bir kontrol mekanizması ile tespit edilen SPARQL uç noktalarının güncel durumları bağlantılı veri tüketicilerine sunulabilmektedir. Sistemin mimarisi Şekil 4.1'de görülmektedir. Şekil 4.1'de görülebileceği üzere tarama işlemi Bing, Yahoo, Google ve Yandex gibi arama motorları üzerinden HtmlUnit tarama simülatörü kullanarak tarama gerçekleştirmektedir. İlk taramadan sonra tespit edilen tüm URL'ler analiz edilmek için, URL analizi aşamasına gönderilerek Jena Framework desteğiyle SPARQL uç noktası tespiti yapılmaktadır. Taramada bulunan tüm URL'ler, alan adı analizi için Google Guava kitaplıklarını kullanarak "Alan Adı Öğrenicisi" aşamasında tekrar taranmak üzere alan adlarına ayrıştırılmaktadır.

SpEnD, çok iş parçacıklı bir Java uygulaması olarak inşa edilmiştir. Şekil 4.1'de açıklanan bağlantılı veri bulma sürecinin aşamaları ana iş parçacığının dışında 3 paralel aşama olarak uygulanmaktadır: (a) tarayıcı, (b) uç nokta çıkarıcı ve (c) istatistiksel analizci. Şekil 4.2'de, bu iş parçacıkları arasındaki bilgi akışı bir etkinlik diyagramı olarak tanımlanmaktadır.



Şekil 4.1. SpEnD sistem diyagramı (Yumusak ve ark., 2017)



Şekil 4.2. İş Parçacığı aktivite diyagramı (Yumusak ve ark., 2017)

Ana iş parçacığı, kullanıcı etkileşimlerini ve çalışan iş parçacıklarını denetlemektedir. İşçi iş parçacıkları aşağıda açıklanmıştır:

- Tarayıcı: Arama motoru meta tarama işleri yapar ve bulduğu adresler ile uç nokta çıkarıcı iş parçacığını besler.
- Uç Nokta Çıkarıcı: Bulunan her adrese SPARQL sorguları göndererek tarayıcı iş parçacığı tarafından üretilen aday bağlantıların analizini gerçekleştirir.
- İstatistiksel Analiz: VoID tanımlarında listelenen istatistiksel SPARQL sorgularını kullanarak SPARQL uç noktalarını analiz eder. Sorgular, iş parçacığı çalıştığı sürece periyodik olarak gönderilir.

SpEnD sistemi bünyesinde, arama motorlarında meta tarama gerçekleştirmek, URL'leri analiz etmek ve keşfedilen SPARQL uç noktalarında istatistiksel analiz yapmak amacıyla "SPARQL Endpoint Crawler and Analyzer" (SPECAN) ismi verilen masaüstü uygulama geliştirilmiştir. Şekil 4.3, SpEnD masaüstü uygulamasının ana ekranını göstermektedir. Ekranın iki sekmesi vardır: (a) Crawler (Tarayıcı), ve (b) Analysis (Analiz).

Tarayıcı bölümünde, kullanıcı tarafından girilen arama sorgularını kullanarak çok iş parçacıklı arama motoru taraması gerçekleştirilir. Analiz bölümünde, keşfedilen SPARQL uç noktaları üzerinde istatistiksel analiz yapılır. Bu yazılım, tüm SPARQL uç noktalarının sürekli taranmasını ve analiz edilmesini sağlayacak şekilde döngüsel olarak çalışabilmektedir.

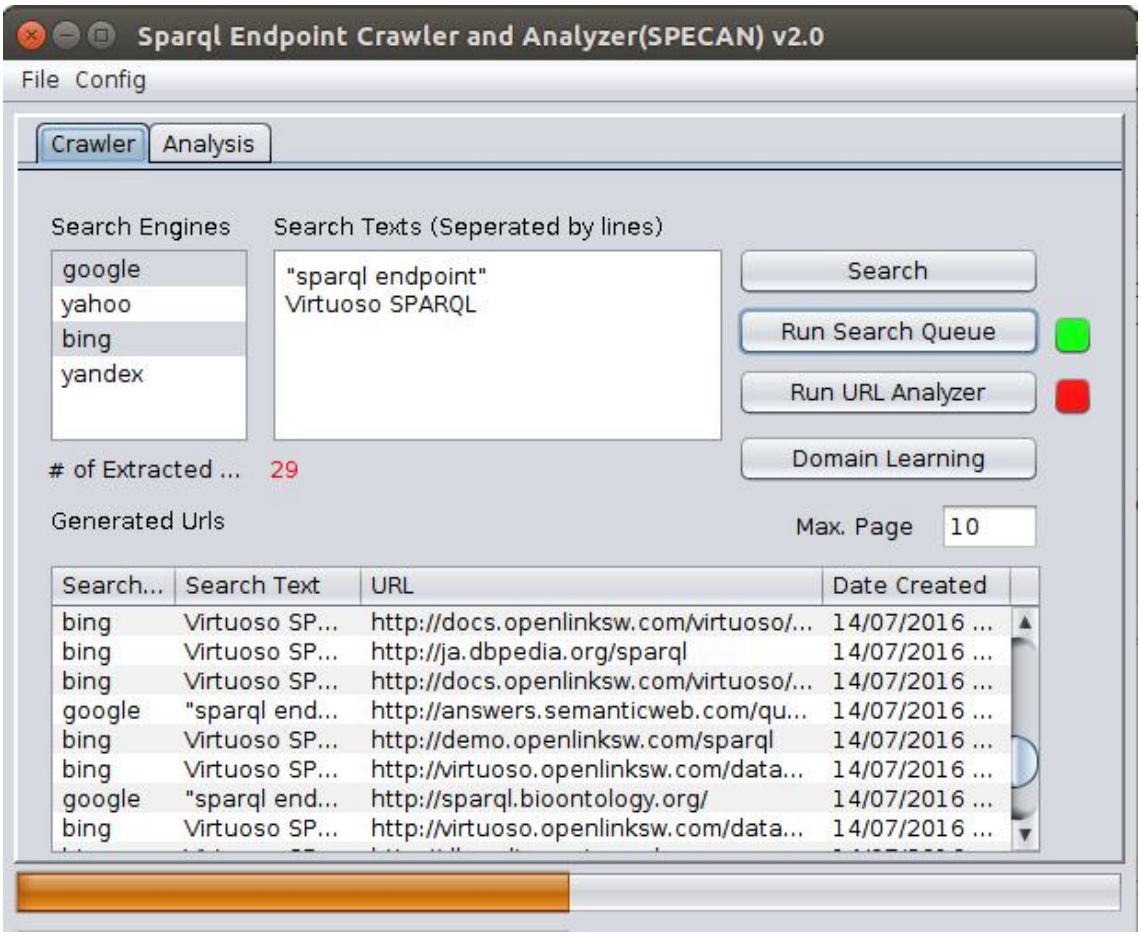
4.1. Tarayıcı Grafik Ara Yüzü

Bu ara yüzde (Şekil 4.3), arama motoru taraması ve URL analizi gerçekleştirilir. Ekran, arama anahtar kelimeleri ve sorguları için bir arama metin giriş kutusunu, seçilecek arama motorları için bir listeyi ve devam eden arama sonuçlarını görüntülemek için bir alt tabloyu içermektedir. Meta taramayı gerçekleştirmek için aşağıdaki adımlar gereklidir:

1. “Search” düğmesine tıklanarak, tüm arama anahtar kelimeleri ve tüm arama motorları için arama görevleri de dahil olmak üzere bir arama sırası oluşturulur. Yazılım, tarama motorunu çalıştırmadan önce birçok farklı kayıt üretilmesine olanak tanır.
2. “Run Search Queue” düğmesine tıklayarak, seçilen her arama motoru için bir arama iş parçacığı başlatılır. Arama iş parçacıkları, arama motorlarının

sınırlamasına uymakta ve Çizelge 3.4'de açıklanan algoritmayı kullanarak, aday URL'ler kuyruğuna sürekli bağlantılar çıkarabilmektedir.

3. “Run URL Analyzer” düğmesine tıklayarak aday link analizci iş parçacığı asenkron olarak çalıştırılabilir. Bu düğme, arama parçacıkları tarafından aday bir SPARQL uç nokta URL'si olarak eklenen her sayfayı ziyaret etmek için önceden tanımlanmış sayıda iş parçacığı oluşturur. Aday URL'leri Çizelge 3.5’de tanımlanan yönteme göre bitiş noktaları olarak işaretlenir.
4. Arama motoru tarama ve URL analiz adımlarından sonra, alan adı öğrenme yöntemi (bkz. Çizelge 3.5) “Domain Learning” düğmesi kullanılarak uygulanır.

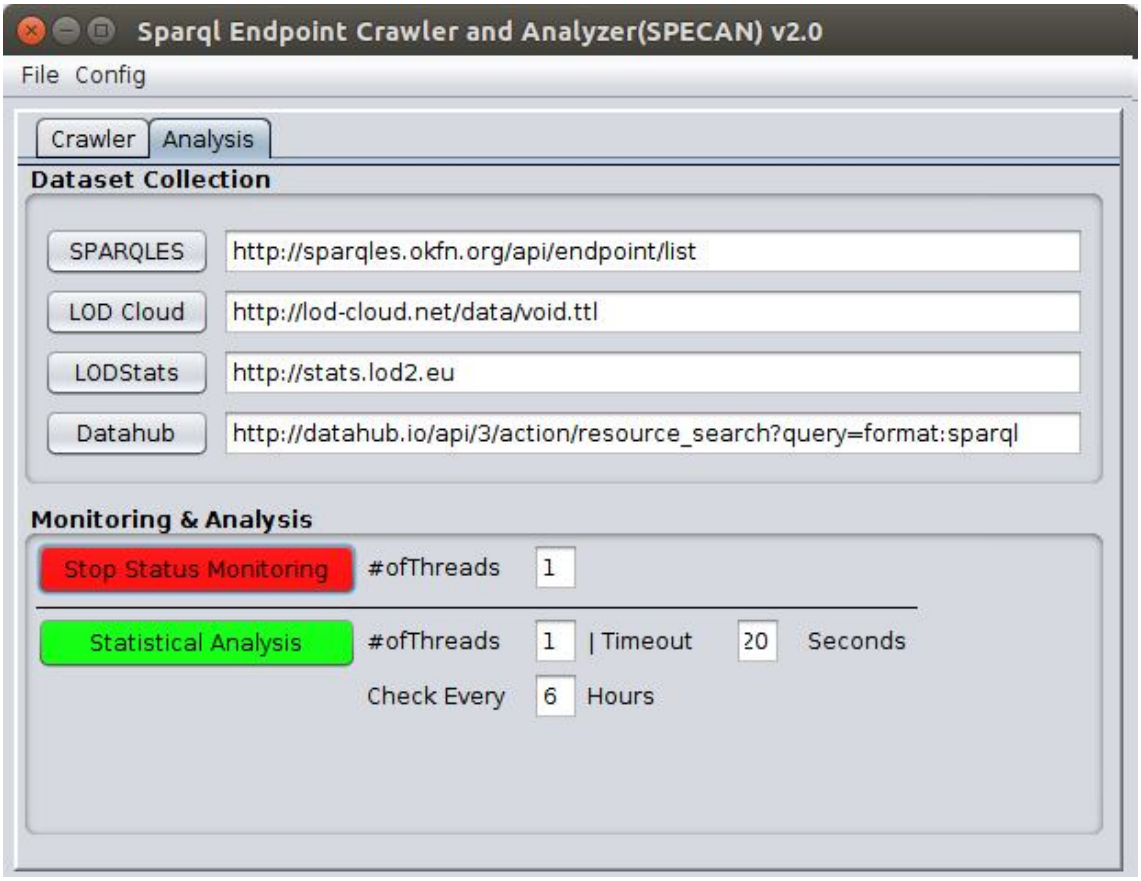


Şekil 4.3. SPECAN v2.0: SpEnD projesi masaüstü yazılımı tarama penceresi

4.2. Analiz Grafik Ara Yüzü

Analiz sekmesinde, diğer projelerden elde edilen SPARQL uç noktalarının çevrim içi olarak içe aktarımı işlemi gerçekleştirilmektedir (Şekil 4.4). Ek olarak (İzleme ve Analiz bölümünde), Çizelge 3.8'de listelenen SPARQL sorgularını kullanarak her bir SPARQL uç noktanın kullanılabilirliği ve her bir SPARQL uç

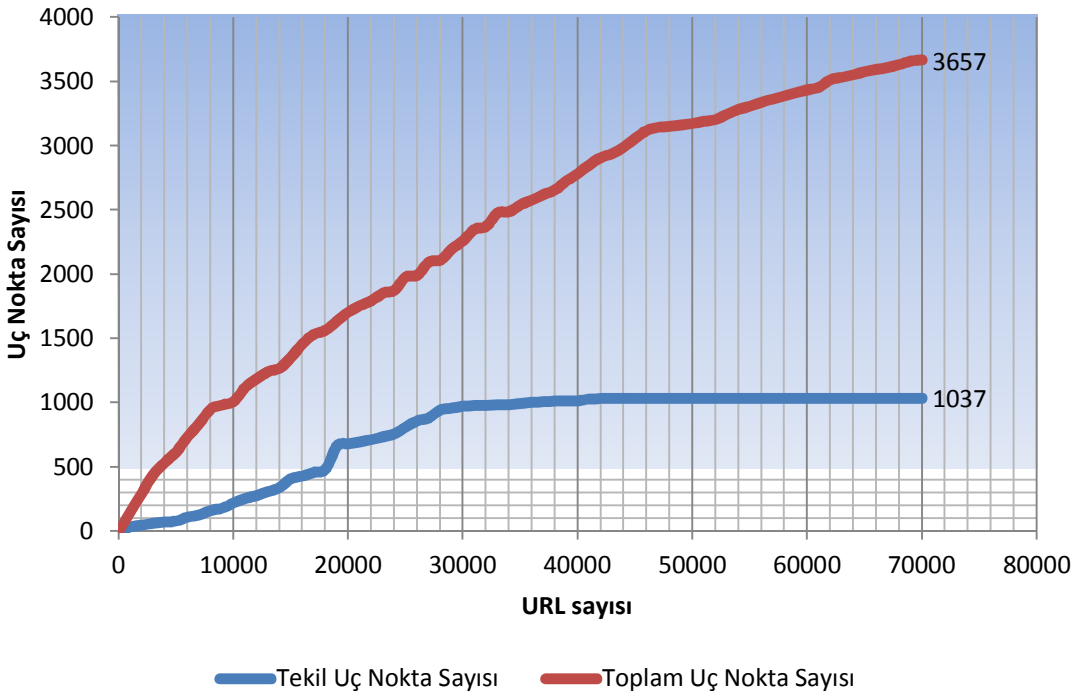
noktanın istatistiksel analizi işlemlerini gerçekleştiren sorgulamalar çalıştırılabilmektedir. Böylece internet üzerinde bulunan SPARQL uç noktalarının sürekli olarak taranması ve eş zamanlı olarak analizinin yapılabilmesi sağlanabilmektedir.



Şekil 4.4. SPECAN v2.0: SpEnD projesi masaüstü yazılımı analiz penceresi

5. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Tüm tarama ve çıkarım faaliyetlerinin sonunda, toplamda 100 binin üzerinde benzersiz URL arama motorları yardımıyla tespit edilmiştir. Şekil 5.1, bulunan SPARQL uç noktalarının toplam sayısına (tekil ve toplam) göre taranan URL'lerin sayısını göstermektedir. Şekil 5.1'de özetlenen tarama sırasında sistemin çalıştırılmasından itibaren yaklaşık 44 bin URL toplandıktan sonra keşfedilen benzersiz son nokta sayısının artmadığı görülmektedir. Yine Şekil 5.1'de 18-19 bin aralığında, keşfedilen benzersiz uç nokta sayısında belirgin bir artış görülmektedir. Bunun nedeni, arama metinlerinin ilk sorgulamasından sonra başlayan alan adı öğrenme görevinin bu aşamada devreye girmiş olmasıdır. Bu deney sırasında toplamda 1.037 benzersiz SPARQL uç noktası keşfedilmiştir. Keşif sürecinden sonra, bu 1.037 benzersiz uç nokta, kullanılabilirliği ve meta bilgileri dikkate alınarak analiz edilmiştir. Bunlardan 211'inin analiz aşamasında erişilemediği tespit edilmiştir (arama motoru sonuçlarında listeleniyor, ancak erişilemiyor). Kalan bitiş noktalarının 168'i, bir defadan fazla sonuç listesinde yer aldıklarından dolayı daha ileri analiz sonrasında listeden çıkarılmıştır. Sonuç olarak, toplam 658 adet uç nokta içeren bir çevrimiçi uç nokta listesi oluşturulmuştur. Çizelge 5.1'de arama terimlerinin kaçar adet uç nokta tespit ettiğinin listesi bulunmaktadır. Bu çizelgede, örneğin tüm arama motorlarında "sparql query" ifadesinin aranarak, 207 benzersiz SPARQL uç nokta tespit edilebildiği görülebilmektedir. Birden fazla arama terimi tarafından tekrar tekrar keşfedilen bazı uç noktalar da bulunmaktadır. Örneğin, "sparql -w3" ve "sparql -wiki" arama terimleri, neredeyse aynı SPARQL bitiş noktası URL'lerinin çıkarımı ile sonuçlanmıştır.



Şekil 5.1. Uç nokta sayısı ile taranan URL sayılarının karşılaştırılması

Çizelge 5.1. Arama sorgularına göre kaydedilen uç nokta sayısı

Arama Sorgusu	**Toplam Uç Nokta Sayısı**
"sparql endpoint site:org	30
"sparql endpoint"	179
"Virtuoso SPARQL Query Editor"	65
allintext: sparql query	37
allintitle: sparql query	45
allinurl: sparql data	23
sparql -language	55
sparql query	16
"sparql endpoint" site:co.uk	6
"sparql endpoint" site:com	10
"sparql endpoint" site:edu	11
"sparql endpoint" site:gov	1
"sparql endpoint" site:org	19
allintitle: sparql data	14
intitle:sparql	15
inurl:PoolParty inurl:sparql	6
inurl:PoolParty sparql	6
inurl:sparql	17
sparql	1

5.1. Tespit Edilen SPARQL Uç Noktalarının Mevcut Listeler ile Karşılaştırması

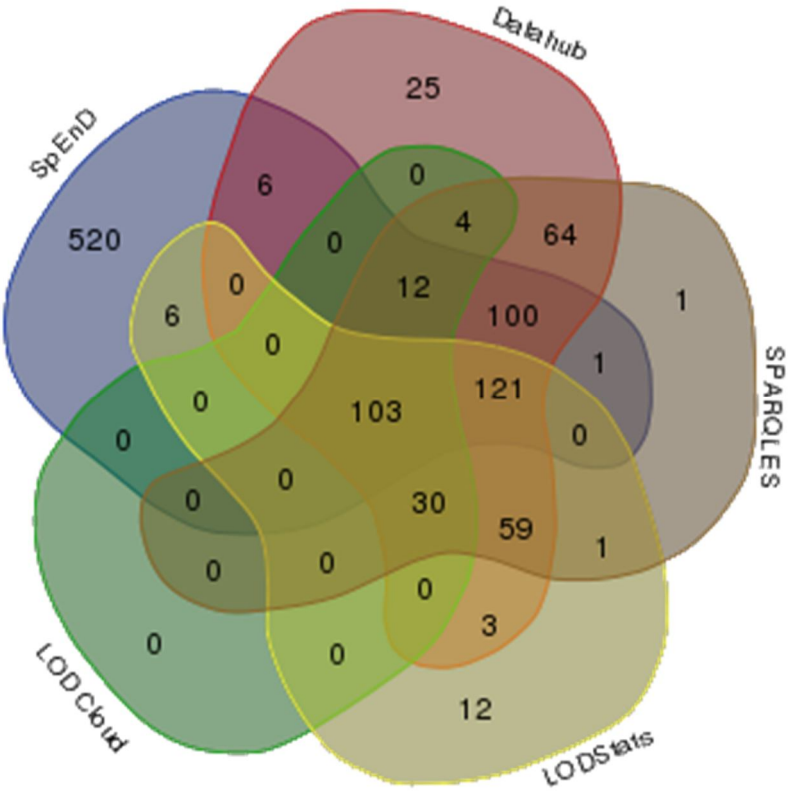
Bu bölümde, SpEnD'nin keşif sürecinden elde edilen sonuçlar, diğer dört büyük uç nokta deposundan (LOD Cloud, LODStats, SPARQLES ve Datahub) elde edilen

veriler ile karşılaştırılmıştır. Bu dört depoda ve SpEnD'de listelenen her uç noktanın durumunu belirlemek için, listelenen her bir SPARQL uç noktaya tez sürecinde basit SPARQL sorguları gönderilmiştir. Çizelge 5.2, bu depolarda ve SpEnD veri kümesindeki çevrimiçi ve çevrimdışı uç noktalarının sayısını listelemektedir. Depolarda listelenen SPARQL uç noktalarının hemen hemen yarısı çevrimdışı (pasif), bu SPARQL son nokta koleksiyonunun bu depolarda sık sık güncellenmediğinin bir göstergesidir. Dahası, SpEnD veri kümesinde de bulunan 211 çevrimdışı son nokta tespit edilmiştir. Şekil 5.2, tüm projeler tarafından bulunan farklı ve ortak uç noktanın toplam sayısını göstermektedir. SpEnD veri kümesinde 520 adet diğerlerinden farklı bitiş noktası vardır. Bununla birlikte, SpEnD tarafından keşfedilemeyen bazı SPARQL uç noktalar da bulunmaktadır, çünkü bunlar arama motorları tarafından keşfedilemeyecek şekilde yayınlanmıştır.

