
GAZİ ÜNİVERSİTESİ
Gazi
1926

İNTERNETTE HETEROJEN VERİ KAYNAKLARINDAN VERİNİN TOPLANMASI, DOĞRULANMASI VE SORGULANMASI

Serdar Kürşat SARIKOZ

DOKTORA TEZİ
BİLGİSAYAR BİLİMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

EYLÜL 2023

Serdar Kürşat SARIKOZ tarafından hazırlanan INTERNETTE HETEROJEN VERİ KAYNAKLARINDAN VERİ TOPLANMASI, DOĞRULANMASI VE SORGULANMASI adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Bilimleri Ana Bilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. M. Ali AKCAYOL

Bilgisayar Bilimleri A.B.D. Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...……..……..

Başkan: Prof. Dr. O. Ayhan ERDEM

Bilgisayar Mühendisliği A.B.D. Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ……..…….

Üye: Prof. Dr. Sevil ŞEN

Yazılım A.B.D, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ……..…….

Üye: Prof. Dr. Hasan Şakir BİLGE

Elektrik-Elektronik Mühendisliği A.B.D. Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...……

Üye: Doç. Dr. A. Talha KABAKUŞ

Bilgisayar Yazılımı A.B.D. Düzce Üniversitesi

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum. ...……

Tez Savunma Tarihi: 06/09/2023

Jüri tarafından kabul edilen bu tezin Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..……

Prof. Dr. Aslıhan TÜFEKÇİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

Bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Serdar Kürşat SARIKOZ

05/10/2023

INTERNETTE HETEROJEN VERİ KAYNAKLARINDAN VERİ TOPLANMASI, DOĞRULANMASI VE SORGULANMASI

(Doktora Tezi)

Serdar Kürşat SARIKOZ

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Eylül 2023

ÖZET

Yapay zekâ alanındaki disiplinler arası araştırmalar ve teknolojik gelişmeler, heterojen veri kaynaklarından gelen büyük hacimli verilerin bilgi grafları sayesinde entegrasyonunu sağlayarak gerçek dünya sorunlarının çözümü ve büyük veri analitiğini olanaklı kılmıştır. Web üzerindeki heterojen veri kaynakları ile çalışmada, bağlantılı veri, anlamsal ağ ve bilgi graflarının entegrasyonu; veri kalitesi ve doğruluk keşfinin birlikte etkileşimini sağlam bir dijital çerçevede sunumunu sağlamaktadır. Bu çerçeve sadece bilginin zenginliğini ve erişilebilirliğini değil, aynı zamanda güvenilirliğini, doğruluğunu ve güvenilirliğini de sağlayarak bilgilerin tutarlılığını ve güvenilirliğini sağlayan dijital çağın önünü açmaktadır. Bir dil için KG oluşturmak, heterojen verilerin karmaşıklığı ve dinamik yapısı nedeniyle zorlukları beraberinde getirmektedir. Heterojen veri kaynakları üzerinden, heterojen bilgi ağlarının oluşturulmasında, saklanmasında, işlenmesinde, analiz edilmesinde ve karmaşık ilişkilerin çıkarılmasında; verimli ve etkin yeni algoritmalar geliştirilmesine ihtiyaç bulunmaktadır. Bu çalışmada, Türkçe diline yönelik olarak heterojen veri kaynaklarından bilgi çıkarımı, makine okunabilir formatta bilgi tabanının oluşturulması ve bilgilerin graf formatında gösterimini sağlayan özgün bir model geliştirilmiştir. Sistem, Wikipedia ve arama motorları üzerinden elde etmiş olduğu veriler üzerinden bilgi çıkarımını, bilgi tabanına eklenmesine yönelik gerekli işlemleri, oluşturmuş olduğu referans bilgi tabanı aracılığı ile bilgilerin doğrulanmasını gerçekleştirilmektedir. Çalışma kapsamında önerilen özgün modelin benzersiz mimarisi, farklı alanlardaki farklı problemlere ve görevlere uyarlanmasına olanak sağlamaktadır. Bu esneklik, modelin çeşitli alanlarda kullanılabilirliğini ve uygulanabilirliğini artırmaktadır. Bununla birlikte Türkçe dilinde Web ölçekli ve genel amaçlı bir bilgi grafı sunması nedeniyle mevcut literatürdeki bir boşluğu doldurmaktadır. Önerilen sistem aynı zamanda arama motorlarına özgün unutulma hakkı ve güvenli arama gösterimini de sunmaktadır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Heterojen Bilgi Ağları, Bilgi Grafları, Gerçekliğin Keşfi
Sayfa Adedi	:	133
Danışman	:	Prof. Dr. M. Ali AKCAYOL

COLLECTING, VERIFYING, AND INQUIRING DATA FROM HETEROGENEOUS DATA SOURCES ON THE INTERNET

(Ph. D. Thesis)

Serdar Kürşat SARIKOZ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF INFORMATICS

September 2023

ABSTRACT

Interdisciplinary research and technological developments in the field of artificial intelligence have enabled the integration of large volumes of data from heterogeneous data sources through knowledge graphs, enabling the solution of real world problems and big data analytics. In working with heterogeneous data sources on the Web, integrating linked data, semantic networks, and knowledge graphs enables the interaction of data quality and truth discovery to be presented in a robust digital framework. This framework paves the way for a digital age that ensures not only the richness and accessibility of information but also its reliability, accuracy, and trustworthiness, ensuring consistency and trustworthiness of information. Creating a QA for a language brings challenges due to heterogeneous data's complexity and dynamic nature. New efficient and effective algorithms are needed for creating, storing, processing, analyzing, and extracting complex relationships from heterogeneous data sources. This study developed a novel model for extracting information from heterogeneous data sources, creating a machine-readable knowledge base, and representing the information in graph format for the Turkish language. The system performs information extraction through the data obtained from Wikipedia and search engines, the necessary operations for adding to the knowledge base, and the verification of the information through the reference knowledge base it has created. The unique architecture of the original model proposed in this study allows it to be adapted to different problems and tasks in different fields. This flexibility increases the usability and applicability of the model in various fields. In addition, it fills a gap in the existing literature by providing a Web-scale and general-purpose knowledge graph in the Turkish language. The proposed system also gives search engines a unique right to be forgotten and a secure search representation.

Science Code	:	92432
Keywords	:	Heterogenous Information Networks, Knowledge Graphs, Truth Discovery
Page Number	:	133
Supervisor	:	Prof. Dr. M. Ali AKCAYOL

TEŞEKKÜR

Doktora çalışmalarım süresince yardımlarını esirgemeyen ve çalışmalarımda tıkandığım tüm noktalarda yönlendirmeleriyle ufkumu açan, değerli bilgi ve birikimleriyle akademik gelişimime katkı sağlayan, kıymetli emek ve destekleri, sabır ve anlayışları nedeni ile danışman hocam Sn. Prof. Dr. M. Ali AKCAYOL'a en içten teşekkürlerimi sunarım. Tez çalışması boyunca desteklerini esirgemeyen ve tezin geliştirilmesine yönelik olumlu katkılarda bulunan tez izleme komitemde yer alan değerli hocalarım Sn. Prof. Dr. Ayhan ERDEM'e ve Sn. Prof. Dr. Sevil ŞEN AKAGÜNDÜZ'e teşekkürlerimi sunarım.

Eğitim hayatım boyunca desteklerini esirgemeyen tüm aileme sonsuz teşekkür ederim. Doktora sürecinin yoğun ve tempolu çalışma süresince sabır ve anlayış gösteren ve her zaman yanımda olan eşim Neslihan SARIKOZ'a, hayatımıza tat katan oğlum Ertuğrul Yusuf SARIKOZ'a ve kızım İpek Zeynep SARIKOZ'a teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

RESİMLERİN LİSTESİ

Resim **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
G	Yönlü bilgi grafı
E	Varlıklar kümesi
R	İlişkiler kümesi
F	Olgular kümesi
(h, r, t)	Özne, ilişki ve nesne üçlüsü
(s, p, o)	Konu, ilişki ve nesne üçlüsü

Kısaltmalar	**Açıklamalar**
API	Uygulama Programlama Arayüzleri (Application Programing Interface)
AI	Yapay Zekâ (Artificial İntelligence)
BT	Bilgi Tabanı
CRF	Koşullu Rasgele Alanlar (Conditional Random Fields)
CWA	Kapalı Dünya Varsayımı (Close World Assumptions)
DOM	Doküman Nesne Veri Modeli (Document Object Model)
EE	Olay Çıkarımı (Event Extraction)
FN	Yanlış Negatif (False Negative)
FP	Yanlış Pozitif (False Positive)
HINs	Heterojen Bilgi Ağları (Heterogeneous information networks)
HMM	Gizli Markov Modelleri (Hidden Markov Models)
HTML	Hiper Metin İşaretleme Dili (Hyper Text Markup Language)
IE	Bilgi Çıkarımı (Information Extraction)
IoT	Nesnelerin Interneti (Internet of Things)
IR	Bilgi Kazanımı **(**Information Retrieval)
IRI	Uluslararasılaştırılmış Kaynak Tanımlayıcı (Internationalized Resource Identifier)