Çizelge 5.2. Keşfedilen SPARQL uç noktalarının erişilebilirlik karşılaştırması

Veri Kümesi	**Çevrimiçi Erişilebilir**			**Çevrim Dışı**	**Toplam**
	Yüksek	**Düşük**	**Toplam**		
Datahub	210	63	273	254	527
LOD Cloud	69	13	82	67	149
LodStats	136	39	175	160	335
SPARQLES	205	61	266	230	496
SpEnD	**537**	**121**	**658**	**211**	**869**

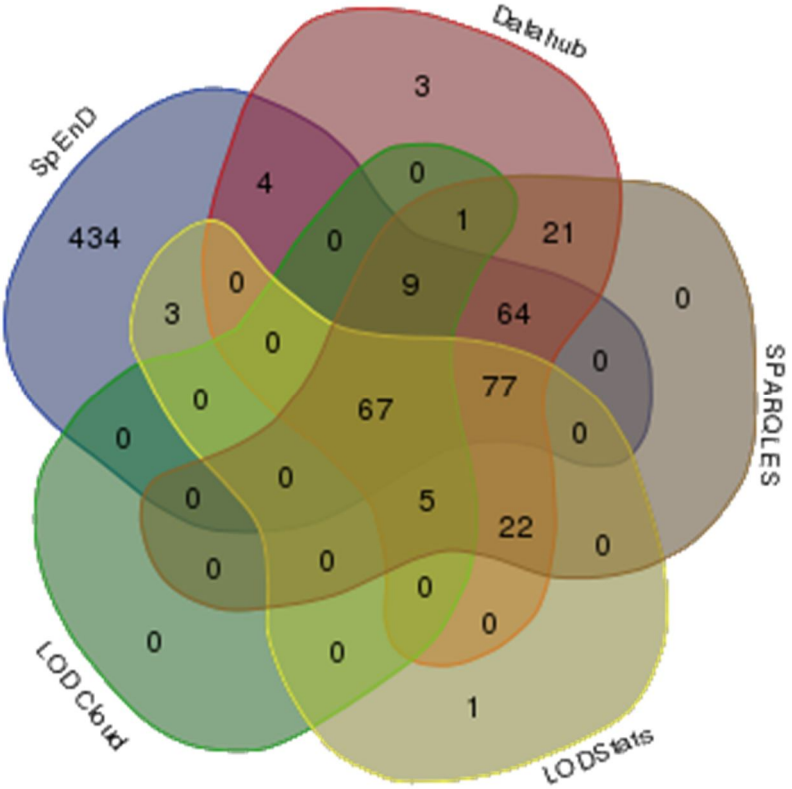
Şekil 5.3, sadece dört depoda ve SpEnD veri kümesinde listelenen aktif ve mevcut uç noktaları göstermektedir. Şekil 5.3'de gösterildiği gibi, diğer koleksiyonlarda listelenen 277 aktif bitiş noktasından 224'ü SpEnD tarafından bulunmuştur (%80,9 Doğruluk). SPARQLES ve LOD Cloud'un ayırt edici URL'leri görülmemekle birlikte; yalnızca üç farklı URL Datahub'da ve LODStats'da listelenen yalnızca bir ayırt edici URL görülmüştür. Datahub deposunun çoğunlukla SpEnD dışındaki etkin URL'ler açısından diğer listeleri kapsamakta olduğu görülmüştür. Şekil 5.2'de görüldüğü gibi, SpEnD veri kümesi toplamda (434 bitiş noktası) önemli miktarda benzersiz URL içermektedir ki bu URL'lerin bazıları aynı alan adı içerisinde bulunmaktadır. Bu nedenle, SPARQL uç nokta URL'lerinin PLD'leri (Pay Level Domain: Satın Alınan Alan Adı) temel alınarak tekrar analiz edilmiştir. Şekil 5.4 aktif PLD'lerin sayısını ve Çizelge EK- 1.2 SPARQL PLD'leri ve bunlara bağlı kaç adet uç nokta bulunduğunu listelemektedir. Alan adı bazında değerlendirildiğinde, SpEnD, diğer koleksiyonlarda listelenen 130 alan adının 119'unu keşfetmiştir (% 91 doğruluk oranı).



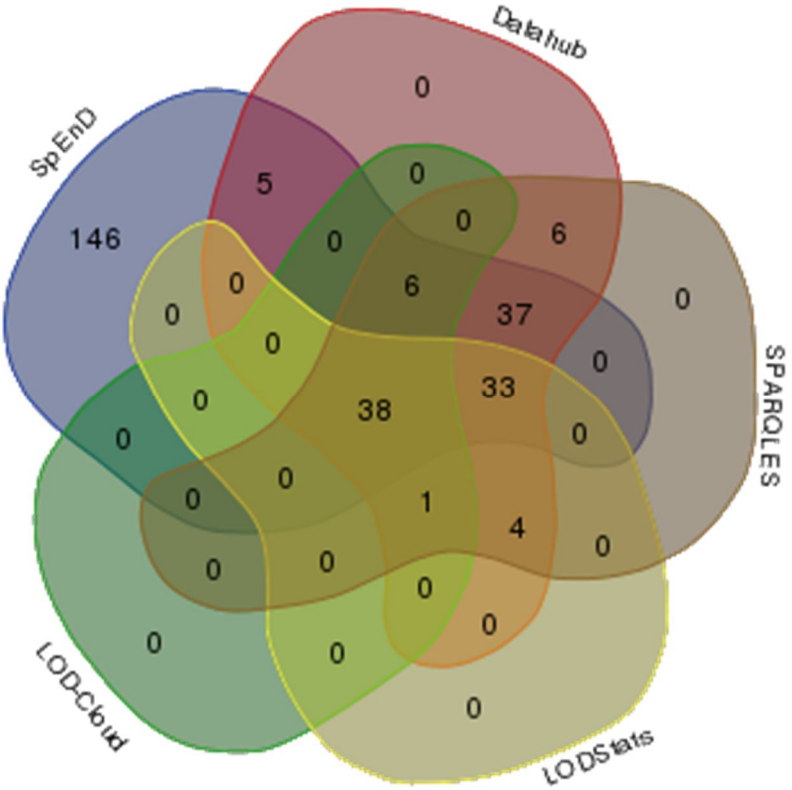
Şekil 5.2. Tüm projelerde bulunan toplam SPARQL uç noktası sayıları

SpEnD projesi tarafından keşfedilen 434 benzersiz uç nokta URL'leri ve 146 alan adı, kalite standartları (Mhleisen ve Bizer, 2012; Acosta ve ark., 2013; Kontokostas ve Westphal, 2014) bakımından değerlendirilmemiştir. Ancak başka hiçbir listede bulunmamaları açısından yeni bulunan uç noktalar araştırılmaya açık bir değer içermektedir.

Şekil 5.3'de, SpEnD diğer dört uç nokta listesi ile karşılaştırılmış, çevrimiçi ve çevrimdışı PLD'lerin ve bitiş noktası URL'lerinin toplam sayısı listelenmiştir. SpEnD veri kümesi en yüksek PLD sayısına (265) ve en fazla uç nokta URL'sine (658) sahiptir.



Şekil 5.3. Tüm projelerde bulunan erişilebilir SPARQL uç noktası sayıları

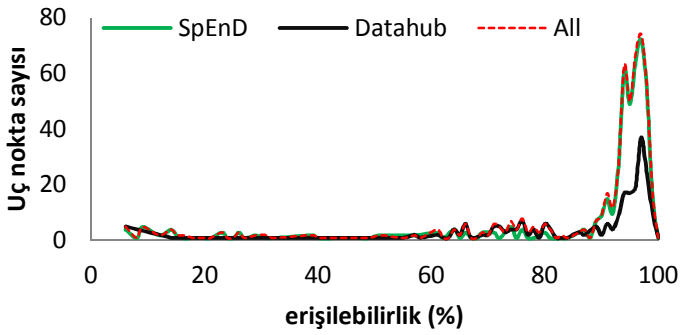


Şekil 5.4. Tüm projelerde bulunan erişilebilir SPARQL uç noktalarına ait tekil alan adlarının sayısı

Çizelge 5.3. Keşfedilen SPARQL uç noktalarının diğer veri kaynakları ile karşılaştırması

Proje adı	Alan Adı Sayısı		Uç Nokta Sayısı	
	Çevrimiçi	Çevrimdışı	Çevrimiçi	Çevrimdışı
SpEnD	265	59	658	211
SPARQLES	125	110	266	230
LODStats	76	78	175	160
LOD Cloud	45	34	82	67
Datahub	130	119	273	254

Haziran 2016'dan Eylül 2016'ya kadar, uç noktaların ortalama kullanılabilirlik yüzdelerini kaydetmek için birden fazla durum izleme görevi gerçekleştirilmiştir. Bu uç noktaların erişilebilirlik yüzdeleri Şekil 5.5'de gösterilmektedir. SpEnD veri kümesinde Datahub veri kümesine göre yüksek erişilebilirliğe sahip uç noktaların genel yüzdeleri daha yüksek olmakla birlikte, geçici veya kalıcı olarak çevrimdışı hale gelmiş uç noktalar da bulunmaktadır.



Şekil 5.5. Erişilebilirlik aralıklarına göre uç nokta sayıları (Yumusak ve ark., 2017)

5.1.1. Karşılaştırmalı istatistiksel sonuçlar

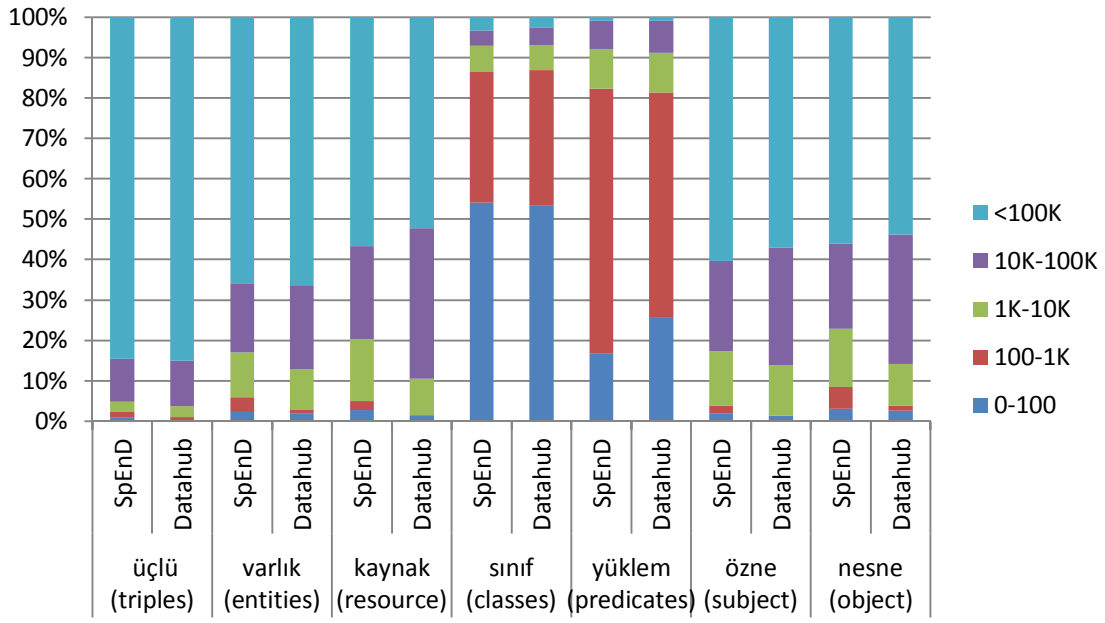
Çizelge 3.8'de listelenen istatistiksel sorgular, dört depoda listelenen bitiş noktalarının her birine ve SpEnD veri kümesindeki listeye uygulanmıştır. Sonuçlar, Çizelge 5.4'de listelenmiştir. SpEnD projesinin, diğer çevrimiçi SPARQL uç nokta kümelerinden daha fazla üçlü, varlık, kaynak, sınıf, öneri, konu ve nesne içerdiği görülmektedir. SpEnD tarafından listelenen 434 benzersiz SPARQL uç noktası (bkz. Şekil 5.2) bulunmaktadır ve 100 milyon'dan fazla üçlü içeren SPARQL bitiş noktası URL'leri

Çizelge EK- 1.3'de listelenmiştir.

Çizelge 5.4. Keşfedilen SPARQL uç noktalarının diğer veri kaynakları ile istatistiksel olarak karşılaştırması

Query ID	Datahub	LOD Cloud	LODStats	SPARQLES	SpEnD
1 (#triple)	60.643.885.274	3.973.511.913	12.525.555.224	58.700.752.022	**72.306.722.184**
2 (#entity)	552.288.747	143.577.593	242.227.863	549.407.079	**1.033.929.102**
3 (#resource)	106.799.760	13.304.311	24.135.529	44.108.276	**190.750.797**
4 (#class)	2.672.043	2.067.494	2.552.804	2.670.811	**4.865.883**
5 (#predicate)	448.478	70.644	152.403	446.420	**1.008.361**
6 (#subject)	94.873.508	14.509.470	33.566.833	91.516.952	**256.355.941**
7 (#object)	162.712.822	29.885.982	69.072.710	160.137.687	**308.017.891**

Şekil 5.6'da benzer sayıda özellik içeren bitiş noktalarının yüzdesi gruplandırılmış ve yığılmış bir sütun diyagramı olarak gösterilmiştir. Şekil 5.6'ya göre SpEnD uç nokta deposu, Datahub deposuyla karşılaştırıldığında tüm istatistiksel sonuçların benzer yüzdelerini içerir. Her varlık için, yüz bin'den fazla sonucu içeren uç noktaların yüzdesi, Datahub deposundan daha fazladır. Yani, SpEnD deposunun, üçlüler, kaynaklar, sınıflar, konular ve nesneler açısından Datahub'dan yüzde olarak daha yüksek hacimde uç noktalar içermektedir.



Şekil 5.6. İstatistiksel karşılaştırmalı analiz sonuçlarına göre uç nokta adetlerinin yüzdesel dağılımı (Yumusak ve ark., 2017)

5.2. Servis Özellikleri

Bir uç noktanın hizmet özelliklerini analiz etmek için RDF formatında veri talep

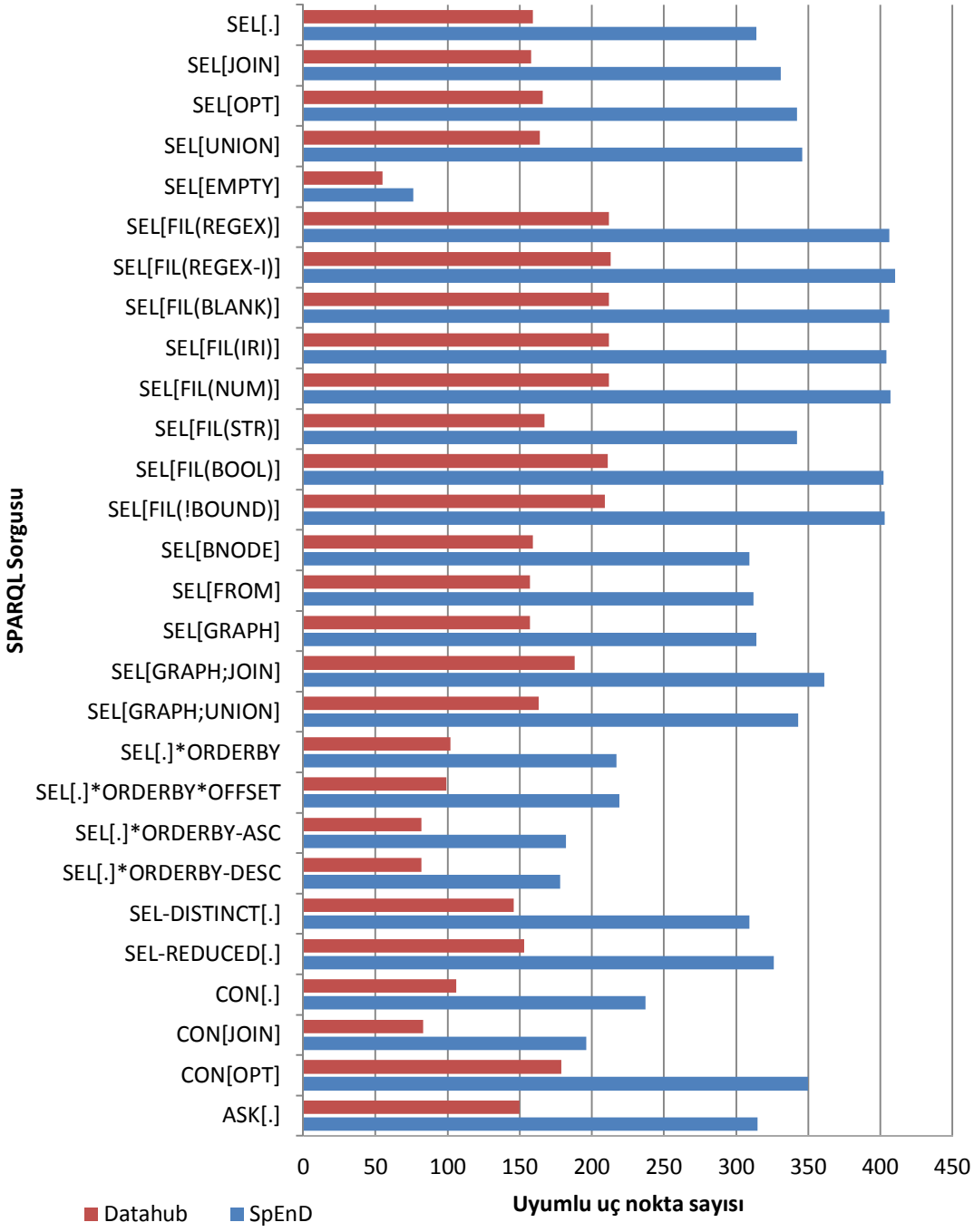
ederek SpEnD listesinde bulunan tüm uç noktalara basit bir GET isteği gönderilerek servis özellikleri tespit edilebilmektedir. Her bir sunucu için üst bilgisi yanıtı elde edilebilmektedir. Genel olarak, 527 bitiş noktası, sorgulamalara "200 OK" yanıtı vermiştir. Yanıtlar, Datahub ve SpEnD için bitiş noktası URL'leri ve PLD'ler açısından Çizelge 5.5'de sınıflandırılmıştır. Bu sonuçlarda, RDF verilerini geri gönderen uç noktaların çoğunda Apache, Virtuoso ve nginx sunucuları kullanılmaktadır.

Çizelge 5.5. Keşfedilen SPARQL uç noktası ve alan adı sunucu bilgileri karşılaştırması (Yumusak ve ark., 2017)

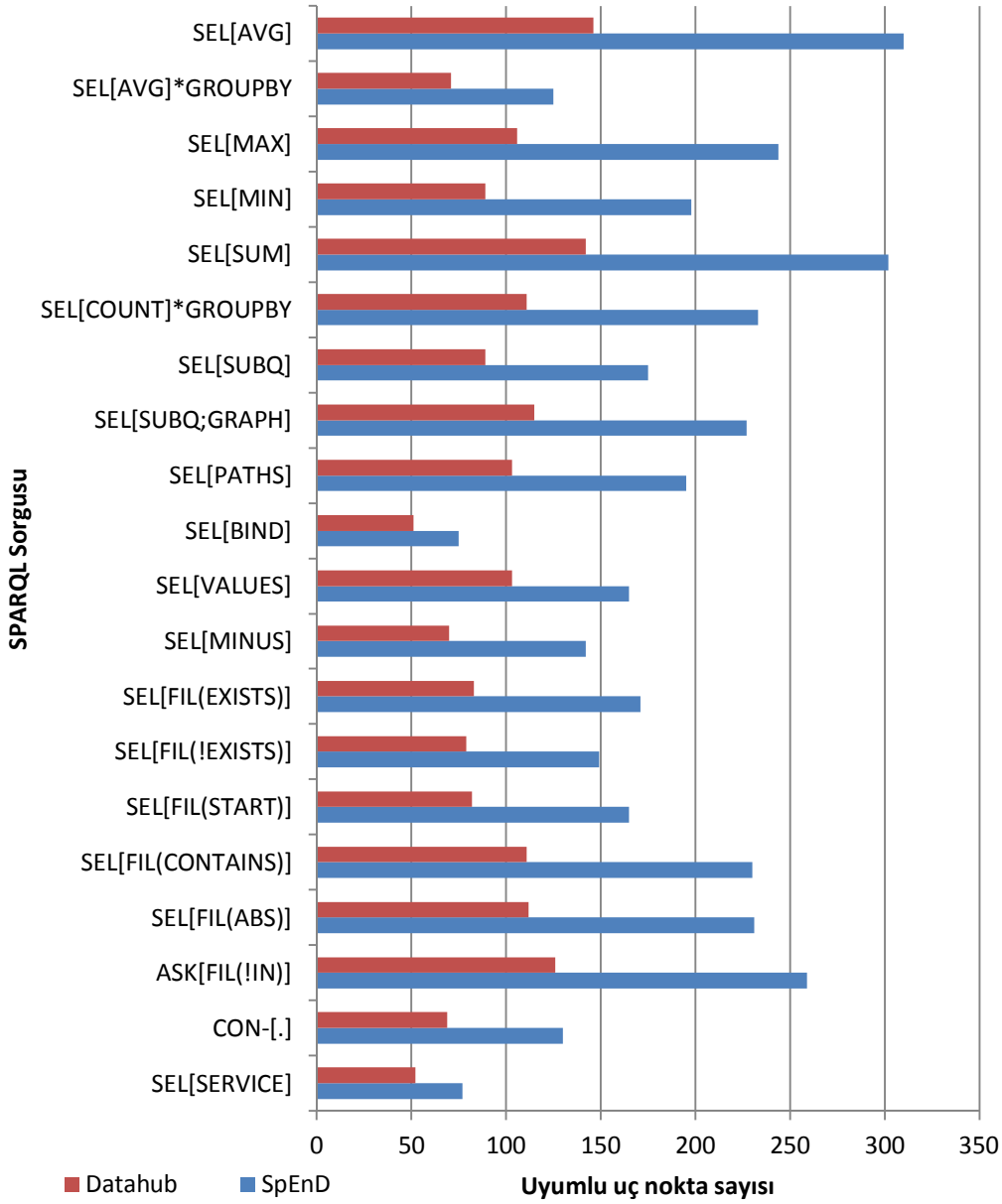
	Sunucu Tipi Sayısı			
	Datahub		SpEnD	
Sunucu Bilgisi	**Alan Adı**	**Uç Nokta**	**Alan Adı**	**Uç Nokta**
Apache	32	108	51	285
Virtuoso	39	71	63	119
nginx	18	16	31	58
Jetty	3	0	3	3
AllegroServe	0	0	2	2
Europa	0	0	1	2
HTTP::Server::PSGI	0	0	2	2
Oracle-Application-Server-10g	0	0	1	2
Fuseki	1	0	1	1
Sesame	1	0	1	1
INSEE	0	0	1	1
Koala Web Server	0	0	1	1
Microsoft-IIS	0	0	1	1
4s-httpd	1	1	0	0

5.3. İnteroperabilite (SPARQL 1.0 ve 1.1 desteği)

Datahub ve SpEnD'de listelenen tüm uç noktalar için, her SPARQL uç noktasına SPARQL 1.0 ve SPARQL 1.1 uyumluluklarının anlaşılabilmesi amacıyla test SPARQL sorguları gönderilmiştir. Test sorguları, standartları tanımlanan eski bir çalışmadan (C Buil-Aranda & Hogan, 2013) alınmıştır. Şekil 5.7 ve Şekil 5.8'de gösterildiği gibi, her SPARQL özelliği için uyumlu uç noktalarının yüzdesi açısından Datahub ve SpEnD karşılaştırması gerçekleştirilir. SpEnD, SPARQL 1.0 standardını destekleyen Datahub'ın iki katı oranında bitiş noktası barındırırken, her iki SPARQL uç nokta kümesi de SPARQL standartlarına uygunluk açısından benzerlik göstermektedir. Genel olarak, SPARQL 1.1 desteği SPARQL 1.0 desteğinden daha düşüktür. Şekil 5.8'de listelenen yüzde değerlerine dayanarak, her iki listedeki uç noktalar da benzer dağılım göstermektedir.



Şekil 5.7. SPARQL sorgu dili v1.0 uyumluluk sonuçları (Yumusak ve ark., 2017)



Şekil 5.8. SPARQL sorgu dili v1.1 uyumluluk sonuçları (Yumusak ve ark., 2017)

5.4. Performans Değerlendirmeleri

5.4.1. Sonuç akış (streaming) performansı

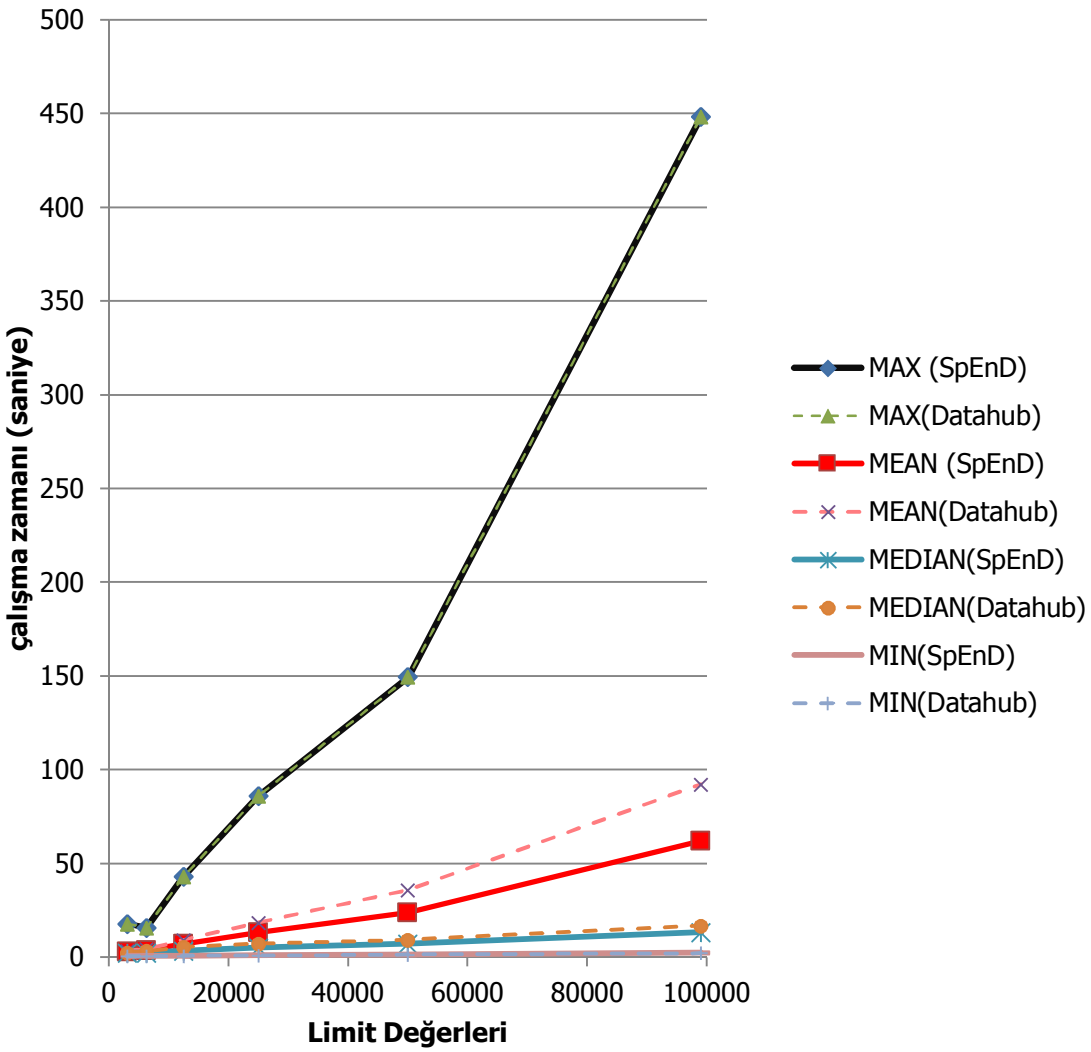
Uç noktalar için sonuç boyutu eşikleri, sonuç boyutunu 100.000'den fazla kayıtla sınırlandırarak bir SELECT sorgusu gönderme yoluyla incelenir. Uç noktalardan 398 boş olmayan sonuç dönüşü olmuştur. 199 uç nokta, 100.000'den fazla sonuç döndürmüş, bu da büyük olasılıkla sonuç büyüklüğü eşiğinin olmadığı anlamına gelmektedir. Uç noktalardan 126'sı (C Buil-Aranda & Hogan, 2013) 'da açıklandığı üzere belirtilen eşiklerde sonuçlar döndürmüştür. Bu geri dönüşler Çizelge 5.6'da; SpEnD, Datahub ve

birleşim uç nokta sayılarını verecek şekilde, sonuç büyüklüğü eşikleri bulunan uç noktaların sayısını karşılaştırmalı olarak listelemektedir.

Çizelge 5.6. Sonuç kısıtının sınırları (Yumusak ve ark., 2017)

	Uç Nokta Sayısı		
Sonuç sayısı (Limit)	**Birleşim Kümesi**	**Datahub**	**SpEnD**
500	5	1	5
1000	11	6	7
1500	1	1	1
10000	59	32	56
20000	3	1	3
50000	8	4	8
100000	39	20	38
Toplam	**126**	**65**	**118**

Belirtilen sınırların %99'undan fazlasını geri getiren uç noktalar için çalışma zamanı açısından çalışma süreleri Şekil 5.9'da gösterilmektedir. Maksimum, minimum ve medyan çalışma süreleri hem SpEnD hem de Datahub uç noktalarda aynı olmasına rağmen SpEnD uç noktaları için ortalama yürütme süresi, Datahub bitiş noktalarına göre daha düşüktür.

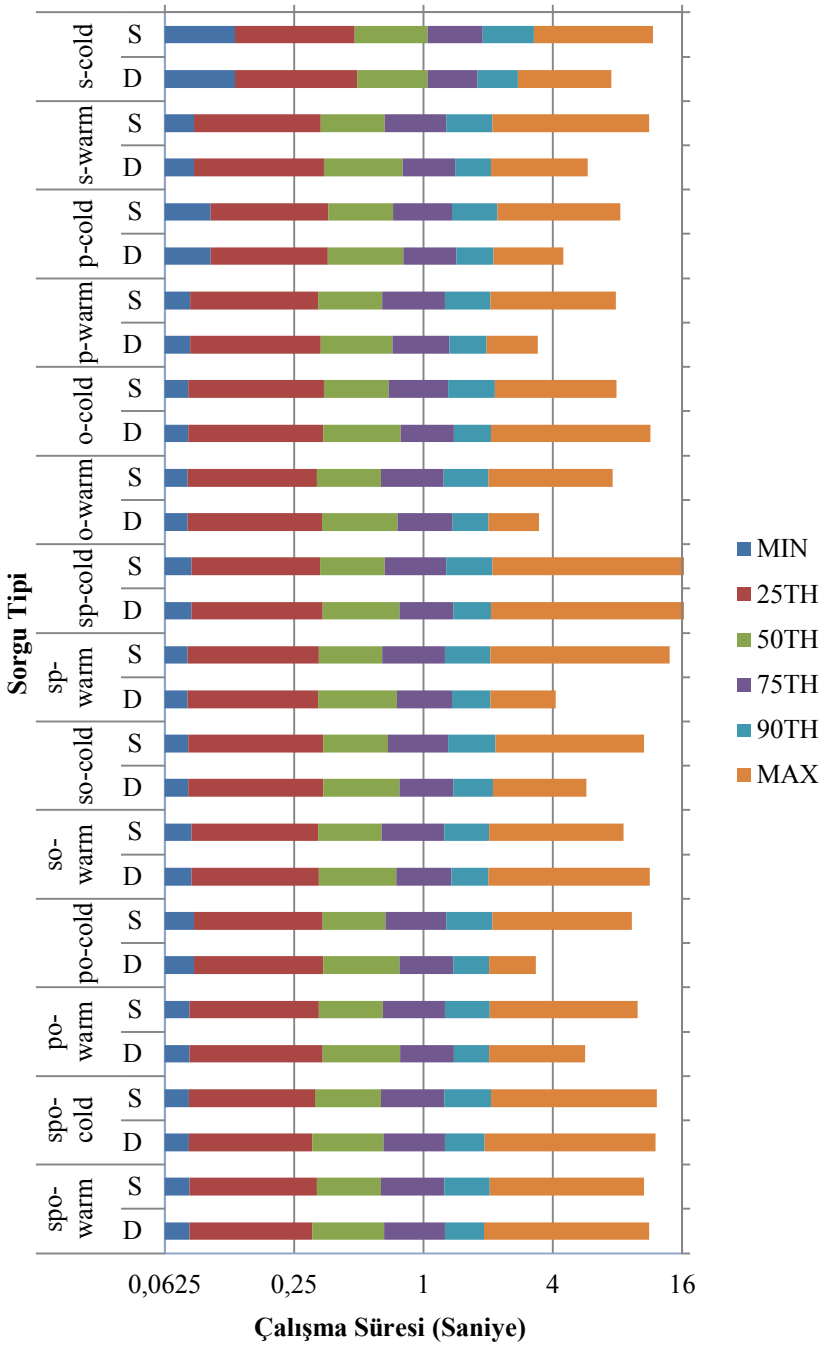


Şekil 5.9. Farklı limit büyüklüklerinin karşılaştırılması (Yumusak ve ark., 2017)

5.4.2. Atomik arama (lookup) ve katılma (join) performansı

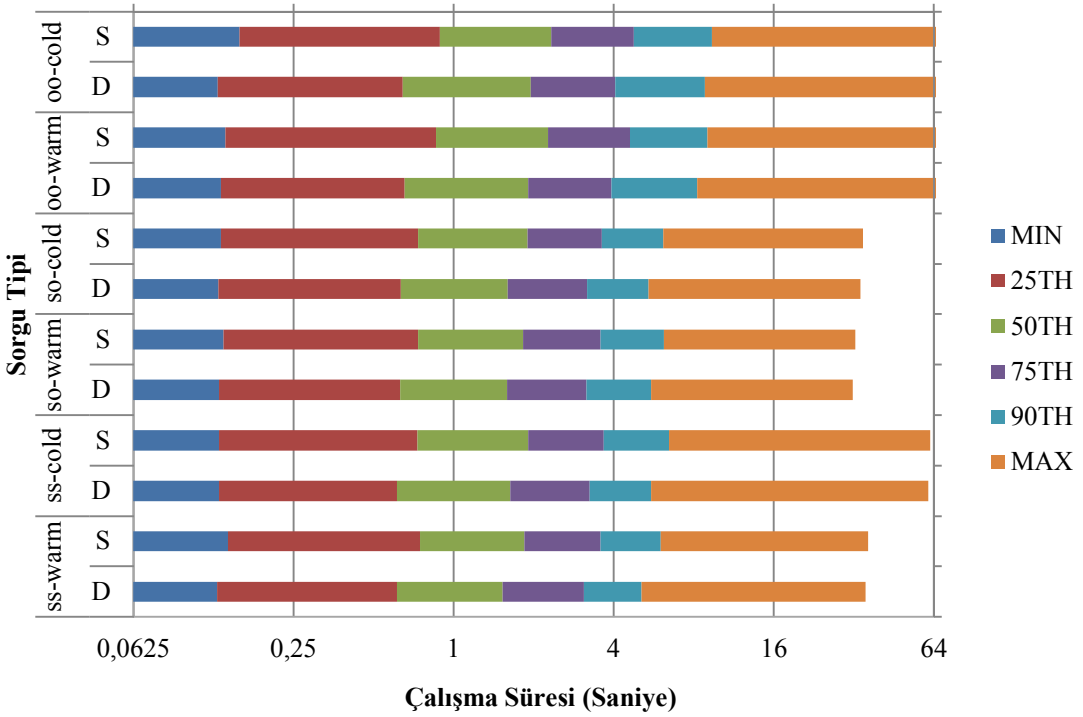
Bu bölümde, (C Buil-Aranda & Hogan, 2013) 'da belirtilen basit ASK sorguları için atomik seviye çalışma süreleri araştırılmıştır. Sorgular; konu, yüklem ve nesne özelliklerinin her bileşimini istemek üzere hazırlanmıştır. Örneğin, bir konu (s) sorgulanırsa, nesne ve yüklem koşulları <a> ve <b> olarak yazılır. Bu yolla, bir uç noktasının her bir üçlüsünü izlemesi gerekir veya ilgili endeksler geçer ve bu tür sorgular için maksimum çalışma süresini temsil eder. Sorgu sonuçlarındaki önbelleğe alma etkisini incelemek için; sorgu gönderimleri, ilk sorguyu *cold* (soğuk), ikinci sorguyu *warm* (ılık) olarak etiketleyerek her özellik için iki kez gerçekleştirilmekte bu sayede sunucuların önbellek özellikleri test edilmektedir. ASK sorguları, Datahub ve SpEnD'de ayrı listelenen tüm uç noktalara gönderilmiş ve sonuçlar Şekil 5.10'da

gösterilmiştir. ASK sorgularına yanıt veren SpEnD'den 373 uç nokta ve Datahub'dan 193 uç nokta test edilmiştir. Her bir çubuktaki renkli bölümler, verilen zaman periyodu (sn) içinde sonuçları döndüren uç noktalarının ilgili yüzdelerini temsil etmektedir. Uç noktaların yaklaşık %60'ı hemen hemen tüm sorgular için 1 saniye içinde yanıt vermiştir, maksimum yanıt süresi ise 16 saniye olarak ölçülmüştür.



Şekil 5.10. ASK sorguları için çalışma zamanı persentil değerleri (Yumusak ve ark., 2017)

Daha önce açıklanan ASK sorgularına benzer şekilde katılım (JOIN) performansları, (C Buil-Aranda & Hogan, 2013) 'da belirtildiği gibi özne-özne, özne-nesne ve nesne-nesne birleşimleri için birleştirme sorguları göndererek sınanmıştır. Sonuçlar, Şekil 5.11'de Datahub ve SpEnD için gösterilmiştir. Uç noktaların %40-%50'si, katılım sorguları için 1 saniye içinde yanıt dönüşü yapmıştır. SpEnD ve Datahub'daki veri kümeleri arasında anlamlı bir fark görülememiştir. So- ve ss- tipi sorgularının oo- tipi sorgulardan daha kısa sürede çalıştığı görülmüştür.



Şekil 5.11. JOIN sorguları için çalışma zamanı persentil değerleri (Yumusak ve ark., 2017)

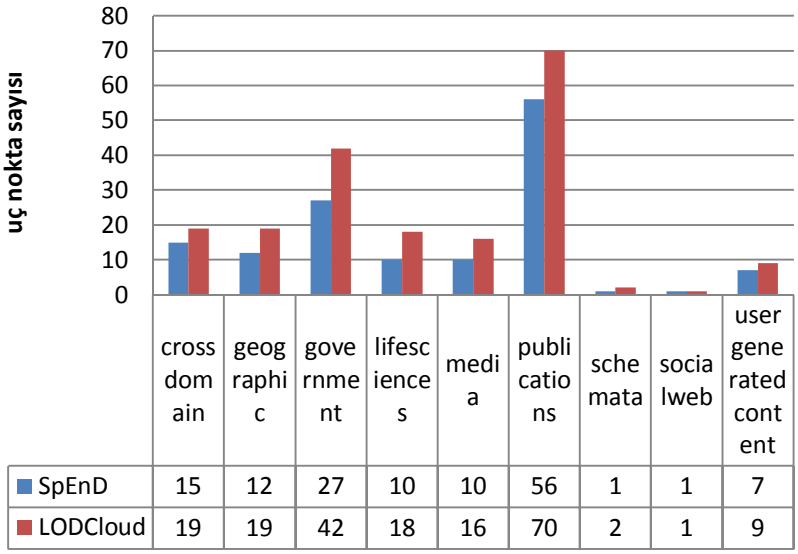
5.5. SPARQL Uç Noktalarının Değerlendirilmesi

Bu bölümde, uç noktaların içeriğini anlamak için iki deneysel analiz sonuçları verilmektedir. İlk olarak, mevcut uç noktaların kategorik analizi sonuçları; sonrasında IoT ontolojisi kullanım durumu hakkında kelime analizi sonuçları verilmektedir.

5.5.1. İçerik değerlendirme

LOD Bulut Diyagramı, bağlantılı açık veri kaynaklarının mevcut durumunu göstermektedir. Diyagramda, tüm veri kümeleri dokuz farklı kategoride Datahub'daki yayıncılar tarafından elle etiketlenerek renklendirilmiştir. SpEnD'de bulunan uç noktaları bu etiketli LOD veri kümesiyle karşılaştırılmıştır. LOD Cloud'da geçerli bir

SPARQL uç noktasına sahip 196 veri kümesi bulunmaktadır ve bunlardan 51'i kontrol edildiğinde erişilemediği tespit edildiğinden kapsam dışı bırakılmıştır. Şekil 5.12, SpEnD ve LOD'da bulunan uç noktalar için kategorilerin dağılımını göstermektedir.



Şekil 5.12. Kategorilere göre keşfedilen uç noktaları

5.5.2. Sözlük ve ontoloji değerlendirmeleri

Anlamsal ağ cümlecikleri veya ontolojileri "Belirli bir uygulamada kullanılabilen terimlerin sınıflandırılması, muhtemel ilişkilerin karakterize edilmesi ve bu terimlerin kullanılmasına ilişkin olası kısıtlamaların tanımlanması için kullanılmaktadır." (Gómez-Pérez & Corcho, 2002). Bağlantılı veri dünyasında Dublin Core, FOAF, SKOS, vb. gibi yaygın olarak kullanılan ve alana özgü bazı ontolojiler bulunmaktadır. Bu sözlüklerin bir veri kümesindeki kullanımı söz konusu alanlarda veri olduğuna işaret etmektedir. Bağlantılı verilerin yaratıldığı uygulama alanlarından birisi Internet of Things (IoT) teknolojileridir ve IoT alanı için zaten yaygın olarak kullanılan pek çok sözlük bulunmaktadır. Bu sözlük ve ontolojilerden bazıları (Gyrard, Atemezing, Bonnet, Boudaoud, & Serrano, 2016) tarafından incelenmiştir. Bu çalışmadan 37 IoT ile ilgili sözlük belirlenmiş, varlıklarını ve dolayısıyla SpEnD ve Datahub'da bulunan bağlantılı veri kümelerindeki kullanımları kontrol edilmiştir. Bu kontrollerde elde edilen sonuçlar IoT alanında anlamsal ağ kaynaklarını tespit etmek amacıyla gerçekleştirilen başka bir ortak çalışmada kullanılmıştır (Kamilaris ve ark., 2016).