JSON	JavaScript Nesne Gösterimi (JavaScript Object Notation)
KB	Bilgi Tabanı (Knowledge base)
KE	Bilgi Çıkarımı (Knowledge Extraction)
KGs	Bilgi Grafları (Knowledge Graphs)
KR	Bilginin Temsili (Knowledge Representation)
LD	Bağlantılı Veri (Linked Data)
ML	Makine Öğrenimi (Machine Learning)
NER	İsimlendirilmiş Varlık Tespiti (Named Entity Recognition)
NERC	İsimlendirilmiş Varlık Tespiti ve Sınıflandırılması (Named Entity Recognition and Classification)
NED	İsimlendirilmiş Varlık Belirsizliğin Giderimi (Named Entity Disambugution)
NLP	Doğal Dil İşleme (Natural Language Processing)
ODP	Ontology Design Patterns
OIE	Açık Bilgi Çıkarımı (Open Information Extraction)
OWA	Açık Dünya Varsayımı (Open World Assumptions)
OWL	Web Ontoloji Dili (Web Ontology Language)
RDF	Kaynak Tanımlama Çerçevesi (Resource Description Framework)
RDF KG	Kaynak Tanımlama Çerçevesi Bilgi Grafları
RDFS	Kaynak Tanımlama Çerçevesi Şeması (Resource Description Framework Schema)
RE	İlişki Çıkarımı (Relation Extraction)
REGEX	Düzenli İfadeler (Regular Expression)
RSS	Zengin Site Özeti **(**Rich Site Summary)
SQL	Yapılandırılmış Sorgu Dili (Structured Query Language)
SRL	Ardışık Rol Etiketleme (Sequential Role Labeling)
SVM	Karar Destek Makineleri (Support Vector Machines)
SW	Anlamsal Ağ (Semantic Web)
TF-IDF	Terim Frekans- Ters Terim Frekansı (Term Frequency – Inverse Term Frequency)
URI	Tekdüzen Kaynak Tanımlayıcı (Uniform Resource Identifier)
VT	Veri Tabanı
WSD	Kelime Anlam Ayrımı (Word Sense Disambiguation)
XML	eXtensible Markup Language

XPath	Biçim Sayfası Dil Dönüşümü (Stylesheet Language Transformation)
WSD	Kelime Anlamı Ayrıştırma (Word Sense Disambiguation)
WWW	Dünya Geneli Ağ (World Wide Web)
YZ	Yapay Zekâ (Artificial Intelligence)

1. GİRİŞ

Kullanıldığı ilk günden itibaren İnternet, etkileyici bir şekilde gelişmeye ve teknolojik fırsatları da beraberinde getirerek hayatımızın birçok kesiminde yerini almaya devam etmektedir. İnternet ve internet teknolojilerinin, dünyanın dört bir yanındaki insanlar üzerinde olağanüstü bir etkisi olmuştur. İletişim teknolojilerindeki gelişmeler ile birlikte inovasyon, büyüme ve yeni iş modellerinin gelişmesine katkı sağladığı, insan girişimciliğini ve yaratıcılığını teşvik etmekte merkezi bir rol oynadığı, toplumların gelişmesine eşsiz katkılar sunduğu; bilgi paylaşımı, haberleşme, ulaşım, sağlık, eğitim, finans, güvenlik vb. pek çok alanda yirmi birinci yüzyıl uygarlığının vazgeçilmez parçası haline geldiği görülmektedir.

Web, İnternetin bir bileşeni olup, farklı türdeki yapısal ve yapısal olmayan veri yapıları ile bilgiye erişimi ve paylaşımı olanaklı hale getirmiştir. Sürekli artan gürültülü veri içeriği ile birlikte dünyanın en büyük heterojen bilgi ağı ve kaynağı olarak kabul edilen World Wide Web (WWW) 1991'den günümüze kadar geçen süre zarfında, dünyanın herhangi bir köşesinde, yeni bilgilerin sunumunda ve bilgiye ulaşımda birincil kaynak haline gelmiştir. Web üzerinde doğal dilde yer verilen, gürültülü ve kullanıcı davranışlarını yansıtan ve sürekli artan verilerin mevcudiyeti, büyük veri analitiğine yönelik olarak yapay zekâ alanında bilgi kazanımı (Information Retrieval - IR), bilgi sunumu (Knowledge Representation - KR), doğal dil işleme (Natural Language Processing - NLP), makine öğrenimi (Machine Learning - ML) gibi çok sayıda alanda disiplinler arası araştırma çalışmalarına katkı sunmuştur [1].

Dijital teknolojilerdeki hızlı ilerlemeler, nesnelerin interneti (IoT) sosyal ağ uygulamaları, video, ses ve coğrafi konum servislerinin popülerliği, geniş miktarda veri toplama ve analiz yapmaya olanak sağlamıştır. Bulut hizmetlerinin ortaya çıkışıyla birlikte, bulut bilişim hızla daha güvenilir, ölçeklenebilir ve maliyet etkin bir bilgi teknolojileri (BT) çözümü haline gelmiş ve geleneksel sistemlerin yerini almaya başlamıştır. Web üzerindeki heterojen veri kaynakları ile birlikte, bulut teknolojisinde yapılandırılmış ve yapılandırılmamış verilerin, farklı yerlerde dağıtılmış olması ve çeşitli veri kaynaklarının mevcudiyeti, veri ve bilgi temsili, veri entegrasyonu, veri sorgulama, iş analizi ve bilgi keşfi konusunda zorluklar yaratmaktadır.

Diğer taraftan gerçek hayatımızda kullanmakta olduğumuz pek çok bilgi sisteminin, heterojen bileşenlerden oluşmasına rağmen, homojen bilgi ağları yapısında temsil edilmesi bilginin temsili, sunumu ve analizinde kayıpları beraberinde getirmiştir [2,3].