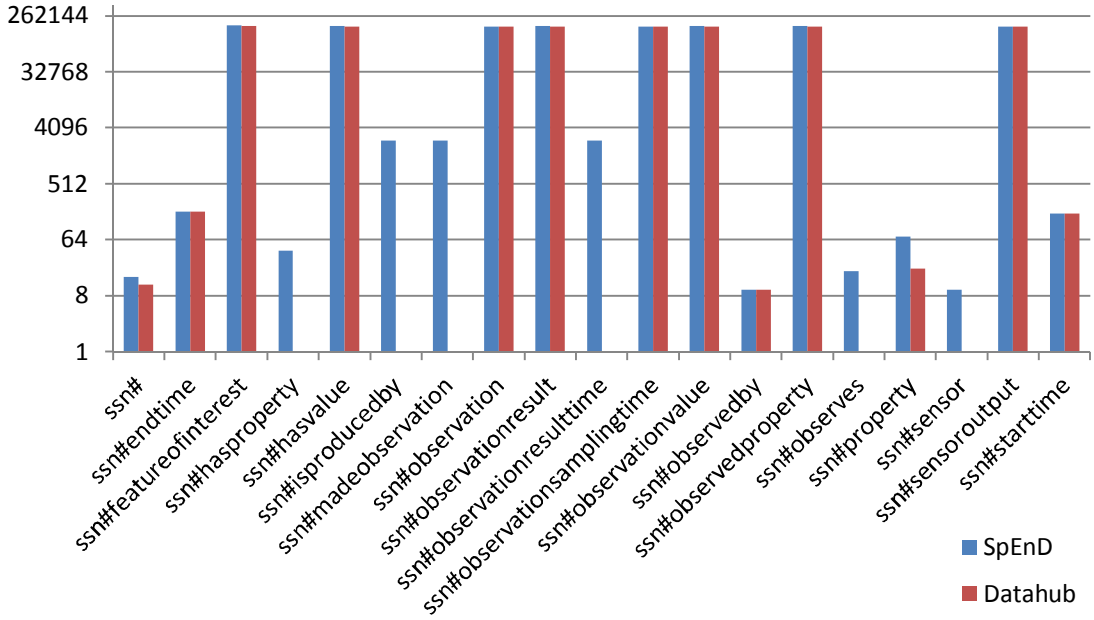
Sözlüklerin 9'unun kullanıldığını ve bunları kullanan bağlı veri kümelerinin

sayısı Çizelge 5.7'de listelenmiştir. Bu sözlükleri kullanan Datahub'dan 10 tane daha veri kümesi olduğu tespit edilmiştir. Bağlantılı veri kümelerindeki en çok kullanılan ontoloji olan SSN (Semantik Sensör Ağları) ontolojisi daha detaylı olarak incelenmiş ve bağlı özellikleri (ssn: hasvalue, vb.) karşılaştırarak veri kümelerindeki SNN ontolojisinin kullanımı kontrol edilmiştir.

Çizelge 5.7. IoT alanında geliştirilmiş tespit edilen ontolojilerin sayısı

IoT Ontoloji	Birleşim Kümesi	SpEnD	Datahub
SSN	7	7	3
DUL	4	4	1
SmartBuilding	2	2	1
Km4City	2	2	1
DogOnt	2	2	1
OpenIot	1	1	1
Fiemser	1	1	1
Fanfpai	1	1	1
Saref	1	1	0
Total	21	21	10

Şekil 5.13, uç noktalarda kullanılan SSN özelliklerinin sayısını (üçlülerin sayısını) göstermektedir. SpEnD tarafından keşfedilen veri kümelerinde, Datahub'dan daha fazla özellik kullanıldığı tespit edilmiştir. Örneğin, ssn: sensor, ssn: observes, ssn: producedBy gibi önemli özellikler Datahub'da bulunan uç noktalarda hiç kullanılmamıştır.



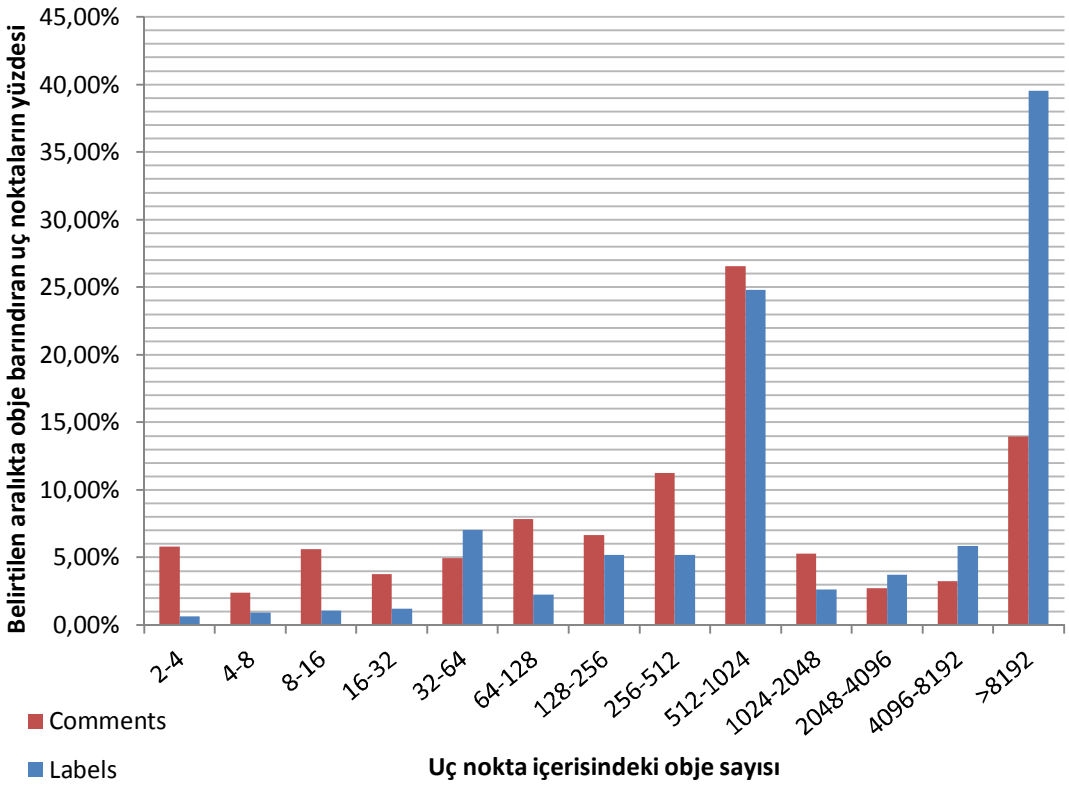
Şekil 5.13. SSN ontolojisi özelliklerinin sayısı

5.6. SPARQL Uç Noktalarının Sınıflandırılma Sonuçları

Toplanan tüm SPARQL uç noktaları erişilebilir olmamakla beraber ve yeterli veri de içermeyebilmektedir. Bu nedenle, çalışma için yararlı olmayan uç noktalar kapsam dışına çıkarılmıştır. Filtreleme işlemi ve elde edilen listeler aşağıdaki şekilde özetlenebilir.

- 1,328 SPARQL uç nokta, başlangıçta ilgili koleksiyonlardan toplanmıştır.
- 1,328 SPARQL uç noktasının 676'sı çevrim içi olarak erişilebilir ve *rdfs: comment* veya *rdfs: label* verisi içermektedir.
- 676 mevcut SPARQL bitiş noktasının 533'ü, en az 10 veya daha fazla *comment* veya *label* nesnesi içermektedir. 10'dan az içerenler kapsam dışı bırakılmıştır.
- 676 mevcut SPARQL uç noktasının 435'inde 1000'den fazla kelime içeren değerler tespit edilmiştir. Uzun içerikler 1000 kelime ile sınırlandırılmıştır.
- 676 SPARQL uç noktasından 77'si, 10.000'den fazla yorum nesnesi içermektedir. Bu nedenle, yalnızca ilk 10.000'i örnek alınmış ve gerisi yok sayılmıştır.
- Bu etiketlerden ve yorumlardan toplam 21.553.998 kelime çıkarılmıştır.

Kalan 533 bitiş noktasındaki etiket ve yorum kullanım dağılımı Şekil 5.14'de gösterilmektedir. Uç noktaların neredeyse yarısı 8.192'den fazla etiket içermekte ve bitiş noktalarının %15'i 8.192'den fazla yorum içermektedir. Ayrıca, bitiş noktalarının %25'inde 500-1.000 arasında "label" ve "comment" kaydı bulunduğu görülmektedir.



Şekil 5.14. Uç noktaların içerildikleri "label" ve "comment" sayılarına göre dağılımı (Yumusak ve ark., 2017)

5.6.1.Bağlantılı veri kaynakları için konu tavsiye yöntemi

Stf-Idf skorlama yönteminin SPARQL uç noktalarından elde edilen hipernim ve konu terimlerine uygulanması ile çeşitli terimlerin ön plana çıkması sağlanabilmektedir. İlgili uç noktalarla birlikte en yüksek skorlama hipernimi ve konu terimleri, Çizelge EK- 1.1'de listelenmiştir. Çizelge EK- 1.1'de listelendiği gibi, yüksek Stf-Idf skorlarına sahip hiperenim terimlerinin, muhtemelen bağlı veri kaynağının içeriğiyle ilişkili olduğu düşünülmektedir. Böylece, bir hipernim teriminin, uygun bir filtreleme kriteriyle bağlantılı veri kümesi için bir etiket olarak kullanılabilmesi öngörülmektedir. Yüksek Stf-Idf puanı olan konu terimleri, bu uç noktalar için LOD Bulutu'nda listelenen kategorilere benzer şekilde, bağlantılı veri kaynağının ana başlıklarıyla ilgili olma olasılığını da arttırmaktadır. Böylece, bir konu teriminin, bağlı bir veri kümesi için uygun bir filtreleme kriteriyle bir kategori adı olarak kullanılabileceği öngörülmektedir.

5.6.2. Bağlantılı veri kaynaklarının sınıflandırılması

Önerilen Tf-Idf skorlamasının sınıflandırma görevleri üzerindeki etkisini ölçmek için Stf-Idf ve Ctf-Idf skorlama fonksiyonları doküman vektörleri üzerinde (Bölüm

3.4'te açıklandığı gibi bağlantılı veri kaynakları ile oluşturulmuştur) uygulanmaktadır. Bu veri kaynakları daha sonra yedi farklı sınıflandırma tekniği kullanılarak sınıflandırılmış ve test edilmiştir. Bu sınıflandırma teknikleri Python v3.5.1[43] platformunda Scikit-learn v0.18[44] kütüphanesi kullanılarak uygulanmıştır, kaynak kodlar[45] açık kaynak olarak yayınlanmıştır. Kullanılan sınıflandırma yöntemlerinin nesnesini oluşturmak amacıyla Python dilinde geliştirilen yapıcı fonksiyonlar bu bölümde bulunan grafiklerde kullanıldıkları isimleriyle aşağıda listelenmiştir:

1. AdaBoost: AdaBoostClassifier()
2. Decision Tree: DecisionTreeClassifier(max_depth=5)
3. Linear SVM: SVC(kernel="linear", C=0.025, probability=True)
4. Naive Bayes: GaussianNB()
5. Nearest Neighbors: KNeighborsClassifier(3)
6. Random Forest: RandomForestClassifier (max_depth=5, n_estimators= 10, max_features=1)
7. RBF SVM: SVC(gamma=2, C=1)

Şekil 5.15'ten Şekil 5.18'e kadar olan tüm şekillerde, farklı sınıflandırıcıların ve puanlama yöntemlerinin maksimum / ortalama doğruluk sonuçları gösterilmektedir. Grafikleri daha anlaşılır kılmak için, özellikler ve puanlama yöntemleri alt çizgi (_) ile ayrılmış şekliyle aşağıdaki kısaltmalar kullanılarak belirtilmiştir:

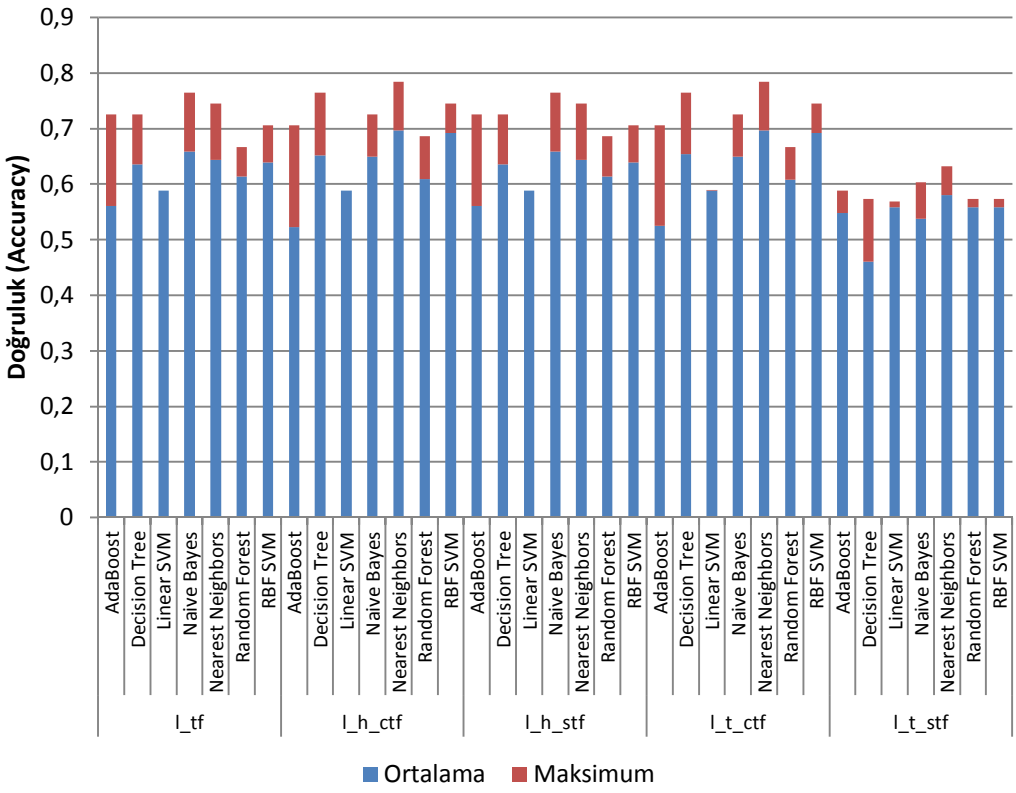
- c: rdf:comment
- l: rdf:label
- h: WordNet Hypernym
- t: Wordnet Topic
- tf: Tf-Idf score
- stf: Stf-Idf score
- ctf: Ctf-Idf score
- lvl: Wordnet Second Level terms

Şekil 5.15'de, yorumlarda farklı sınıflandırıcılar çalıştırmadan önce Ctf-Idf ve Stf-Idf skorlaması uygulanmaktadır. Doğruluk sonuçları hem Wordnet hipernimleri hem de Wordnet konuları için hesaplanmıştır. Bu şekile göre, anlamsal puanlama (Stf-Idf ve Cffidf), Tf-Idf skorlamasına kıyasla doğruluk sonuçlarını önemli ölçüde arttırmaktadır.

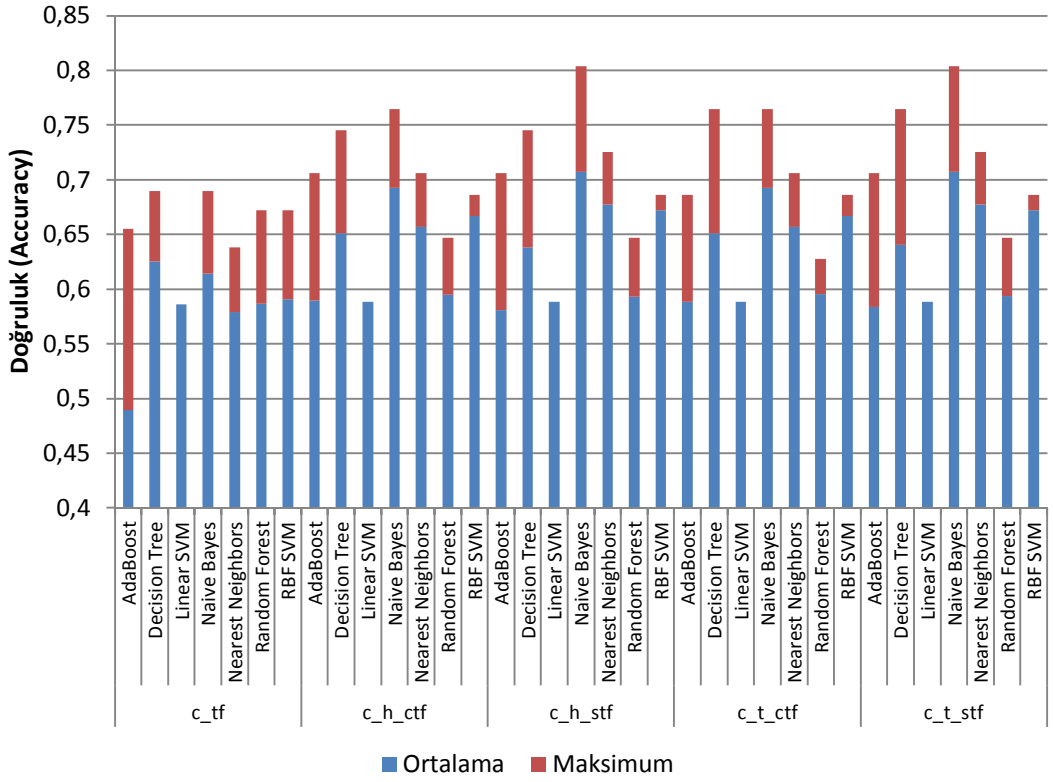
[43] https://www.python.org/downloads/release/python-351/
[44] http://scikit-learn.org
[45] https://github.com/semihyumusak/SparqlEndpointClassification

Hipernim parametresi ve Stf-Idf skorlaması (c_h_stf) ile doğruluk, Naive Bayes sınıflandırıcısı için %80'e kadar artabilir. Ayrıca, hipernim ve konu bazlı Stf-Idf skorlaması, tüm sınıflandırma algoritmaları için anlamlı bir fark göstermemektedir.

Şekil 5.15'te "label" özellikleri üzerinde, Şekil 5.16'da "comment" özellikleri üzerinde farklı sınıflandırıcılar çalıştırmadan önce Ctf-Idf ve Stf-Idf skorlaması uygulanmaktadır. Doğruluk sonuçları hem Wordnet hipernim hem de Wordnet konular için hesaplanmaktadır. Şekil 5.15'e göre, anlamsal puanlama, standart Tf-Idf skorlamasına kıyasla doğruluk sonuçlarını hafifçe artırmaktadır. Hipernim veya konu parametresi ve Ctf-Idf puanlaması (l_h_ctf, l_t_ctf) ile için doğruluk %78'e kadar artabilmektedir. Şekil 5.16'ya göre, anlamsal puanlama, Tf-Idf skorlamasına kıyasla doğruluk sonuçlarını artırdığı gözlemlenmiştir. Hipernim veya konu parametresi ve Stf-Idf puanlaması (c_h_stf, c_t_stf) ile için doğruluk %81'e kadar artabilmektedir.



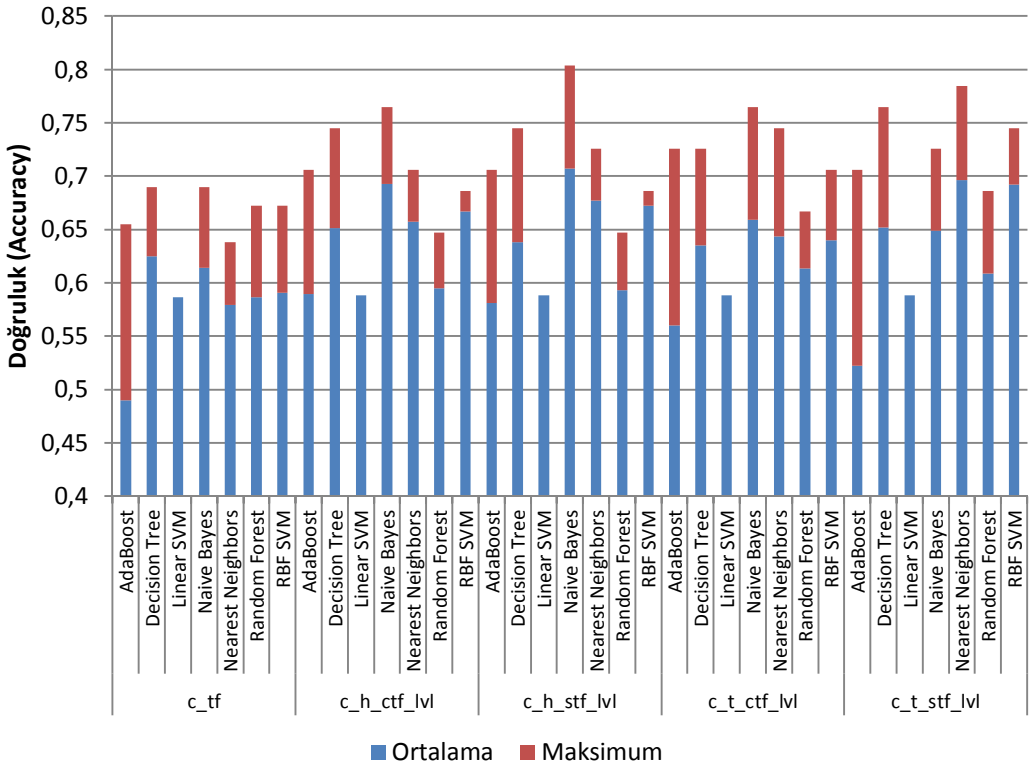
Şekil 5.15. Sınıflandırma yöntemleri ve skorlama yöntemlerinin "label" özelliğine göre doğruluk değerleri (Yumusak ve ark., 2017)



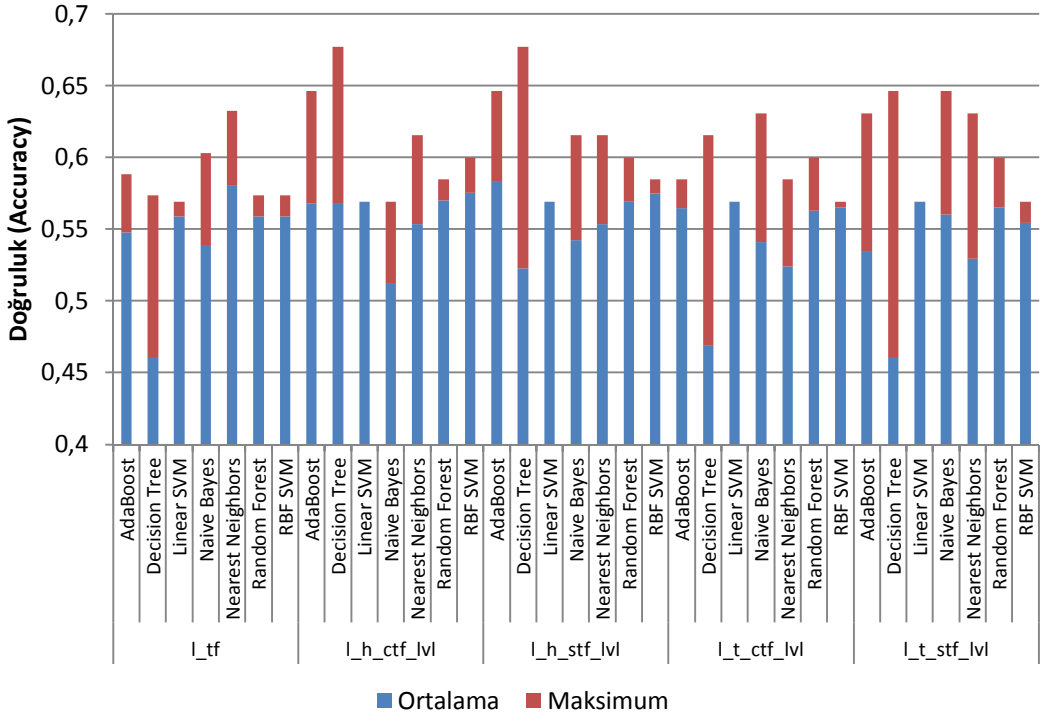
Şekil 5.16. Sınıflandırma yöntemleri ve skorlama yöntemlerinin "comment" özelliğine göre doğruluk değerleri (Yumusak ve ark., 2017)

Şekil 5.17'de, "comment" üzerinde farklı sınıflandırıcılar çalıştırılmadan önce Ctf-Idf ve Stf-Idf skorlaması uygulanmaktadır. Doğruluk sonuçları hem Wordnet ikinci düzey hipernimler hem de WordNet ikinci düzey konular için hesaplanmıştır. Bu şekile göre, Stf-Idf skorlaması Tf-Idf skorlamasına kıyasla doğruluk sonuçlarını artırmaktadır. Hipernim parametresi ve Stf-Idf skorlama (c_h_stf_lvl) ile doğruluk %80'e kadar artırılabildiği gözlemlenmiştir. Daha detaylı farklılık analizi ilerleyen bölümlerde gerçekleştirilmiştir.

Şekil 5.18'de, hem Ctf-Idf hem de Stf-Idf skorlaması, etiketler üzerinde farklı sınıflandırıcılar çalıştırmadan önce uygulanmıştır. Doğruluk sonuçları hem Wordnet ikinci düzey hipernimler hem de Wordnet ikinci düzey konular için hesaplanmıştır. Bu şekile göre, Stf-Idf skorlaması Tf-Idf skorlamasına kıyasla doğruluk sonuçlarını hafifçe arttırmıştır. Hipernim parametresi ve anlamsal skorlama (l_h_stf_lvl ve l_h_ctf_lvl) ile doğruluk değerinin %68'e kadar artırılabildiği gözlemlenmiştir. Daha detaylı farklılık analizi ilerleyen bölümlerde gerçekleştirilmiştir.



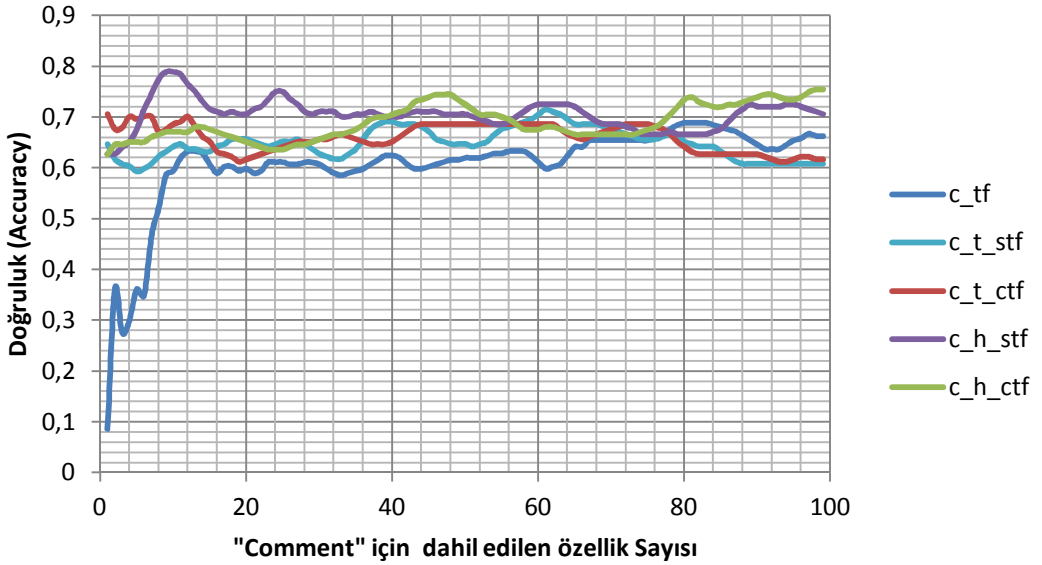
Şekil 5.17. Sınıflandırma yöntemleri ve skorlama yöntemlerine göre "comment" özelliklerinin ikinci seviye anlamsal ilişkilerinin doğruluk sonuçları (Yumusak ve ark., 2017)



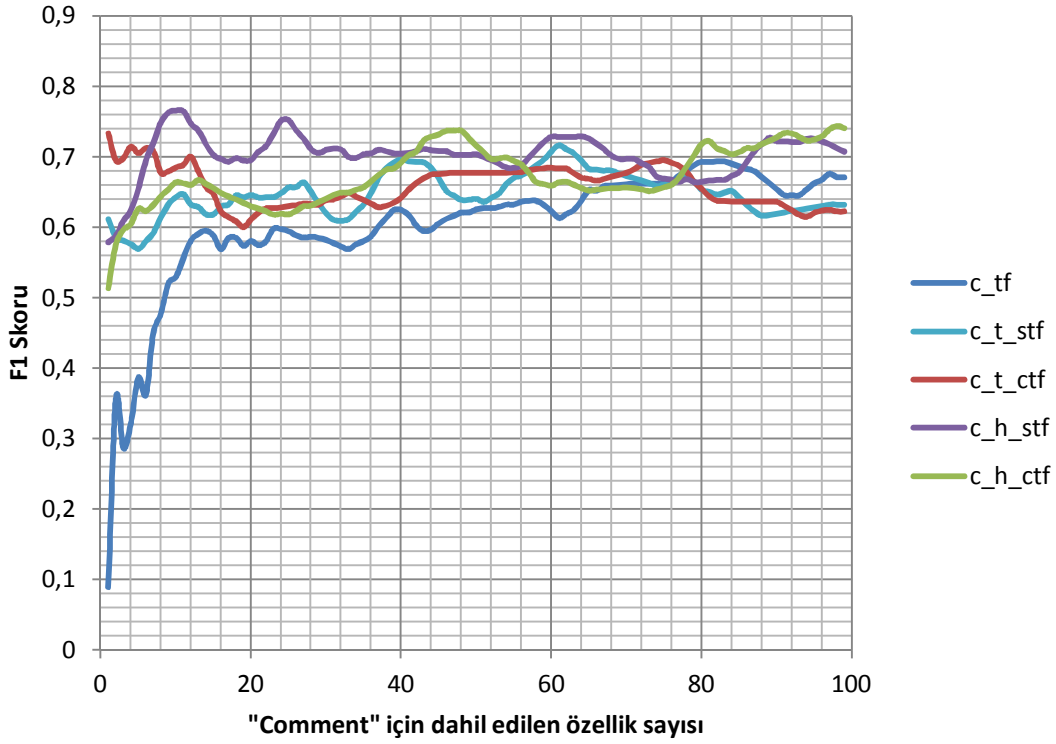
Şekil 5.18. Sınıflandırma yöntemleri ve skorlama yöntemlerine göre "label" özelliklerinin ikinci seviye anlamsal ilişkilerinin doğruluk sonuçları (Yumusak ve ark., 2017)

Şekil 5.15'den Şekil 5.18'e kadar olan şekillerde farklı girdilerin ve puanlama yöntemlerinin farklı sınıflandırıcılar üzerindeki etkisi özetlemektedir. Bu sonuçlara dayanarak, Naive Bayes sınıflandırıcısının denemelerin çoğunda iyi performans gösterdiği anlaşılmıştır. Stf-Idf ve Ctf-Idf yöntemleri ile skorlanan özellik sayısının etkisini incelemek amacıyla Naive Bayes sınıflandırıcısı için doğruluk ve F1 skorlarının ayrıntılı grafikleri oluşturulmuştur. Bu grafikler, Stf-Idf ve Ctf-Idf skorları ile skorlanan, skor değerlerine göre artımlı özellik ekleme yöntemi kullanılarak oluşturulmuştur. Bu deneylerin her birinde sınıflandırma sonuçları ilk 100 özellik için hesaplanmıştır.

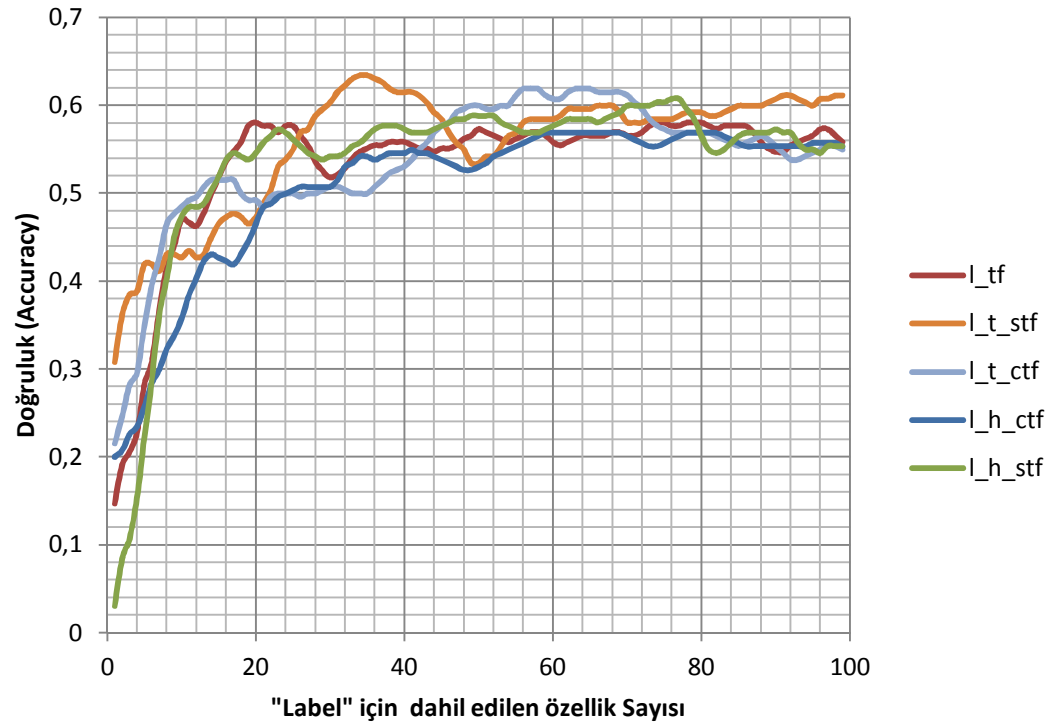
Şekil 5.19 ve Şekil 5.20'ye göre, yorumlardaki standart Tf-Idf skorlaması daha düşük doğruluk ve F1 skorlarıyla sonuçlanırken, hipernim ve konu geliştirilmiş semantik skorlama daha yüksek skorlar vermektedir. Öte yandan, etiketleri özellik olarak kullanarak sınıflandırma sonuçları skorlama yöntemleri arasında önemli bir değişiklik göstermemektedir. Etiket özellikleri altındaki metin içeriği genellikle yorum özelliklerinden daha kısadır. Etiket cümlelerinin eksikliği nedeniyle, etiketler sınıflandırma görevi için yeterli girdi değerini sağlayamamaktadır. Bununla birlikte, Şekil 5.21 ve Şekil 5.22'deki 20-40 özellik eklenmesi arasında Stf-Idf (l_t_stf) sınıflandırma sonuçlarının diğer skorlama yöntemlerinden belirgin şekilde yüksek olduğu gözlemlenmiştir. Şekil 5.26'ya kadar olan diğer tüm grafiklerde gerçekleştirilmiş olan tüm analiz sonuçlarının artırımsal özellik seçimi yöntemiyle Naive Bayes sınıflandırıcısı için doğruluk ve F1 skoru değerleri gösterilmektedir.



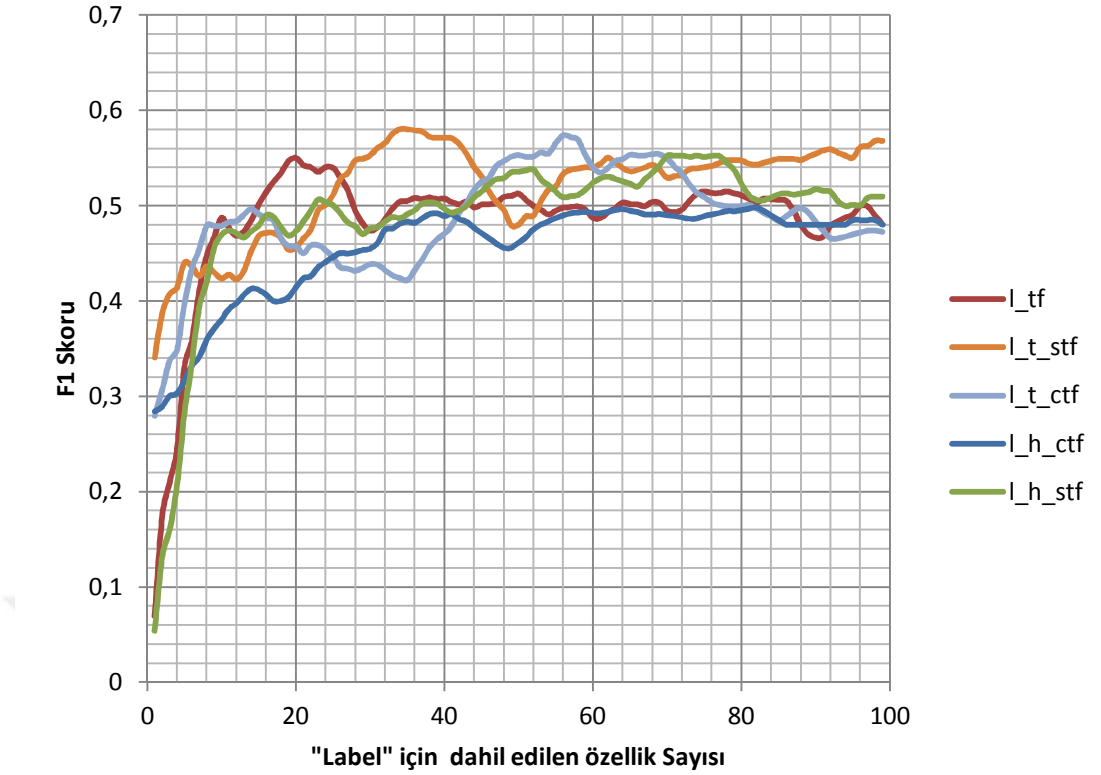
Şekil 5.19. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, "comment" için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017)



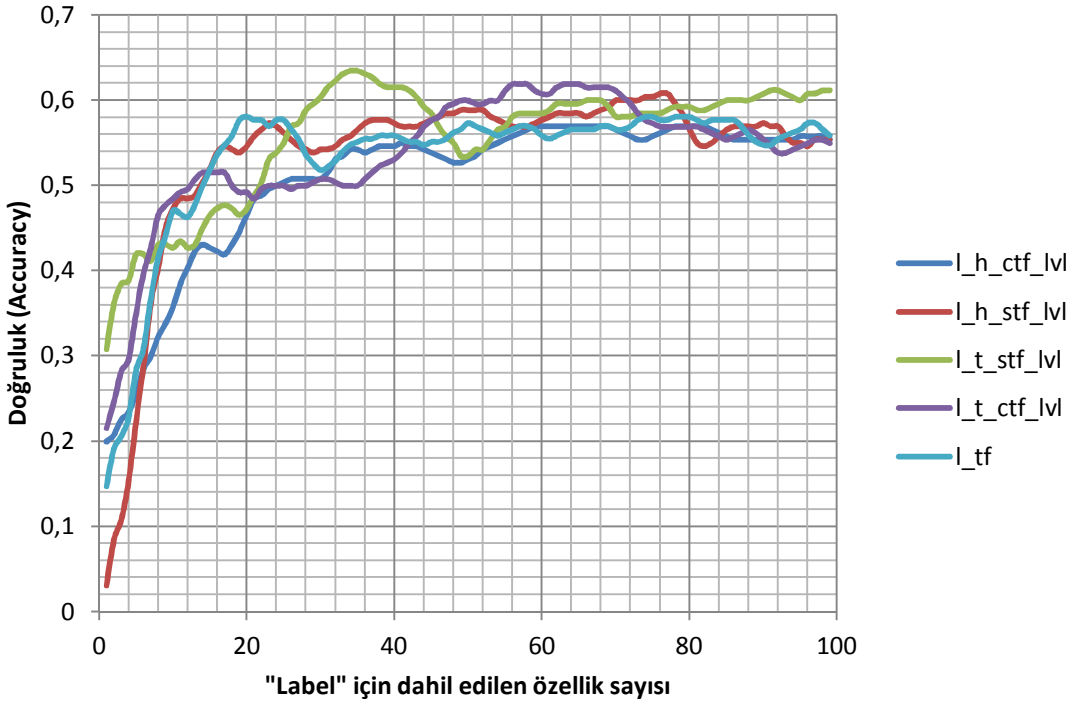
Şekil 5.20. Naive Bayes sınıflandırıcısına göre F1 değerlerinin, "comment" için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017)



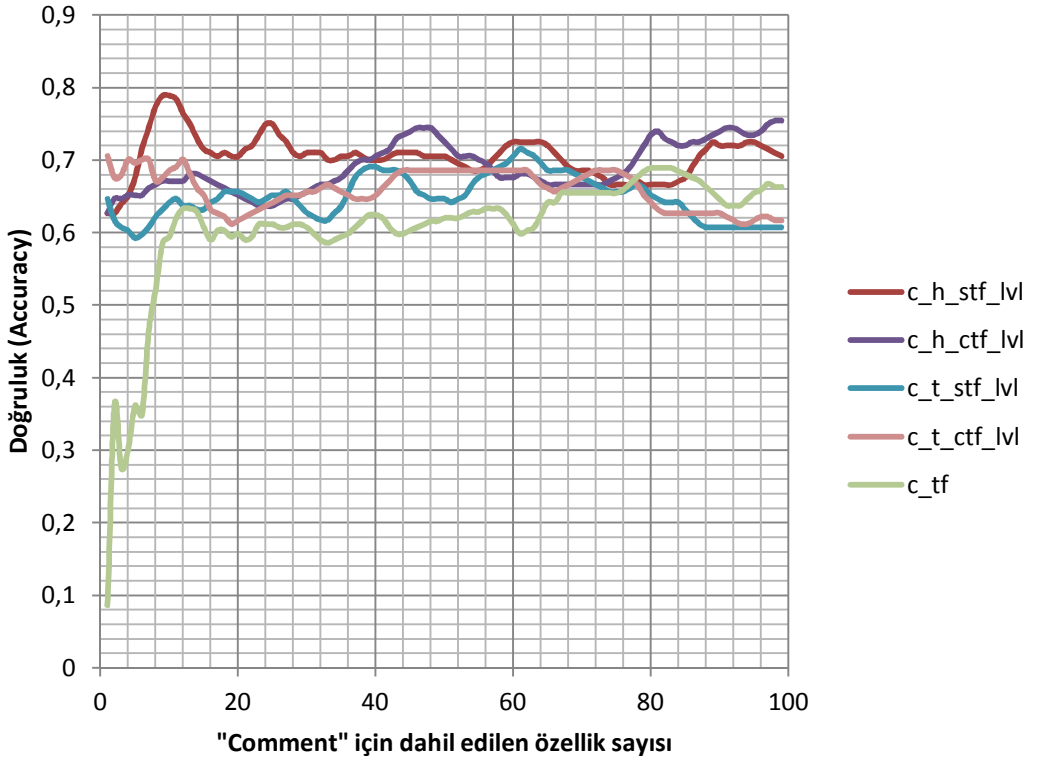
Şekil 5.21. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, "label" için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017)