Heterojenlik ifadesi, büyük verinin önemli özelliklerinden birisi olup, aynı zamanda farklı türlerde ve formatlarda bulunan verileri ifade etmektedir. Heterojen ağlar farklı türde düğümler veya bağlantıları içerirken, homojen ağlarda sadece bir tür nesne ve bağlantı bulunmaktadır. Veri heterojenliğinin temel nedeni, farklı kaynaklardan gelen verilerin farklı biçimlerde oluşturulması olup, heterojen verilerin siber güvenlikten, sosyal medya analizine, sağlık sektöründen, akademik ağlara ve e-ticaret gibi pek çok bilgi ağında kullanılmakta olduğu görülmektedir [4].

Dünyanın en büyük heterojen veri ve bilgi ağı olan Web üzerinde yer alan gürültülü ve düşük kaliteli, yapısal olmayan veri ve medya içerikleri, değerli bilimsel, kültürel ve genel hayata ilişkin bilgilerin insanların keşfetmesini zorlaştırmaktadır. Web üzerindeki yapılandırılmamış verilerden bilginin çıkarımı, bilgi tabanlarının oluşturulması ve gösterimine yönelik son yirmi sene içerisinde İngilizce dilinde anlamsal ilişkiler ile zenginleştirilmiş otomatik bilgi tabanlarının ve bilgi graflarının oluşturulmasına yönelik çok sayıda araştırma yapılmıştır [5,8,9,10].

Farklı türdeki milyonlarca düğüm ve düğümler arası ilişkileri içeren büyük hacimli verilerde mevcut analiz yöntemleri yetersiz kalmaktadır. Bu nedenle heterojen veri kaynaklarından alınan bilgilerin, heterojen bilgi ağları (Heterogenous Information Networks- HINs) ve bilgi grafları (Knowledge Graphs – KGs) ile temsil edilerek gerçek dünya problemlerinin çözümüne yönelik çalışmalar önem kazanmıştır. KGs ve HINs benzer şekilde yapılandırılmış, yarı yapılandırılmış veya yapılandırılmamış farklı veri türleri arasındaki çoklu tür ilişkilerin temsili, sunumu, saklanması ve gösterimi amacı ile karşımıza çıkmaktadır. Bununla birlikte KGs, sahip olduğu hiyerarşik zengin şema içeriği, varlıklar arasındaki ontolojilerin sunumu ve bilgi temsilinin rahatlıkla yapılabilmesi yönleriyle HINs'den ayrışmaktadır [6].

Bilgi graflarına yönelik çalışmalar 1970'li yıllardan itibaren yapılıyor olmasına rağmen, bilgi grafları terimi, Google'ın arama motorunda semantik arama yeteneklerinin artırılması amacı ile 2012 yılında bilgi graflarının kullanmaya başlamasıyla yaygın olarak bilinir hale

gelmiştir. Google, bilgi grafları ile arama motoru kullanıcı sorgularını, Web üzerindeki belgeler yerine gerçek dünyaya ait varlıklar üzerinden eşleştirilmesini sağlamıştır. Gerek Web ölçeğindeki çalışmalar gerekse alan özel endüstriyel çalışmalar için anlamsal tabanlı arama, öneri, kişiselleştirme, reklam gibi pek çok yeni nesil uygulamaların gerçekleştirilebilmesine olanak sağlaması, makine öğrenmesi ve doğal dil işleme çalışmalarının kapsamını genişletiyor olması nedeniyle bilgi graflarının popülaritesi son yıllarda giderek artmıştır [7,8].

Web ölçekli bilgi tabanları ve bilgi grafları, başta arama motoru teknik alt yapısı olmak üzere, soru cevaplama sistemleri, kişiselleştirilmiş öneri sistemleri, makine öğrenimi, doğal dil işleme vd, çok çeşitli bilgi merkezli bilişsel uygulamalar için kritik öneme sahip konuma gelmiştir [8].

KG'ları, bilgilerin kolay bir şekilde organize edilmesini, yapılandırılmasını ve anlaşılmasını mümkün kılması, sezgisel ve özlü bir soyutlama sağlaması nedeni ile basit kavramlardan karmaşık fikirlere ve teorilere kadar geniş bir bilgi yelpazesini temsil etmek için kullanılabilmektedir [8,9].

KG'ları, gerçek dünya olgularını ve anlamsal ilişkilerini, düğümlerden ve kenarlardan oluşan, düğümlerin varlığı ve düğümler arasındaki kenarın ise varlıklar arasındaki anlamsal ilişkinin temsil edildiği yönlendirilmiş bir graf olarak tanımlanmaktadır. KG'larının oluşturulması; bilgi çıkarımı ve doğal dil işleme teknolojileri kullanılarak, bilginin varlıkların ve ilişkilerin tespiti, bilginin temsili, birleştirilmesi, tutarlılığının ve kalite parametrelerinin sağlanmasını içeren çok çalışmayı bir araya getirmektedir [9].

Doğal dilde yer alan metinlerden bilgilerin çıkarımı, yapay zekâ ve doğal dil işleme yöntemleriyle uzun süredir devam eden bir araştırma alanıdır [10].

Web üzerinde bulunan yapılandırılmış veri miktarı geçtiğimiz yıllarda önemli ölçüde artmış olsa da yapılandırılmış ve yapılandırılmamış verilerin kapsamı arasında hala önemli bir boşluk bulunmaktadır. Çeşitli derlemlerin yapısını otomatik olarak çıkarma veya geliştirme yöntemleri anlamsal ağ (Semantic Web - SW) disiplininde önemini korumaktadır [10].

Tim Bernars Lee'nin anlamsal ağ (The Semantic Web) [11] vizyonuna uyumlu olarak, makineleri dünyadaki varlıklar ve bunlar arasındaki ilişkiler hakkında kapsamlı bilgilerle donatmak amacı ile Web üzerindeki heterojen veri kaynaklarından, verilerin toplanması,

bilgilerin çıkarımı ve bilgi graflarının otomatik olarak oluşturulmasına yönelik 2004 yılından günümüze kadar literatürde tamamına yakını İngilizce dilinde olmak üzere, çok sayıda araştırma yapılmıştır [13].

Web üzerindeki yapılandırılmamış heterojen veri kaynaklarında yer alan veriler üzerinden milyonlarca varlık ve varlıklar arasındaki milyonlarca ilişkiyi içeren ConceptNet [14,15], KnowItAll [16], TextRunner [17], Yago [18], Freebase [19], DbPedia [20], Nell [21], Wikidata [22] gibi bilgi graflarının oluşturulmasına yönelik teknikler üzerine yapılan akademik araştırmalar; IBM Watsons [23], IBM Watsons Healtcare [24], Google Knowledge Vault [25], Yahoo KG [26] gibi domain spesifik endüstriyel bilgi tabanlarının ve bilgi graflarının otomatik olarak oluşturulmasına, oluşturulan bilgilerin makine okunabilir formatta gösterimine yönelik çalışmalar bulunmaktadır [27,28].

ConceptNet [14], "Open Mind Common Sense (OMCS)" projesi kapsamında anlamsal ağ ve doğal dil işleme çalışmalarında kullanılmak amacıyla 15.000'den fazla katılımcının katkısı ile 2004 yılında oluşturulmuş bir bilgi tabanıdır. 2017 yılında yer verilen çalışmada 83 dil desteği ile 8 milyon varlığa ait 17 milyon ilişkiyi barındırdığı belirtilmiştir [15].

KnowItAll [16] çalışması kapsamında, İngilizce dilinde Web üzerinden belirli etki alanlarına yönelik olarak bilgi çıkarımı gerçekleştirilmiştir.

TextRunner [17] çalışmasında Web üzerinden Open Information Extraction (OIE) modeli önerildiği ve söz konusu çalışma kapsamında 9.000.000 web sitesi üzerinden bilgi çıkarımı yapıldığı ve KnowITAll'dan daha başarılı sonuçlar elde edildiği belirtilmiştir

Yago [18] çalışmasında, Suchanek ve arkadaşları tarafından Wikipedia Infobox verileri üzerinden Wordnet ile taksonomi eşleştirmesi yapılarak; varlık ve ilişkilerin tespiti, bilgi çıkarımı gerçekleştirilmiştir. Araştırmacılar tarafından yapılan deneysel çalışmalarda doğruluk oranı %95 olarak belirtilmiştir.

Freebase [19] çalışması kapsamında, büyük bir kullanıcı kitlesi tarafından 4.000'den fazla varlık türüne ait 125.000.000 adet üçlü değeri üretilmiştir.

DBpedia [20] çalışması kapsamında, Wikipedia Infobox verileri üzerinden bilgi çıkarımlarının gerçekleştirildiği ve kaynak tanımlama çerçevelerinin (Resource Description Framework) araştırmacıların kullanımına sunulduğu; DBPedia'nın 2.600.000 varlığa ait 4.700.000 üçlü değerini içerdiği belirtilmiştir. Söz konusu çalışma kapsamında semantik

web teknolojileri kullanılarak Wikipedia verilerinin yapılandırılmış, anlaşılabilir ve kullanılabilir bir formata dönüştürülmesinin amaçlandığı görülmektedir.

NELL [21]; çalışması, Web üzerinden çevrim içi bilgi çıkarımını gerçekleştirmek amacı ile oluşturulmuştur. Araştırmacılar tarafından 541 tür ve 310 sınıfa ait 5.805.102 adet triple üretildiği, DBPedia ile kıyaslandığında gerek içerik gerekse güvenilirlik anlamında eksikliklerin bulunduğu belirtilmiştir.