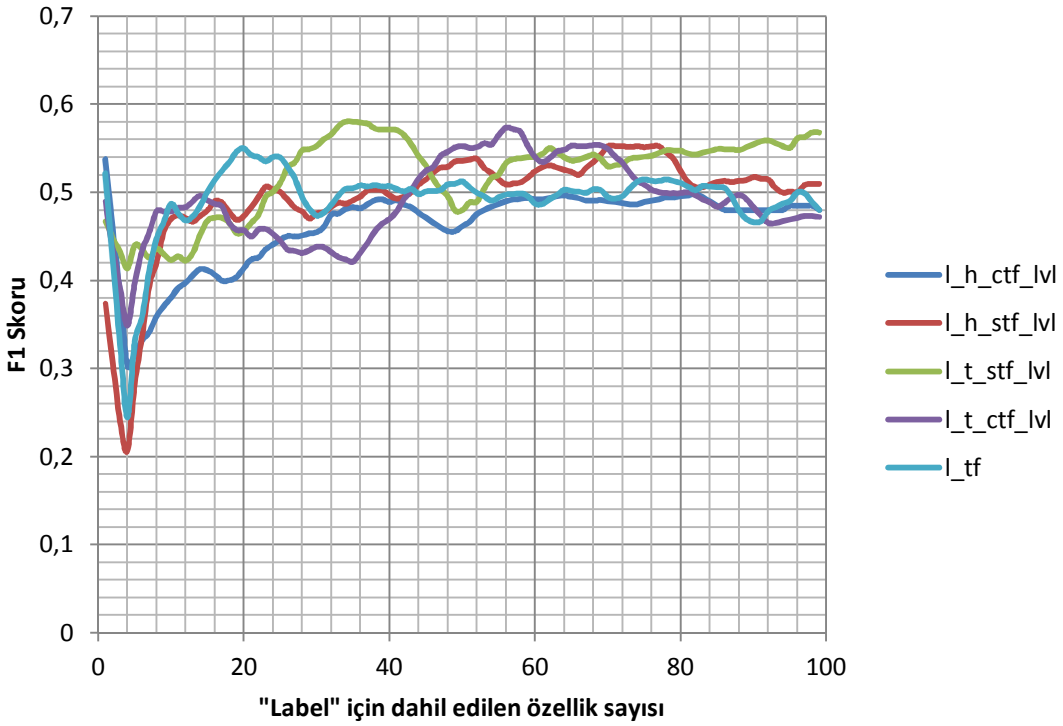
Şekil 5.22. Naive Bayes sınıflandırıcısına göre F1 değerlerinin, "label" için skorlanan özellik sayısının artışına göre değişimi (Yumusak ve ark., 2017)



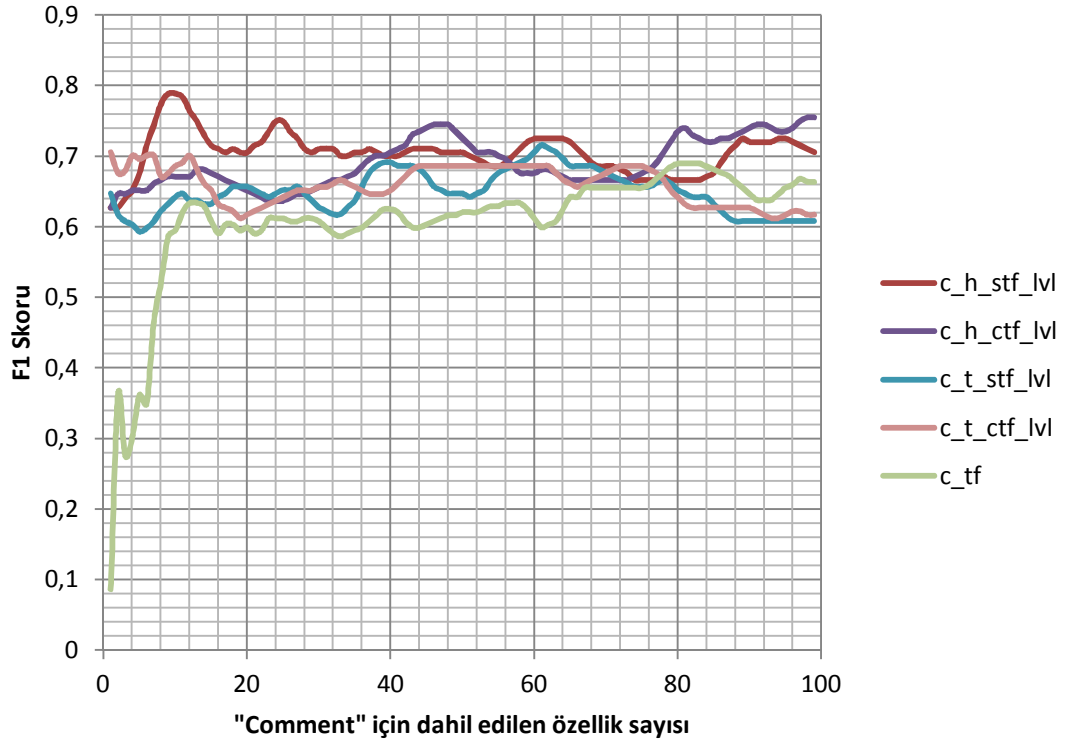
Şekil 5.23. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, semantik ikinci seviye "label" için skorlanan özellik sayısının artışına göre değişimi



Şekil 5.24. Naive Bayes sınıflandırıcısına göre doğruluk değerlerinin, semantik ikinci seviye "comment" için skorlanan özellik sayısının artışına göre değişimi



Şekil 5.25. Naive Bayes sınıflandırıcısına göre F1 skoru değerlerinin, semantik ikinci seviye "label" için skorlanan özellik sayısının artışına göre değişimi



Şekil 5.26. Naive Bayes sınıflandırıcısına göre F1 skoru değerlerinin, semantik ikinci seviye "comment" için skorlanan özellik sayısının artışına göre değişimi

5.6.3. Bağlantılı veri kaynaklarının sınıflandırılma sonuçlarının istatistiksel analizi

Bu bölümde, tahmin doğruluk değerlerinin farklı skorlama yöntemleri için istatistiksel anlamlılıkları analiz edilmektedir. Anlamlı farklılık analizi için Kruskal-Wallis H test (Kruskal ve Wallis, 1987) ve Mann-Whitney U testleri (Mann ve Whitney, 1947), tahmin doğruluk değerleri üzerine uygulanmıştır. Kruskal-Wallis H testi sıralama tabanlı parametrik olmayan bir test olarak açıklanmıştır ve bu test bağımsız değişkenin iki veya daha çok grupları için istatistiksel anlamlı farklılık olup olmadığını tespit etmek için kullanılır. Mann-Whitney U testi ise iki bağımsız grup arasında istatistiksel anlamlı farklılık olup olmadığını tespit etmek amacıyla karşılaştırma yapmaktadır. Bu iki test ilk önce ortalama doğruluk skorları üzerine uygulanmış, sonrasında da maksimum doğruluk skorlarına uygulanmıştır ve sonuçları bu bölümde listelenmiştir.

5.6.3.1. Ortalama tahmin doğruluğu değerlerinin analizi

Anlamlı farklılık analizi için Kruskal-Wallis H testi ortalama doğruluk değerlerine uygulanmıştır. Sonrasında, farklılığın kaynağının tespiti amacıyla Mann-

Whitney U test uygulanmıştır ve her bir yöntem için ikili karşılaştırma gerçekleştirilmiştir.

Bu bölümde gerçekleştirilen Kruskal Wallis H testi sonuç tablolarında bulunan çıktı kısaltmalarının açıklamaları aşağıdadır:

- N: doğruluk değerleri karşılaştırılan örneklem sayısı
- df (Degress of freedom) : özgürlük derecesi
- ki-kare (chi-squared statistics): ki-kare istatistiksel değeri
- p (Asymp. Sig.): istatistiksel anlamlılık değeri

Gerçekleştirilen Mann-Whitney U testi sonuç tablolarında bulunan çıktı kısaltmalarının açıklamaları aşağıdadır:

- N: doğruluk değerleri karşılaştırılan örneklem sayısı
- U: istatistiksel U değeri
- p (Asymp. Sig.): istatistiksel anlamlılık değeri

Çizelge 5.8'de, ortalama sıralama değerlerine göre, Ctf-Idf doğruluk sonuçları Stf-Idf ve Tf-Idf doğruluk değerlerinden fazla çıkmıştır.

Çizelge 5.8. Kruskal-Wallis H test: farklı skorlama tekniklerine göre ortalama doğruluk değerlerinin farklılıkları (Yumusak ve ark., 2017)

Yöntem	N	Sıra ortalaması	df	ki-kare	p
Tf-Idf	56	70,64285714	2	6,853863393	*0,032
Stf-Idf	56	90,58928571			
Ctf-Idf	56	92,26785714			
Toplam	168				

* İstatistiksel anlamlılık değeri p<0,05

Kruskal-Wallis H test sonucuna göre, bu skorlama teknikleri arasında doğruluk değerlerinde anlamlı farklılık tespit edilmiştir. (p<0,05). Anlamlı farklılığın kaynağını tespit etmek amacıyla, Mann-Whitney U testi uygulanmış ve sonuçlar Çizelge 5.9'da listelenmiştir.

Çizelge 5.9'da ikili karşılaştırma sonuçları listelenmektedir. Bu tablodaki sonuçlara göre, Stf-Idf skorlaması Tf-Idf skorlamasından anlamlı oranda yüksek doğruluk vermektedir (U=1232; p=0,05; p<0,05). Ctf-Idf ve Tf-Idf skorlama yöntemleri arasında Ctf-Idf lehine anlamlı farklılık tespit edilmiştir (U=1128; p=0,01; p<0,05).

Ancak, Stf-Idf ve Ctf-Idf skorlamaları arasında anlamlı bir farklılık tespit edilememiştir (U=1563; p=0,98; p>0,05).

Çizelge 5.9. Mann-Whitney U test: Farklı skorlama tekniklerinin ortalama doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017)

Yöntemler	N	Sıra ortalaması	Sıraların Toplamı	U	p
Tf-Idf	56	50,50	2828	1232	*0,05
Stf-Idf	56	62,50	3500		
Tf-Idf	56	48,64	2724	1128	*0,01
Ctf-Idf	56	64,36	3604		
Stf-Idf	56	56,59	3169	1563	0,98
Ctf-Idf	56	56,41	3159		

* İstatistiksel anlamlılık değeri p<0,05

Birinci seviye anlamsal terimler ile ikinci seviye anlamsal terimler arasındaki ortalama doğruluk anlamlı farklılığını analiz etmek için, Mann-Whitney U testi uygulanmıştır. Bağımsız değişkenin semantik terim seviyesi olduğu bu analiz sonuçları Çizelge 5.10'da listelenmiştir.

Çizelge 5.10'daki sonuçlara göre, birinci seviye semantik terimlerin sınıflandırması ile ikinci seviye semantik terimlerin sınıflandırması arasında anlamlı bir farklılık tespit edilmemiştir (U=1544; p=0,89; p>0,05).

Çizelge 5.10. Mann-Whitney U test: Farklı semantik seviyelerin ortalama doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017)

Semantik Seviye	N	Sıra ortalaması	Sıraların Toplamı	U	p
1. seviye	56	56,93	3188	1544	0,89
2. seviye	56	56,07	3140		

* İstatistiksel anlamlılık değeri p<0,05

5.6.3.2. Maksimum tahmin doğruluğu değerlerinin analizi

Bir önceki bölümde gerçekleştirilen analizlere benzer bir şekilde, Kruskal-Wallis H testi bu defa maksimum doğruluk değerlerine uygulanmıştır. Mann-Whitney U testi de aynı şekilde uygulanarak farklılığın kaynağı araştırılmıştır.

Çizelge 5.11'de, sıra ortalaması değerine göre, Stf-Idf doğruluk değerleri sırasıyla Ctf-Idf ve Tf-Idf değerlerinden yüksektir. Kruskal-Wallis testine göre, yöntemler arasında anlamlı farklılık mevcuttur. Anlamlı farklılığın kaynağını tespit

etmek amacıyla gerçekleştirilen Mann-Whitney U testinin sonuçları aşağıda listelenmiştir.

Çizelge 5.11. Kruskal-Wallis H test: farklı skorlama tekniklerine göre maksimum doğruluk değerlerinin farklılıkları (Yumusak ve ark., 2017)

Yöntem	N	Sıra ortalaması	df	ki-kare	p
Tf-Idf	56	70,5	2	7,885	*0,019
Stf-Idf	56	95,86607143			
Ctf-Idf	56	87,13392857			
Toplam	168				

* İstatistiksel anlamlılık değeri p<0,05

Çizelge 5.12'de, Stf-Idf skorlamasının Tf-Idf skorlamasından anlamlı olarak yüksek doğruluk sonuçları verdiği görülmektedir (U=1100; p=0,01; p<0,05). Ctf-Idf ve Tf-Idf skorlamaları arasında anlamlı farklılık tespit edilmemiştir. (U=1252; p=0,065; p>0,05). Benzer şekilde Stf-Idf ve Ctf-Idf skorlamaları arasında anlamlı farklılık tespit edilmemiştir (U=1399; p=0,326; p>0,05).

Çizelge 5.12. Mann-Whitney U test: Farklı skorlama tekniklerinin maksimum doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017)

Yöntemler	N	Sıra ortalaması	Sıraların Toplamı	U	p
Tf-Idf	56	48,1428571	2696	1100	*0,006
Stf-Idf	56	64,8571429	3632		
Tf-Idf	56	50,8571429	2848	1252	0,065
Ctf-Idf	56	62,1428571	3480		
Stf-Idf	56	59,5089286	3332,5	1399	0,326
Ctf-Idf	56	53,4910714	2995,5		

* İstatistiksel anlamlılık değeri p<0,05

Maksimum doğruluk değerlerine göre birinci ve ikinci seviye semantik terimlerin kullanımının anlamlı bir farklılık oluşturup oluşturmadığını tespit etme amaçlı gerçekleştirilen Mann-Whitney U testinin sonuçları Çizelge 5.13'de gösterilmektedir.

Çizelge 5.13. Mann-Whitney U test: Farklı semantik seviyelerin maksimum doğruluk değerlerin farklılıklarına göre ikili karşılaştırılması (Yumusak ve ark., 2017)

Semantik Seviye	N	Sıra ortalaması	Sıraların Toplamı	U	p
1. seviye	56	56,57	3168	1564	0,98
2. seviye	56	56,43	3160		

* İstatistiksel anlamlılık değeri p<0,05

Çizelge 5.13'deki sonuçlara göre, birinci ve ikinci seviye semantik terimler arasında anlamlı bir farklılık bulunmamaktadır. ($U=1564$; $p=0,98$; $p>0,05$).

Özetle, yapılan istatistiksel karşılaştırmalar neticesinde Stf-Idf ve Ctf-Idf skorlama yöntemlerinin standart Tf-Idf skorlamasına göre sınıflandırma algoritmalarına tabi tutulan özellik vektörlerine uygulandığında daha iyi doğruluk verdiği tespit edilmiştir. Ancak Stf-Idf ve Ctf-Idf yöntemleri arasında istatistiksel bir farklılık gözlenememiş olup, Ctf-Idf skorlamasına göre daha az hesaplama gerektiren Stf-Idf skorlamasının tek başına uygulanmasının yeterli olduğu tespit edilmiştir. Bu anlamsal skorlama yöntemlerini uygularken kullanılan Wordnet anlamsal bağlantılarının birinci veya ikinci seviye olmasına dayalı yapılan analizlerde ise, kelimenin birinci veya ikinci anlamsal bağlantılarının kullanılmasının herhangi bir istatistiksel farklılık oluşturmadığı tespit edilmiştir.

6. SONUÇLAR VE ÖNERİLER

Bağlantılı veri kullanımının yaygınlaştırılması ile hedeflenen temel amaç çevrim içi veri akışlarının düzenlenmesi ve veri yığınlarının anlamlı hale getirilmesidir. Bunun sonucunda kolay sorgulanabilir ve erişilebilir veri kaynakları ile insanların olduğu kadar bilgisayarların da anlamlı veri alışverişi yapabilmesi amaçlanmaktadır. İnternet sunucularında saklanan veriler anlamlı ve standart bir biçime uyarlandığı zaman, dağıtık veri kaynağı haline dönüşmekte ve farklı yerlerden anlamsal sorgular ile sorgulanabilmektedir. Bu veri kaynaklarının sorgulanması amacıyla geliştirilmiş olan SPARQL sorgulama dili temelinde geliştirdiğimiz bu tez çalışmasında, anlamsal ağların temel veri biçimi olan RDF veri tabanlarının sorgulanabilmesini kolaylaştıracak ve bu veriye ihtiyaç duyanlara yol gösterecek bir araç geliştirilmiştir.

Bahsedilen SPARQL sorgulamaları, temelinde çevrim içi bağlantılı veri kaynaklarının uzak sorgulamalara olanak vermesi amacıyla SPARQL uç noktaları ismi verilen bağlantı noktaları üzerinden HTTP veya SOAP benzeri protokolleri ile gerçekleştirilmektedir. Çevrim içi bağlantılı veri kaynaklarının sorgulanmasına olanak veren bu uç noktaların, internet üzerinde dağınık şekilde web sayfaları olarak bulunmakta olduğu ve kullanıcılar tarafından kolaylıkla tespit edilemediği bu tez çalışmasında gösterilmektedir. Bu veri kaynaklarını ve bağlı olan uç noktaları, kullanıcılar tarafından kolay erişilebilir kılmak amacıyla listeleyen çeşitli çalışmaların varlığı tespit edilmiş ancak bu çalışmaların da oldukça yetersiz oldukları içerik analizleriyle ortaya çıkarılmıştır. Bu yetersizliği giderebilmek amacıyla, SPARQL uç noktalarını otomatik olarak tespit eden, sürekli gözlem ve analizlerini gerçekleştirebilen bir meta-arama ve analiz aracı geliştirilmiştir. Devamında, tespit edilen SPARQL uç noktaları, kullanıcılar tarafından kullanılabilmesini sağlamak amacıyla sınıflandırma, konu önerme, etiketleme gibi prosedürlere tabi tutulmuştur. Bu tez çalışmasında açıklanan yöntem ve uygulamanın, SPARQL uç noktalarının tespit edilmesi aşamasından, içerik analizi yapılarak kullanıcılara sunulabilmesi aşamasına kadar olan tüm süreçler anlatılmıştır.

Tespit edilen tüm uç noktalar ve bunlara bağlı çıkan sonuçların, mevcut listeleme çalışmalarıyla karşılaştırmalı analizleri gerçekleştirilmiştir. Bu analizler neticesinde bağlantılı veri koleksiyonları çoğunlukla Datahub gibi merkezi depolarda saklansalar da bu yaklaşımların yeni çevrimiçi SPARQL uç noktalarını keşfetmek, izlemek ve belirli bir süre sonra çevrimdışı olanları belirlemek için yeterince etkili ve

dinamik olmadığı açıkça gösterilmiştir. Bu eksiklikleri giderebilmek amacıyla, tez çalışması sürecinde geliştirilen SPECAN tarama ve analiz yazılımı tüm yönleriyle açıklamıştır. Bu tez çalışmasında açıklanan yöntemler kullanılarak, SPARQL uç noktalarının sürekli keşfi, analizi ve tanımlanması mümkün olabilmektedir. Tez çalışması sürecinde geliştirilen tüm yazılımlar ve keşfedilen tüm SPARQL uç noktası listeleri, SpEnD adıyla tanıtılan proje sayfalarında ve açık kaynak kütüphanelerinde[46] kullanıma sunulmuştur.

Bu tez çalışmasında geliştirilen yazılım ve yöntemlere ek olarak iki konuda geliştirilme yapılması önerilmektedir. Birincisi, SPARQL uç noktalarının haricinde sorgulanabilir bağlantılı verilerin de bu yöntemlere dahil edilmesinin daha geniş kapsamlı bir veri erişimi sağlayacağı öngörülmektedir. Bu kapsamda geliştirilecek olan yöntemler, tez çalışmasında elde edilen yöntemler ile birlikte kullanılarak bağlantılı veri kullanımının artırılmasını sağlayacaktır. İkincisi, tespit edilen bağlantılı veri kaynaklarının anlamsal analizleri bu kaynakların kullanılabilirliği için önem arzetmektedir. Bu tez çalışmasıyla ön çalışmaları ve sonuçları açıklanmış olan anlamsal veri analizinin daha derin kapsamlı olarak incelenmesinin bağlantılı veri kullanımının kolaylaşmasını ve artmasını sağlayacağı öngörülmektedir.

[46] https://github.com/semihyumusak/SpEnD

KAYNAKLAR

Acosta, M., Vidal, M. E., Lampo, T., Castillo, J. ve Ruckhaus, E., 2011, ANAPSID: An adaptive query processing engine for SPARQL endpoints, 7031 LNCS (PART 1), 18-34.

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S. ve Lehmann, J., 2013, Crowdsourcing Linked Data Quality Assessment, *International Semantic Web Conference*, 8219, 260-276.

Akar, Z. ve Hala, 2012, Querying the Web of Interlinked Datasets using VOID Descriptions., *Ldow*.

Alani, H., Brewster, C. ve Shadbolt, N., 2006, Ranking ontologies with AKTiveRank, 1-15.

Aleman-Meza, B., Halaschek-Weiner, C., Arpinar, I. B., Ramakrishnan, C. ve Sheth, A. P., 2005, Ranking complex relationships on the semantic web, *IEEE Internet Computing*, 9 (3), 37-44.

Alexander, K. ve Hausenblas, M., 2009, Describing linked datasets-on the design and usage of void, the'vocabulary of interlinked datasets.

Alexander, K., Cyganiak, R., Hausenblas, M. ve Zhao, J., 2011, Describing Linked Datasets with the VoID Vocabulary.

Bai, X., Delbru, R. ve Tummarello, G., 2008, RDF snippets for Semantic Web search engines, 1304-1318.

Balmin, A., Hristidis, V. ve Papakonstantinou, Y., 2004, Objectrank: Authority-based keyword search in databases, 564-575.

Balog, K., Serdyukov, P. ve Vries, A. P., 2010, Overview of the TREC 2010 entity track.

Batzios, A., Dimou, C., Symeonidis, A. L. ve Mitkas, P. A., 2008, BioCrawler: An intelligent crawler for the semantic web, *Expert Systems With Applications*, 35 (1-2), 524-530.

Batzios, A. ve Mitkas, P. a., 2012, WebOWL: A Semantic Web search engine development experiment, *Expert Systems With Applications*, 39 (5), 5052-5060.

Berners-lee, B. T. ve Hendler, J., 2001, The Semantic Web, *Scientific American* (May 2001).

Berners-Lee, T., Cailliau, R., Groff, J. F. ve Pollermann, B., 1992, World-Wide Web: The Information Universe, *Internet Research*, 2 (1), 52-58.

Berners-Lee, T., Hendler, J. ve Lassila, O., 2001, The Semantic Web, *Scientific American* (May).

Berners-Lee, T., 2006, Linked Data, http://www.w3.org/DesignIssues/LinkedData.html:

Berners-lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A. ve Sheets, D., 2006, Tabulator : Exploring and Analyzing linked data on the Semantic Web.

Berton, D., Klock, B., Glover, E. ve Kordik, S., 2004, United States Patent US20040143644 A1-Meta-search engine architecture.

Bizer, C., Heath, T. ve Berners-Lee, T., 2009, Linked data-the story so far, *International Journal on Semantic Web and Information Systems*, 5 (3), 1-22.

Bojars, U., Passant, A., Cyganiak, R. ve Breslin, J., 2008, Weaving sioc into the web of linked data.

Boldi, P., Codenotti, B., Santini, M. ve Vigna, S., 2004, UbiCrawler: a scalable fully distributed Web crawler, *Software: Practice and Experience*, 34 (8), 711-726.

Bosch, T., Cyganiak, R., Gregory, A. ve Wackerow, J., 2013, DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data.

Brin, S. ve Page, L., 1998, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30 (1), 107-117.

Buil-Aranda, C., 2012, Federated Query Processing for the Semantic Web, (January).

Buil-Aranda, C. ve Hogan, A., 2013, SPARQL Web-Querying Infrastructure: Ready for Action?, 277-293.

Campinas, S. ve Ceccarelli, D., 2011, The Sindice-2011 dataset for entity-oriented search in the web of data, 26-32.

Chakrabarti, S., van den Berg, M. ve Dom, B., 1999, Focused crawling: a new approach to topic-specific Web resource discovery, *Computer Networks*, 31 (11-16), 1623-1640.

Chebolu, P. ve Melsted, P., 2008, PageRank and the random surfer model, 1010-1018.

Cheng, G. ve Qu, Y., 2009, Searching Linked Objects with Falcons, *International Journal on Semantic Web and Information Systems*, 5 (3), 49-70.

Craswell, N. ve Soboroff, I., 2005, Overview of the TREC-2005 Enterprise Track Email search task, 1-7.

Cyganiak, R., Catasta, M. ve Tummarello, G., 2009, Towards ECSSE : live Web of Data search and integration.

Cyganiak, R. ve Jentzsch, A., 2014, The Linking Open Data cloud diagram.

Cyganiak, R. ve Jentzsch, A., 2017, The Linking Open Data cloud diagram.

D'Aquin, M., Motta, E., Euzenat, J., Rhne-alpes, I. G., Hall, W. ve Keynes, M., 2011, Watson , more than a Semantic Web search engine, *Semantic Web*, 2 (1), 55-63.

Davies, J. ve Weeks, R., 2004, QuizRDF: Search technology for the semantic web, 1-8.

Delbru, R., Campinas, S. ve Tummarello, G., 2012, Searching web data: An entity retrieval and high-performance indexing model, *Web Semantics: Science, Services and Agents on the World Wide Web*, 10, 33-58.

Demartini, G., Iofciu, T. ve Vries, A. P. D., 2010, Overview of the INEX 2009 Entity Ranking, 254-264.

Ding, L., Pan, R., Finin, T. ve Joshi, A., 2005, Finding and ranking knowledge on the semantic web, (November), 156-170.

Dodds, L., 2006, Slug: A semantic web crawler, *Proceedings of Jena User Conference*.

Du, Y. ve Hai, Y., 2013, Semantic ranking of web pages based on formal concept analysis, *Journal of Systems and Software*, 86 (1), 187-197.

Elbassuoni, S., Ramanath, M., Schenkel, R., Marcin, S. ve Weikum, G., 2009, Language-model-based ranking for queries on RDF-graphs, 977-986.

Elbassuoni, S., Ramanath, M., Schenkel, R. ve Weikum, G., 2010, Searching RDF Graphs with SPARQL and Keywords, *IEEE Data Eng. Bull*, 33 (1), 16-24.

Elbassuoni, S. ve Blanco, R., 2011, Keyword search over RDF graphs, *Proceedings of the 20th ACM international \ldots*, 237-242.

Elbassuoni, S., Ramanath, M. ve Weikum, G., 2012, RDF Xpress: a flexible expressive RDF search engine, 1013-1013.

Ermilov, I., Lehmann, J., Martin, M. ve Auer, S., 2016, LODStats: The Data Web Census Dataset, 38-46.

Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E. ve Castells, P., 2008, Semantic Search Meets the Web, *2008 IEEE International Conference on Semantic Computing*, 253-260.

Ferrara, A., Informatica, D., Genta, L. ve Montanelli, S., 2013, Linked Data Classification : a Feature-based Approach, *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, 75-82.

Finin, T., Ding, L., Finnin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. ve Sachs, J., 2004, Swoogle : A Search and Metadata Engine for the Semantic Web, 652-659.

Franz, T., Schultz, A., Sizov, S. ve Staab, S., 2009, Triplerank: Ranking semantic web data by tensor decomposition, 213-228.

Grlitz, O. ve Staab, S., 2011, SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions., *Cold*.

Gulli, a. ve Signorini, A., 2005, Building an open source meta-search engine, *Special interest tracks and posters of the 14th international conference on World Wide Web - WWW '05*, 1004.

Harth, A., Umbrich, J. ve Decker, S., 2006, Multicrawler: A pipelined architecture for crawling and indexing semantic web data, 258-271.

Harth, A., Hogan, A., Delbru, R., Riain, S. O. ve Decker, S., 2007, SWSE : Answers Before Links !

Harth, A., Kinsella, S. ve Decker, S., 2009, Using naming authority to rank data and ontologies for web search, *International Semantic Web Conference*, 277-292.

Hogan, A., Harth, A. ve Decker, S., 2006, Reconrank: A scalable ranking method for semantic web data with context.

Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A. ve Decker, S., 2011, Semantic Search- Reading - Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine, *Web Semantics: Science, Services and Agents on the World Wide Web*, 9 (4), 365-401.

Howe, A. E. ve Dreilinger, D., 1997, SavvySearch: A Meta-Search Engine that Learns which Search Engines to Query, *AI Magazine*, 18 (2), 12-25.

Isele, R., Bizer, C. ve Harth, A., 2010, LDSpider An open-source crawling framework for the Web of Linked Data, In: Proceedings of the 2010 International Conference on Posters \& Demonstrations Track-Volume 658, Eds: CEUR-WS. org, 29-32.

Kafer, T., Abdelrahman, A., Umbrich, J. ve O'Byrne, P., 2013, Observing linked data dynamics, *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*.

Kamilaris, A., Yumusak, S. ve Ali, M. İ., 2016, WOTS2E: A search engine for a Semantic Web of Things, *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, 436-441.

Kaptein, R. ve Kamps, J., 2013, Exploiting the category structure of Wikipedia for entity ranking, *Artificial Intelligence*, 194, 111-129.

Karnstedt, M., Sattler, K.-U. ve Hauswirth, M., 2012, Scalable distributed indexing and query processing over Linked Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, 10, 3-32.

Kasneci, G., Suchanek, F. M., Ifrim, G., Ramanath, M. ve Weikum, G., 2008, NAGA : Searching and Ranking Knowledge, 00, 953-962.

Kenneth, A., McMahon, J. ve Us, C. A., 2012, United States Patent US7805432 B2 Meta Search Engine. 1.

Kim, J. Y. ve Croft, W. B., 2012, A field relevance model for structured document retrieval, 97-108.

Kleinberg, J. M., 1999, Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, 46 (5), 604-632.

Knight, J. P., 1996, Resource discovery on the internet, *New Review of Information Networking*, 2, 3-14.

Kontokostas, D. ve Westphal, P., 2014, Test-driven evaluation of linked data quality, 747-757.

Kruskal, W. H. ve Wallis, W. A., 1987, Citation Classic - Use of Ranks in One-Criterion Variance Analysis, *Current Contents/Arts & Humanities* (40), 20-20.

Lalithsena, S., Hitzler, P., Sheth, A. ve Jain, P., 2013, Automatic Domain Identification for Linked Open Data, *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 205-212.

Lawrence, S. R. ve Giles, C. L., 2006, United States Patent US6999959 B1-Meta search engine. 1.

Lei, Y., Uren, V. ve Motta, E., 2006, SemSearch : A Search Engine for the Semantic Web, In: EKAW, Eds: Springer, 238-245.

Liu, W. ve Du, Y., 2014, A Novel Focused Crawler Based On Cell-Like Membrane Computing Optimization Algorithm, *Neurocomputing*, 123, 266-280.

Mann, H. B. ve Whitney, D. R., 1947, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *The Annals of Mathematical Statistics*, 18 (1), 50-60.

Materne, A. ve Sleightholme, G., 2013, Methods of ranking search results for searches based on multiple search concepts carried out in multiple databases, *World Patent Information*, 1-12.

Melo, G. D., Hose, K. ve de Melo, G., 2013, Searching the web of data, *European Conference on Information Retrieval*, 869-873.

Meusel, R., Spahiu, B., Bizer, C. ve Paulheim, H., 2015, Towards Automatic Topical Classification of LOD Datasets, *CEUR Workshop Proceedings*, 1409.

Mhleisen, H. ve Bizer, C., 2012, Web Data Commons-Extracting Structured Data from Two Large Web Corpora., *Ldow*, 2-5.

Miller, G. a., 1995, WordNet: a lexical database for English, *Communications of the ACM*, 38 (11), 39-41.

Miller, R. C. ve Bharat, K., 1998, SPHINX: a framework for creating personal, site-specific Web crawlers, *Computer Networks and ISDN Systems*, 30, 119-130.

Mirizzi, R., Ragone, A., Noia, T. D. ve Sciascio, E. D., 2010, Ranking the linked data: the case of dbpedia, 337-354.

Patel, C., Supekar, K., Lee, Y. ve Park, E. K., 2003, OntoKhoj: a semantic web portal for ontology searching, ranking and classification, In: Proceedings of the 5th ACM international workshop on Web information and data management, Eds: ACM, 58-61.

Ponte, J. M. ve Croft, W. B., 1998, A language modeling approach to information retrieval, In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Eds: ACM, 275-281.

Pound, J., Mika, P. ve Zaragoza, H., 2010, Ad-hoc object retrieval in the web of data, *Proceedings of the 19th international conference on World wide web - WWW '10*, 771.

Radu, I.-G. ve Rebedea, T., 2014, A focused crawler for Romanian words discovery, *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference*, 1-6.

Raghavan, S. ve Garcia-Molina, H., 2000, Crawling the hidden web, *Stanford*, *Technical Report*.

Robertson, S., Zaragoza, H. ve Taylor, M., 2004, Simple BM25 extension to multiple weighted fields, In: Proceedings of the thirteenth ACM international conference on Information and knowledge management, Eds: ACM, 42-49.

Rocha, C., Schwabe, D. ve Arago, M. P., 2004, A Hybrid Approach for Searching in the Semantic Web, In: Proceedings of the 13th international conference on World Wide Web, Eds: ACM, 374-383.

Roder, M., Ngomo, A.-C. N., Ermilov, I. ve Both, A., 2015, Detecting Similar Linked Datasets Using Topic Modelling, *International Semantic Web Conference*, 3-19.

Rungsawang, A. ve Angkawattanawit, N., 2005, Learnable topic-specific web crawler, *Journal of Network and Computer Applications*, 28, 97-114.

Saleem, M. a., 2014, HiBISCuS: Hypergraph-based source selection for SPARQL endpoint federation, *European Semantic Web Conference*, 176-191.

Scaiella, U., Informatica, D., Ferragina, P., Marino, A. ve Ciaramita, M., 2012, Topical Clustering of Search Results, *Proceedings of the fifth ACM international conference on Web search and data mining* (May), 223-232.

Shah, V., Patni, R., Patani, V. ve Shah, R., 2014, Understanding Focused Crawler, *International Journal of Computer Science \& Information Technologies*, 5 (5), 6849-6852.

Sheldon, M. a., Duda, A., Weiss, R. ve Gifford, D. K., 1995, Discover: a resource discovery system based on content routing, *Computer Networks and ISDN Systems*, 27 (95), 953-972.

Shi, B., 2010, Semantic focused crawler based on Q-learning and Bayes classifier, *2010 3rd International Conference on Computer Science and Information Technology*, 420-423.

Shkapenyuk, V. ve Suel, T., 2002, Design and implementation of a high-performance distributed Web crawler, *Proceedings 18th International Conference on Data Engineering*, 357-368.

Steiner, T. ve Mhleisen, H., 2014, Weaving the Web (VTT) of Data, *Ldow*.

Stojanovic, N., Studer, R. ve Stojanovic, L., 2003, An Approach for the Ranking of Query Results in the Semantic Web, 500-516.

Stoyanovich, J., Bedathur, S., Berberich, K. ve Weikum, G., 2007, EntityAuthority: Semantically Enriched Graph-Based Authority Propagation., (WebDB).

Tonon, A. ve Catasta, M., 2013, TRank: Ranking Entity Types Using the Web of Data, *International Semantic Web Conference*, 640-656.

Tran, T., Wang, H. ve Haase, P., 2009, Hermes: Data Web search on a pay-as-you-go integration infrastructure, *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3), 189-203.

Tran, T., Herzig, D. M. D. M. ve Ladwig, G., 2011, SemSearchPro–using semantics throughout the search process, *Web Semantics: Science, Services and Agents on the World Wide Web*, 9 (4), 349-364.

Tuarob, S., Pouchard, L. C., Mitra, P. ve Giles, C. L., 2015, A generalized topic modeling approach for automatic document annotation, *International Journal on Digital Libraries*.

Tummarello, G., Delbru, R. ve Oren, E., 2007, Sindice. com: Weaving the open linked data, In: The Semantic Web, Eds: Springer, 552-565.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R. ve Decker, S., 2010, Sig.ma: Live views on the Web of Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, 8 (4), 355-364.

Umbrich, J., Hogan, A., Polleres, A. ve Decker, S., 2014, Link Traversal Querying for a Diverse Web of Data, *Semantic Web*.

Vandenbussche, P.-y., Hogan, A. ve Buil-aranda, C., 2017, SPARQLES : Monitoring Public SPARQL Endpoints, *Semantic Web*, 8 (6), 1049-1065.

Vandenbussche, P. Y., Aranda, C. B., Hogan, A. ve Umbrich, J., 2013, Monitoring the Status of SPARQL Endpoints, 1380 (3130617), 3-6.

Wan, G., Ding, Y., Li, B. ve Tan, X., 2014, VRobot: A crawler of education and vocation, *2014 9th International Conference on Computer Science \& Education* (Iccse), 473-476.

Wang, H., Liu, Q., Penin, T., Fu, L., Zhang, L., Tran, T., Yu, Y. ve Pan, Y., 2009, Semplore: A scalable IR approach to search the Web of Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3), 177-188.

Yang, S.-Y., 2010, OntoCrawler: A focused crawler with ontology-supported website models for information agents, *Expert Systems With Applications*, 37 (7), 5381-5389.

Yumusak, S., Dogdu, E. ve Kodaz, H., 2014, A Short Survey of Linked Data Ranking, *2014 ACM Southeast Regional Conference*, Kennesaw, Georgia, 48.

Yumusak, S., Dogdu, E., Kodaz, H. ve Kamilaris, A., 2017, SpEnD: Linked Data SPARQL Endpoints Discovery Using Search Engines, *IEICE Transactions on Information and Systems*, E100.D (4), 758-767.

Yumusak, S., Dogdu, E. ve Kodaz, H., 2018, Classification of Linked Data Sources Using Semantic Scoring, *IEICE Transactions on Information and Systems*, E101-D (1), -.

EKLER

EK-1 Detay Tablolar

Çizelge EK- 1.1. Uç noktalar için Stf-Idf skoru en yüksek olan terimlerin detaylı listesi (Yumusak ve ark., 2018)

		c_t		l_t		c_h		l_h	
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
1	bioportal.bio2rdf.org/sparql	cytochrome	biochemistry	fenestra	otology	epitopes	situation	apophysis	outgrowth
2	linkedspl.bio2rdf.org/sparql	insert	film	ringers	quoits	insert	break	balm	remedy
3	pubmed.bio2rdf.org/sparql	processor	photography	neurology	neurology	processor	worker	neurology	medical specialty
4	internal.opnetendata.cz:8890/sparql	net	field hockey	zombie	voodoo	exchange	capture	blitzkrieg	attack
5	wit.istc.cnr.it:8894/sparql	brother	religion	joseph	Old Testament	advocate	lawyer	piste	ski run
6	data.metamatter.nl/sparql	church	church service	rabbi	Hebrew	hackers	programmer	showrooms	panopticon
7	ruian.linked.opendata.cz/sparql	processor	photography	processor	photography	processor	worker	processor	hardware
8	dbpedia-live.openlinksw.com/sparql	hero	Greek mythology	grace	Christian theology	dip	angle	philosopher	scholar
9	live.dbpedia.org/sparql	hero	Greek mythology	grace	Christian theology	dip	angle	philosopher	scholar
10	linked.opendata.cz/sparql	net	field hockey	adenosine	biochemistry	exchange	capture	bpi	density
11	cr.eionet.europa.eu/sparql	filling	dentistry	renting	car	disclaimer	repudiation	growing	production
12	semantic.etackea.europa.eu/sparql	seafaring	renting	car	pentecost	Jewish holy day	renting	transaction	
13	linkeddata.uriburner.com/sparql	television	television	smash	motor vehicle	routers	device	disclaimer	repudiation
14	proxy.uriburner.com/sparql	television	television	ninja	Nipponese	routers	device	raises	gamble
15	uriburner.com/sparql	television	television	ninja	Nipponese	routers	device	raises	gamble
16	virtuoso.gbpn.org/sp	energy	physics	energy	physics	energy	physical phenome	energy	physical phenome

		c_t		l_t		c_h		l_h	
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
	arql						non		non
17	wiktionary.dbpedia.org/sparql	station	navy	drop	Drug	sentence	final judgment	check	chess move
18	mlode.nlp2rdf.org/sparql	station	navy	drop	drug	sentence	final judgment	check	chess move
19	data.oceandrilling.org/sparql	charge	tax	grace	Christian theology	drill	training	jacobs	patriarch
20	semantic.ckan.net/sparql	charge	tax	tag	tag	inflation	explosion	tag	touch
21	semantic.datahub.io/sparql	charge	tax	tag	tag	inflation	explosion	tag	touch
22	hanne.aksw.org:8892/sparql	baldr	Norse mythology	rabbis	Hebrew	comet	extraterrestrial object	fighter	airplane
23	data.bnf.fr/sparql	images	psychology	nibelungen	Teuton	images	appearance	combats	battle
24	fantom5.nanopub.org/sparql	rna	biochemistry	inversions	counterpoint	insert	break	insert	break
25	data.utpl.edu.ec/utpl/alod/sparql	ambrosi	classical mythology	ambrosia	classical mythology	inclination	angle	amphisbaena	mythical monster
26	serendipity.utpl.edu.ec/lod/sparql	ambrosi	classical mythology	ambrosia	classical mythology	inclination	angle	amphisbaena	mythical monster
27	dbpedia.inria.fr/sparql	sabre	fencing	vampire	folklore	confession	penance	bengali	Asian
28	data.allie.dbcls.jp/sparql	process or	photography	antigen	immunology	processor	worker	mp	lawman
29	linkedstat.spaziodati.eu/sparql	process	photography	book	card game	articles	determiner	section	expanse
30	dati.camera.it/sparql	account	history	dona	Spanish	ai	agency	mafia	organized crime
31	it.dbpedia.org/sparql	anas	antiquity	television	television	salute	greeting	inclination	angle
32	lodlaundromat.org/sparql	header	soccer	header	soccer	client	computer	nodes	point
33	sparql.backend.lodlaundromat.org	header	soccer	header	soccer	client	computer	nodes	point
34	es-la.dbpedia.org/sparql	ishmael	Old Testament	aves	ornithology	umma	community	phylum	social group
35	nl.dbpedia.org/sparql	moses	Old Testament	athene	Greek mythology	libel	defamation	bolt	abandonment