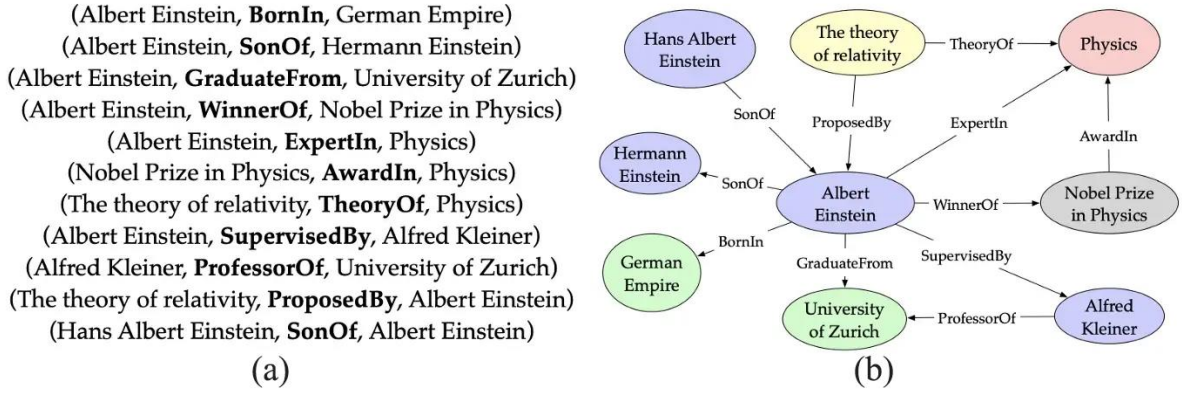
Wikidata [22], Wikipedia'nın milyonlarca varlığı içeren, herkesin katkıda bulunabileceği ve okuma ve düzenleme izinleriyle kullanabileceği açık kaynaklı, makine okunabilir formatta bilgi tabanı olup, metinsel bilgilerin yanı sıra resim, coğrafi koordinatlar ve sayısal verileri de içermektedir. 287 dilde yer alan, 30.000.000 Wikipedia sayfası kullanılarak oluşturulmuştur. DBpedia ve schema.org'dan farklı olarak bilginin doğruluğu açısından kaynağı ve bilginin geçerliliğinin başlangıç zamanını içeren zamansal geçerliliği gibi bilgiler içerdiği, RDF ve JavaScript Object Notation (JSON) formatında verilerin aktarılabildiği belirtilmiştir.

IBM Watsons [23] projesi kapsamında, yapısal olmayan verilerden RDF'ler çıkarılarak bilgi tabanı oluşturulmuş, IBM ürünlerinde kullanılmak üzere soru cevaplama sistemi alt yapısı geliştirilmiştir. IBM Watsons Sağlık Analitiği, Öğrenme ve Semantik ile Güçlendirme Merkezi (Health Empowerment by Analytics, Learning, and Semantics-HEALS) projesi kapsamında merkezi sağlık yönetimileri için analitik, bilgi tabanlı öğrenme ve semantik temelli sorgulama kullanan bilişsel araçlar geliştirilmesinin planlandığı, hastaların kronik hastalıklarını yönetmelerine ve önlem almalarına yardımcı olmanın amaçlandığı belirtilmiştir [24].

Google Knowledge Vault [25], Google tarafından geliştirilen ve Web üzerinde geniş bir olasılıksal bilgi tabanı oluşturan otomatik bir sistemdir. Diğer bilgi tabanlarının aksine, İnternet üzerindeki Hyper Text Markup Language (HTML) tabloları, yapılandırılmış işaretlemeler ve metin belgeleri gibi farklı kaynaklardan bilgi çıkarır. Sistem, grafik tabanlı öncül verileri, çıkarıcıları ve bilgi birleştirme bileşenlerinden oluşur. RDF üçlüleri ile temsil edilen bilgiler, olasılık tabanlı bir güven değeri ile ilişkilendirilir. Yüksek güven seviyesine sahip olan bilgiler, bilgi tabanında depolanarak, bilgilerin kalite ve doğruluğunun sağlanması amaçlanmıştır.

Yahoo Knowledge Graph [26], Yahoo tarafından geliştirilen gerçek dünyaya ait çeşitli nesne ve varlıkları kapsayan kapsamlı bir bilgi grafıdır. Nesneler hakkında kavramlar, ilişkiler, özellikler ve bağlantılar gibi temel bilgileri kapsar. Yahoo bilgi grafı, Freebase ve Wikipedia gibi açık kaynaklardan oluşturulmuştur.

Bir bilgi grafı, gerçek dünya olgularını ve anlamsal ilişkilerini, düğümlerden ve kenarlardan oluşan, düğümlerin varlığı ve düğümler arasındaki kenarın ise varlıklar arasındaki anlamsal ilişkiyi temsil ettiği yönlendirilmiş bir graf olarak tanımlanmakta olup, literatürde öne çıkan çalışmalarda bilgiler üçlüler şeklinde subject-predicate-object (S-P-O) formatında sunulmuştur. Örnek olarak "Albert Einstein ExpertIn Physics" ifadesinde, "Albert Einstein" konuyu, "ExpertIn" ilişkiyi ve "Physics" değeri temsil eder. Bu format, karmaşık ilişkileri ve bilgiyi yapılandırılmış bir şekilde temsil etmek için kullanılır. Şekil 1-a, üçlü örneğini gösterirken, Şekil 1-b bilgi grafındaki varlık ilişkilerinin temsilini göstermektedir [29].



Şekil 1.1. KG Örneği

Şekil 1.1 (a)'da Albert Einstein'a ait değerlerin yer aldığı üçlülere, Şekil 1.1 (b)'de ise söz konusu üçlülerin bilgi grafı formatında gösterimine yer verilmiştir.