		c_t		l_t		c_h		l_h	
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
36	kaiko.getalp.org/sparql	process	photography	processor	photography	processor	worker	processor	hardware
37	lab.environment.data.gov.au/sparql	stations	navy	pilot	aircraft	percentile	mark	horn	noisemaker
38	open-data.europa.eu/sparql	telecommunication	telecom	games	game	indicator	coloring material	exchange	capture
39	sparql.hegroup.org/sparql	rna	biochemistry	nodule	mineralogy	spasms	constriction	tonicity	tension
40	bis.270a.info/sparql	quarter	professional basketball	education	education	clauses	grammatical construction	subjects	term
41	fao.270a.info/sparql	quarters	professional basketball	education	education	clauses	grammatical construction	subjects	term
42	lod.sztaki.hu/sparql	charge	tax	book	card game	charge	liabilities	email	electronic communication
43	pt.dbpedia.org/sparql	tv	television	television	television	tv	receiving system	fortes	volume
44	imf.270a.info/sparql	quarters	professional basketball	education	education	clauses	grammatical construction	subjects	term
45	ichoose.twc.rpi.edu/sparql	charge	tax	body	homo	charge	liabilities	column	file
46	cs.dbpedia.org/sparql	hymen	Greek mythology	tristan	legend	offside	mistake	relativity	scientific theory
47	rdf.imim.es/sparql	gene	molecular biology	article	contract	medicine	drug	article	determiner
48	sparql.opennmobilenetwork.org	bengali	Hinduism	article	contract	bengali	Asian	subjects	term
49	en.openei.org/sparql	energy	physics	utilities	economics	waste	deed	easements	prerogative
50	data.clarosnet.org/sparql	fathers	Christianity	mihrab	Islam	fathers	theologian	amphitheater	gallery
51	data.cubis.nl/sparql	accounts	history	alcides	classical mythology	accounts	record	punt	kick
52	id.dbpedia.org/sparql	jati	Hinduism	obi	West Indies	guru	religious leader	jati	caste
53	data.bbib.no/sparql	synagogue	Judaism	posting	bookkeeping	synagogue	place of worship	contraindication	reason
54	epo.publicdata.eu/sparql	work	physics	foils	fencing	conviction	final judgment	vibrations	wave
55	linked-	tao	Taoism	optative	Sanskrit	abaya	robe	clauses	grammati

		c_t		l_t		c_h		l_h
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
	data.org/sparql								cal construction
56	opendata-bundestag.de/sparql	tao	Taoism	optative	Sanskrit	abaya	robe	clauses	grammatical construction
57	glycoinfo.org/lodestar/sparql	charge	psychoanalysis	rna	biochemistry	affinity	kinship	argument	computer address
58	linkeddata.ge.imati.cnr.it:8890/sparql	relationship	anthropology	cultivation	farming	curry	dish	accretion	increment
59	healthdata.tw.rpi.edus/sparql	channel	river	lot	Old Testament	column	file	indicators	coloring material
60	data.sepa.org.uk	water	river	don	Spanish	water	thing	bail	legal system
61	data.linkedtv.eu:8890/sparql	exhibition	art	herr	German	plate	base	mensch	good person
62	sparql.wikipathways.org	set	psychology	rna	biochemistry	set	abstraction	hells	imaginary place
63	environmentstation.data.gov.uk/sparql/bwq/query	navy	shore	lake	colonies	animal group		birling	twirl
64	linguistic.linkeddata.es/sparql	lcostas	vertebrate	translation	genetics	costas	bone	translation	transformation
65	eu.dbpedia.org/sparql	amazon	Greek mythology	mars	Roman mythology	frau	title of respect	camp	military quarters
66	linkeddata.finki.ukim.mk/sparql	diana	Roman mythology	ringer	quoits	infusion	instillation	infusion	instillation
67	ieeevis.tw.rpi.edu/sparql	sides	animal	bidding	bridge	citation	speech act	serf	thrall
68	rdf.neuinfo.org/sparql	editing	literature	inversions	counterpoint	recombination	combining	insert	break
69	lod.bco-dmo.org/sparql	crown	dentistry	transmitter	microorganism	parameter	computer address	squid	seafood
70	sparql.asn.desire2learn.com:8890/sparql	habits	religion	education	education	expectation	mean	correlation	parametric statistic
71	data.linkedu.eu/kis/query	complet	American football	literature	literature	quartile	mark	quartile	mark
72	services.d	insert	film	menorah	Judaism	insert	break	menorah	candelabr

		c_t		l_t		c_h		l_h	
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
	ata.gov.uk/education/sparql								um
73	digital-agenda-data.eu/sparql	quarter	professional basketball	reviews	accounting	indicator	coloring material	reviews	accounting
74	digital-agenda-data.eu/data/sparql	quarter	professional basketball	reviews	accounting	indicator	coloring material	reviews	accounting
75	db.lodc.jp/templessparql	Judaism	circulation	library science	ceramics	instrumentality	immunity		condition
76	dati.san.beniculturali.it/sparql	processor	photography	justice	legislation	processor	worker	curia	administration
77	linked-statistics.gr/sparql	division	botany	community	ecology	citizenship	legal status	code	coding system
78	data.webfoundation.org/sparql	relationship	anthropology	education	education	rectification	refining	computers	machine
79	location.testproject.eu/sparql	port	ship	decision	boxing	exchange	capture	decision	result
80	cpsv.testproject.eu/sparql	port	ship	decision	boxing	exchange	capture	decision	result
81	spcdata.digitpa.gov.it:8899/sparql	pit	auto racing	don	Spanish	tares	counterweight	brig	penal institution
82	data.aalto.fi/sparql	disturbances	psychiatry	consumption	economics	truss	bandage	fatigue	duty assignment
83	bfs.270a.info/sparql	quarters	professional basketball	education	education	clauses	grammatical construction	subjects	term
84	stats.270a.info/sparql	xxx	genetics	repeaters	electrical engineering	xxx	sex chromosome	parities	bit
85	data.logainm.ie/sparql	bishop	Roman Catholic	posting	bookkeeping	guardhouse	headquarters	constituent	syntagma
86	smartcity.linkeddata.es/sparql	council	Christianity	temple	Judaism	pilot	aviator	mess	dining room
87	sadiframework.org/registry/sparql	rna	biochemistry	accession	civil law	insert	break	citations	speech act
88	biordf.net/sparql	rna	biochemistry	accession	civil law	insert	break	citations	speech act
89	portal.chemicalsemantics.com/	charges	tax	h	thermodynamics	nucleus	midpoint	bond	recognizance

		c_t		l_t		c_h		l_h	
#	Uç Nokta	kelime	terim	kelime	terim	kelime	terim	kelime	terim
	cs/sparql								
90	leipzig-data.de:8890/sparql	frau	German	games	game	frau	title of respect	passage	legislation
91	lod.gesis.org/thesoz/sparql	tag	tag	mensch	Yiddish	tag	touch	mensch	good person
92	data.linkedu.eu/ocw/query	margin	corporate finance	cmb	cosmology	factorizations	resolution	terrorists	radical
93	data.globalchange.gov/sparql	literature	literature	hybrid	Latin	manifestation	protest	desktop	screen
94	newt.oerc.ox.ac.uk:8890/sparql	record	photography	cultures	archeology	strains	nervousness	contrast	scope
95	data.ox.ac.uk/sparql	literature	literature	mover	order	fullerenes	carbon	appointments	disposal
96	crashmap.okfn.gr:8890/sparql	processor	photography	processor	photography	processor	worker	driver	utility program
97	semanticlab.jrc.ec.europa.eu:4433/sparql	axiom	logic	veda	Sanskrit	ontology	arrangement	veda	sacred text
98	wordnet.okfn.gr:8890/sparql	record	photography	americana	furniture	head	coil	americana	artifact
99	services.data.gov.uk/statistics/sparql	councils	Christianity	vicars	Episcopal Church	councils	assembly	vicars	clergyman
100	matvocab.org/sparql	pitch	ship	tracer	radiology	reinforcement	stimulation	accelerator	activator
101	aliada.scanbit.net:8890/sparql	judith	Apocrypha	use	economics	manifestation	protest	carriers	immune
102	waes.servusnet.com/sparql	charge	tax	tag	tag	charge	liabilities	tag	touch
103	zbw.eu/beta/sparql/stw/query	range	mathematics	industries	industry	connections	supplier	inflation	explosion
104	lod.nature.go.kr/sparql	phylum	biology	characters	genetics	phylum	social group	characters	attribute
105	eatld.et.tu-dresden.de/sparql	account	history	connections	narcotic	devices	emblem	connections	supplier
106	wiktionary.dbpedia.org/sparql	station	navy	drop	drug	sentence	final judgment	check	chess move
107	wit.istc.cnr.it:8894/sparql	brother	religion	joseph	Old Testament	advocate	lawyer	piste	ski run
108	wordnet.okfn.gr:889	record	photography	americana	furniture	head	coil	americana	artifact

		c_t		l_t		c_h		l_h
#	Uç Nokta kelime	terim	kelime	terim	kelime	terim	kelime	terim
	0/sparql							
109	zbw.eu/be range ta/sparql/s tw/query	mathemat ics	industries	industry	connections	supplier	inflation	explosion

Çizelge EK- 1.2. Keşfedilen SPARQL uç noktalarının alan adları bazında tespit edilen uç nokta sayısı

Alan Adı	Uç nokta sayısı
rkbexplorer.com	65
b3kat.de	58
insee.fr	34
dbpedia.org	27
fundacionctic.org	23
data.gov.uk	21
270a.info	15
ign.fr	13
linkeddata.es	13
eagle-i.net	12
rpi.edu	9
aksw.org	7
geolba.ac.at	7
sepa.org.uk	7
auth.gr	6
europa.eu	6
lod.ac	6
openlinksw.com	6
tso.co.uk	6
bio2rdf.org	5
open.ac.uk	5
cnr.it	4
datahub.kr	4
ordnancesurvey.co.uk	4
uriburner.com	4
dbcls.jp	3
getty.edu	3
linkedu.eu	3
logainm.ie	3
opendata.cz	3
openmobilenetwork.org	3
ox.ac.uk	3
semantic-web.at	3
um.es	3
uniprot.org	3
upm.es	3
202.45.139.84	2
4store.org	2
aalto.fi	2
bartoc.org	2
beniculturali.it	2
cedar-project.nl	2
colinda.org	2
data.gov.au	2
data.gov.ru	2
deri.ie	2
deusto.es	2
dydra.com	2
ebi.ac.uk	2
gbpn.org	2
iringsandbox.org	2
ldf.fi	2
libriotech.no	2
linked-statistics.org	2
mmisw.org	2
nii.ac.jp	2
publicdata.eu	2

reegle.info	2
soton.ac.uk	2
southampton.ac.uk	2
thegazette.co.uk	2
unicamp.br	2
wolterskluwer.de	2
yafjp.org	2
zaragoza.es	2
zbw.eu	2
62.217.127.118	1
apc.gov.tw	1
aragon.es	1
archaeologydataservice.ac.uk	1
archiveshub.ac.uk	1
atted.jp	1
australiancurriculum.edu.au	1
babelnet.org	1
bbib.no	1
bcn.cl	1
bco-dmo.org	1
beef.org.pl	1
bibliotheek.nl	1
biordf.net	1
bl.uk	1
bne.es	1
bnf.fr	1
caceres.es	1
caicyt.gov.ar	1
camera.it	1
ccr.it	1
chemicalsemantics.com	1
ckan.net	1
clarosnet.org	1
colorado.edu	1
contextdatacloud.org	1
coxpresdb.jp	1
creativeartefact.org	1
ctsaconnect.org	1
cubiss.nl	1
culture.fr	1
curriculum.edu.au	1
dariah.eu	1
datahub.io	1
datameti.go.jp	1
datao.net	1
data-observatory.org	1
dbtune.org	1
deichman.no	1
desire2learn.com	1
dewey.info	1
digital-agenda-data.eu	1
digitpa.gov.it	1
disit.org	1
eagle-network.eu	1
edina.ac.uk	1
ekt.gr	1
euscreen.eu	1
factforge.net	1
fer.hr	1
festdb.org	1

freeyourmetadata.org	1
genome.jp	1
geodan.nl	1
gesis.org	1
getalp.org	1
globalchange.gov	1
glycoinfo.org	1
greggkellogg.net	1
hamakei-opendata.com	1
hegroup.org	1
heritagedata.org	1
i2g.pl	1
iana.org	1
idease.info	1
identifiers.org	1
ifmo.ru	1
ifpri.org	1
imim.es	1
indiana.edu	1
influencetracker.com	1
interridge.org	1
invemar.org.co	1
isaf2014.info	1
jesandco.org	1
kdata.kr	1
klappstuhlclub.de	1
kontrax.bg	1
ksharp.net	1
ksu.ru	1
kth.se	1
kupkb.org	1
l3s.de	1
learningsparql.com	1
leipzig-data.de	1
lenka.no	1
linkedarc.net	1
linkedbrainz.org	1
linked-data.org	1
linkeddatahub.com	1
linkedevents.org	1
linkedfood.org	1
linkedgeodata.org	1
linkedlifedata.com	1
linkedmdb.org	1
linked-statistics.gr	1
linkedtv.eu	1
linklion.org	1
lodc.jp	1
lodlaundromat.org	1
lotico.com	1
lter-europe.net	1
mathbiol.org	1
matvocab.org	1
mcu.es	1
meducator3.net	1
metalex.eu	1
metamatter.nl	1
monodzukurilod.org	1
msc2010.org	1
muninn-project.org	1

myexperiment.org	1
nanopub.org	1
nature.go.kr	1
neuinfo.org	1
nexacenter.org	1
nih.gov	1
nlp2rdf.org	1
nobelprize.org	1
nstac.go.jp	1
oceandrilling.org	1
oclc.org	1
okfn.gr	1
okfn.org	1
ontotext.com	1
opendatasupport.eu	1
openei.org	1
openspring.net	1
opmw.org	1
organic-edunet.eu	1
oszk.hu	1
panlex.org	1
p-dpa.net	1
poolparty.biz	1
posccaesar.org	1
pubmlst.org	1
rechercheisidore.org	1
rhizomik.net	1
rism.info	1
sadiframework.org	1
scanbit.net	1
semanticweb.org	1
senato.it	1
servusnet.com	1
spaziodati.eu	1
symbolicdata.org	1
sztaki.hu	1
techinvestlab.ru	1
telegraphis.net	1
testproject.eu	1
the-fr.org	1
toby.ink	1
tobyinkster.co.uk	1
tsukuba.ac.jp	1
tudelft.nl	1
tu-dresden.de	1
tuwien.ac.at	1
ucd.ie	1
uec.ac.jp	1
ukim.mk	1
umu.se	1
unibo.it	1
uni-muenster.de	1
univ-nantes.fr	1
url.edu	1
utpl.edu.ec	1
uu.se	1
uwindsor.ca	1
vocabularyserver.com	1
webfoundation.org	1
whoi.edu	1

wikipathways.org	1
worldpece.org	1
xdams.org	1
yovisto.com	1

Çizelge EK- 1.3. Keşfedilen SPARQL uç noktalarının üçlü sayıları (100 milyondan fazla üçlü barındıran)

SPARQL Uç Noktası	Üçlü Sayısı
http://babelnet.org/sparql	1.927.476.268
http://goa.bio2rdf.org/sparql	1.376.988.942
http://internal.opendata.cz:8890/sparql	1.247.756.192
http://commons.dbpedia.org/sparql	1.229.690.546
http://dbpedia.org/sparql	1.223.211.963
http://ldf.fi/corsproxy/dbpedia.org/sparql	1.223.049.594
http://linkedgeodata.org/sparql	1.032.032.408
http://wit.istc.cnr.it/sparql	916.223.754
http://ga.dbpedia.org/sparql	913.144.199
http://data.metamatter.nl/sparql	761.572.297
http://integrator.poolparty.biz:8890/sparql	681.181.544
http://ruian.linked.opendata.cz/sparql	639.144.908
https://www.ebi.ac.uk/rdf/services/biosamples/sparql	638.725.750
http://maiana.lodac.nii.ac.jp/sparql	626.078.786
http://lod.geodan.nl/sparql	622.366.962
http://dbpedia-live.openlinksw.com/sparql	560.400.672
http://live.dbpedia.org/sparql/%22	560.293.669
http://live.dbpedia.org/sparql	560.258.245
http://linked.opendata.cz/sparql	555.666.202
http://babelnet.org/sparql	1927476268
http://goa.bio2rdf.org/sparql	1376988942
http://internal.opendata.cz:8890/sparql	1247756192
http://commons.dbpedia.org/sparql	1229690546
http://dbpedia.org/sparql	1223211963
http://ldf.fi/corsproxy/dbpedia.org/sparql	1223049594
http://linkedgeodata.org/sparql	1032032408
http://wit.istc.cnr.it/sparql	916223754
http://ga.dbpedia.org/sparql	913144199
http://data.metamatter.nl/sparql	761572297
http://integrator.poolparty.biz:8890/sparql	681181544
http://ruian.linked.opendata.cz/sparql	639144908
http://maiana.lodac.nii.ac.jp/sparql	626078786
http://lod.geodan.nl/sparql	622366962
http://dbpedia-live.openlinksw.com/sparql	560400672
http://live.dbpedia.org/sparql/%22	560293669
http://live.dbpedia.org/sparql	560258245
http://linked.opendata.cz/sparql	555666202
http://cr.eionet.europa.eu/sparql	483835680
http://semantic.eea.europa.eu/sparql	462188228
http://linkeddata.uriburner.com/sparql	396381240
http://data.uriburner.com/sparql	368273054
http://proxy.uriburner.com/sparql	368069075
http://uriburner.com/sparql	368066841
http://virtuoso.gbpn.org/sparql	308857208
http://mlode.nlp2rdf.org/sparql	303167429
http://doc.metalex.eu:8000/sparql	294615368
http://data.oceandrilling.org/sparql	284665625
http://sv.dbpedia.org/sparql	280643697
http://semantic.ckan.net/sparql	277294235
http://semantic.datahub.io/sparql	276930375
http://wikidata.dbpedia.org/sparql	246389485
http://rechercheisidore.org/sparql	226592863
http://hanne.aksw.org:8892/sparql	222033941
http://data.bnf.fr/sparql	211992257
http://fantom5.nanopub.org/sparql	198660088
http://data-gov.tw.rpi.edu/sparql	195537010
http://data.utpl.edu.ec/utpl/lod/sparql	195028890

http://fr.dbpedia.org/sparql	185377626
http://es.dbpedia.org/sparql	169285270
http://de.dbpedia.org/sparql	159116548
http://datos.bne.es/sparql	143154176
http://data.allie.dbcls.jp/sparql	140280468
http://linkedstat.spaziodati.eu/sparql	136668993
http://dati.camera.it/sparql	129059608
http://it.dbpedia.org/sparql	121901393
http://europeana-triplestore.isti.cnr.it/sparql	116042632
http://lodlaundromat.org/sparql	112639178
http://linkedspending.aksw.org/sparql	109674521
http://es-la.dbpedia.org/sparql	106364790
http://nl.dbpedia.org/sparql	105071423
http://kaiko.getalp.org/sparql	104216280
http://ja.dbpedia.org/sparql	100098304

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Semih Yumuşak
Uyruğu	**:**	TC
Doğum Yeri ve Tarihi	**:**	Nusaybin 16.02.1983
Telefon	**:**	+90 555 565 55 56
e-mail	**:**	semihyumusak@yahoo.com

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Adana Fen Lisesi, Seyhan, Adana	2000
Üniversite	:	Koç Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği	2005
Yüksek Lisans	:	İstanbul Bilgi Üniversitesi, İşletme Fakültesi, İşletme	2008
Doktora	:	Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği A.B.D.	2017

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2005-2007	SAMPAŞ A.Ş.	Yazılım Geliştirme Uzmanı
2007-2011	Semisoft Ltd.Şti.	Yazılım Proje Yöneticisi
2011-2014	KTO Karatay Üniversitesi	Araştırma Görevlisi
2014-2017	KTO Karatay Üniversitesi	Öğretim Görevlisi

UZMANLIK ALANI: İnternet Teknolojileri, Web Madenciliği, Anlamsal Ağlar, Bağlantılı Veri

YABANCI DİLLER: İngilizce

YAYINLAR

Uluslararası hakemli dergilerde yayınlanan makaleler (SCI, SSCI, Arts and Humanities)

Yumusak, S., Dogdu, E., Kodaz, H., Kamilaris, A., Vandenbussche, P. (2017). SpEnD: Linked Data SPARQL Endpoints Discovery Using Search Engines. *IEICE Transactions on Information and Systems*, *E100.D*(4), 758–767. **(Doktora Tezinden)**

Yumusak, S., Dogdu, E., Kodaz (2018). Classification of Linked Data Sources Using Semantic Scoring. *IEICE Transactions on Information and Systems*, *E101-D, No.1,Jan. 2018.***(Doktora Tezinden, Kabul Edildi-Ön Yayında)**

Uluslararası diğer hakemli dergilerde yayınlanan makaleler

Yumusak, S., Dogdu, E., & Kodaz, H. (2014). Tagging Accuracy Analysis on Part-of-Speech Taggers. *Journal of Computer and Communications*, (March), 157–162. http://doi.org/10.4236/jcc.2014.24021

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitabında basılan bildiriler

Uysal, E., **Yumusak, S**., Oztoprak, K., & Dogdu, E. (2017). Sentiment Analysis for the Social Media: A Case Study for Turkish General Elections Categories and Subject Descriptors. *Proceedings of the SouthEast Conference. ACM*, 215–218.

Kamilaris, A., **Yumusak, S**., & Ali, M. I. (2016). WOTS2E: A search engine for a Semantic Web of Things. *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, (December), 436–441. http://doi.org/10.1109/WF-IoT.2016.7845448

Yumusak, S., Munoz, E., Minervini, P., Dogdu, E., & Kodaz, H. (2016). A Hybrid Method for Rating Prediction Using Linked Data Features and Text Reviews. *5th International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*.

Yumusak, S., Dogdu, E., & Kodaz, H. (2014). A Short Survey of Linked Data Ranking. In *Proceedings of the 2014 ACM Southeast Regional Conference* (pp. 14–17). ACM. **(Doktora Tezinden)**

Dogdu, E., Hakimov, S., & **Yumusak, S.** (2014). A Data-Model Driven Web Application Development Framework. In *Proceedings of the 2014 ACM Southeast Regional Conference*. ACM.

Yumusak, S.; Dogdu E.; Kodaz, H. (16-18 Eylul 2013). An Accuracy Analysis on Using Different Initial Taggers for the Transformation-Based Error-Driven Learning POS Tagger. *6-th International Conference "Advanced Computer Systems and Networks: Design and Application"*. Lviv Polytechnic National University, Lviv, Ukrayna