Bu kapsamda yapılan çalışmalara, kullanılan yöntemlere ve özelliklerine Çizelge 1.1'de kapsamlı olarak yer verilmiştir [28]:

Çizelge 1.1. Literatürde öne çıkan çalışmalar

Organizasyon	Girdi Verişeri	Yöntem	Çıktı Formatı	Oluşturma Şekli	KG Türü	KG Boyutu	Yıl
Cycorp	Manuel iddialar (kurallar ve sağ duyu bilgisi)	Manuel	Bilgi Tabanları	Manuel	Kısmi açık kaynak	0,5 milyon kavram & 17000 ilişki & 7 milyon iddia	2006
CMU	Ontolojiler & Web sayfaları	Örüntü tanıma, POS, Lojistik regresyon	RDF Bilgi Tabanları	Otomatik	Açık Kaynak	123 kategori ve 55 ilişki içeren ontolojiler & 242.000'den fazla üçlü	2010
Metaweb	Metadata, yapılandırılmış veri	Manuel	Çapraz bağlantılı veriler	Manuel	Açık Kaynak	Yaklaşık olarak 44 milyon konu başlığı ve 2.4 milyar adet üçlü	2007
MPII- YAGO	Wikipediakategori sayfaları ve Wordnet bilgitabanı	Şablon & Kural Tabanlı	Zaman ve mekâna bağlı ontoloji	İş birliği	Açık Kaynak	10 milyon varlık ve bunlar ile ilgili 120 milyon üçlü	2007
UC Berkeley	Text özetleme	Yazılım araçları tabanlı (PERL/CGI-based)	Çok dilli sözlüksel veritabanı-asimetrik yönlü ilişki (XML & RDF)	İş birliği	Açık Kaynak	10,000 kelimeye ait anlamsal çıkarımlar ve 170.000 cümler	1998
DBPedia	Wikipedia sayfaları	Wiki ayrıştırıcı, kilgi kutuları, şablonlar	LOD Cloud & RDF	İş birliği	Açık Kaynak	13,7 milyon adet varlığa ait, 1,86 milyar üçlü (111 dil desteği)	2015
UW	Web sayfaları	POS ve Özellik tabanlı sınıflandırma	Domain bağımsız çıktı	Otomatik	Açık Kaynak	54.753 üçlü	2005
MIT	Viki sözlük	NLP teknikleri, graf tabanlı teknikler	Çoklu dil desteğine sahip bilgi grafı	Otomatik	Açık Kaynak	8 milyon düğüm ve 21 milyon ilişki	2004
MSRA	Olasalıksal modeller	NER, WSD	Olasalıklı ontoloji	Otomatik	-	2.7 milyon kavram	2012

Çizelge 1.1. (devam) Literatürde öne çıkan çalışmalar

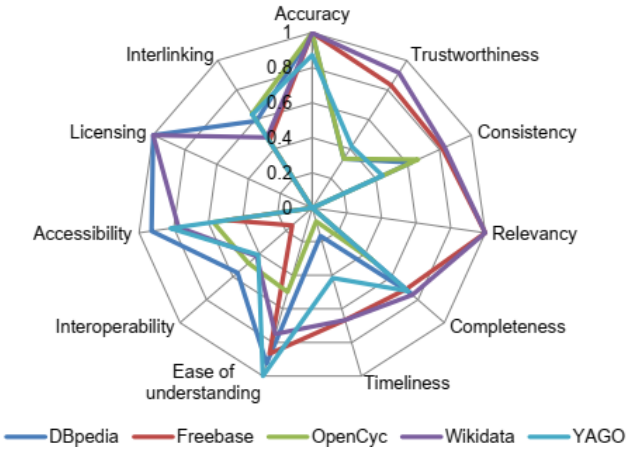
Organizasyon	Girdi Verileri	Yöntem	Çıktı Formatı	Oluşturma Şekli	KG Türü	KG Boyutu	Yıl
Google	Web Verileri	Sınıflandırma Yöntemleri	HTML Web tabloları	Otomatik	Açık Kaynak	1.6 milyar üçlü	2014
Google	Ontoloji, olgular, Freebase, Wikipedia	Kitlesel kaynaklı (CrowdSourcing)	Çoklu dil desteğine sahip anlamsal arama fonksiyonalitesi	Otomatik	Açık Kaynak	3.5 milyar üçlü, 500 Milyon nesne-varlık	2012
Facebook	Facebook bloglarından gönderi ve yorum verileri	NLP teknikleri ve şablon tabanlı yaklaşım	Semantik Arama Servisi	Otomatik	Açık Kaynak	50 Milyon birincil varlık, 500 Milyon üçlü	2013
Microsoft	Bing arama akışları	Öznitelik ve örüntü tabanlı	Yapısal veritabanı ve bilgi grafı	Otomatik	-	300 Milyon varlık, 800 Milyon ilişki içeren üçlüler	N/A
Prospera	Yarı yapısal text	Örüntü tabanlı	Web ölçekli RDF bilgi grafı	İş birliği	-	2 Milyon varlık	2011
DeepDive	Farklı veri kaynakları	Markov zincirleri, danışmanlı yöntemler	Factor Graphs	Otomatik	Açık Kaynak	-	2016
IBM	Yapısal ve yapısal olmayan veri	Watson Bilgi Keşfi servisleri	Müşteri odaklı bilgi grafı	Otomatik	-	100Milyondan fazla doküman, 5 milyar ilişki, 100 Milyon varlık	2019
IBM	Yapısal ve yapısal olmayan veri	Watson Bilgi Keşfi servisleri	Sağlık odaklı bilgi grafı	Otomatik	-	-	2022

Farber ve arkadaşları tarafından literatürde öne çıkan bilgi tabanları ve bilgi grafları farklı kalite parametreleri açısından incelenerek karşılaştırılmış, Çizelge 1.2'de DBPedia, Freebase, OpenCyc, Wikidata ve YAGO RDF KGs'nın istatistiksel karşılaştırmalı sonuçlarına yer verilmiştir [30].

Çizelge 1.2. Literatürde öne çıkan RDF KG'nın karşılaştırmalı sonuçları

	DBpedia	**Freebase**	**OpenCyc**	**Wikidata**	**YAGO**
Üçlü sayısı	411.885.960	3.124.791.156	2.412.520	748.530.833	1.001.461.792
Sınıf sayısı	736	53.092	116.822	302 280	569.751
İlişki sayısı	2.819	70.902	18.028	1.874	106
Tekil yüklem sayısı	60.231	784.977	165	4.839	88.736
Varlık sayısı	4.298.433	49.947.799	41.029	18.697.897	5.130.031
Örnek sayısı	20.764.283	115.880.761	242 383	142.213.806	12.291.250
Her sınıf için ortalama varlık sayısı	5.840	9.408	35	62	9
Tekil konu sayısı	31.391.413	125.144.313	261 097	142.278.154	331.806.927
Nesne konumundaki tekil değişkenlerin sayısı	83.284.634	189.466.866	423 432	101.745.685	17.438.196
Nesne konumunda benzersiz değişmezlerin sayısı	161.398.382	1.782.723.759	1.081.818	308.144.682	682.313.508

Yine söz konusu çalışmada KG'larının sağlaması gerektiği belirtilen kalite parametrelerine yer verilmiştir. Bu parametreler ışığında DBPedia, Freebase, OpenCyc, Wikidata ve Yago RDF bilgi graflarının veri kalitesi metrikleri açısından karşılaştırmalı sonuçları Şekil 1.2'de sunulmuştur [30].



Şekil 1.2. KGs'ın veri kalite parametreleri açısından karşılaştırmalı sonuçları

Şekil 1.2'de yer alan literatürde öne çıkan DBPedia, Freebase, OpenCYC, Wikidata ve YAGO bilgi tabanlarının ve bilgi graflarının doğruluk, tutarlılık, güvenilirlik, güncellik, erişilebilirlik, anlaşılabilirlik, birlikte çalışabilirlik gibi öne çıkan veri kalite parametreleri açısından karşılaştırmalarına yer verilmiştir.

Web ölçekli heterojen veri kaynaklarından verinin toplanması, doğrulanması ve sorgulanması çok sayıda fırsat ve zorluğu beraberinde getirmektedir [5,9]. Veri tabanı ve Web teknolojilerinin olgunlaşması, kullanıcıları bilgiyi büyük miktarlarda ve çeşitli formatlarda kamuya açık hale getirmeye teşvik etmiş ve çok kaynaklı veriler üzerinde büyük ölçekli arama ve karşılaştırmalı analizler yapılmasına olanak sağlamıştır. Web üzerinde yer alan veri hacminin büyüklüğü nedeniyle bilginin doğruluğunu ve güvenilirliğini sağlamak giderek zorlaşmıştır. Temel zorluklardan biri, Web üzerinden çıkarılan bilgilerin gürültülü, eski, yanlış, yanıltıcı ve dolayısıyla güvenilmez olabilmesidir. Bununla beraber, Web üzerinde yer alan çakışan bilgilerin, hatalı ve sahte içeriklerin çok fazla sayıda kaynağa kolayca yayılabilir olması, ulaşılan veri içerisinde neyin doğru neyin yanlış olduğunu ayırt edilmesini de zorlaştırmaktadır [31].

Bağlantılı veri, anlamsal web ve bilgi graflarının birleşik önemi, daha geniş dijital ekosistemde birbirleriyle nasıl etkileşim içinde oldukları ve birbirlerini nasıl tamamladıklarını değerlendirilmesi ile mümkün olacaktır.

Bağlantılı veri, anlamsal web ve bilgi grafları, veri kalitesi, veri doğrulama ve doğruluk keşfinin birlikte etkileşimi sağlam bir dijital çerçeve oluşturmaktadır. Bu çerçeve sadece bilginin zenginliğini ve erişilebilirliğini değil, aynı zamanda güvenilirliğini, doğruluğunu ve güvenilirliğini de sağlamayı amaçlamaktadır.

Veri kalitesi, veri doğrulama ve doğruluk keşfi, bilgi grafları mimarisinde temsil edilen Web ölçekli veri ve bilgilerin yüksek kalitede olmasını sağlamaktadır. Veri doğrulama, verilerin doğruluğunu ve tutarlılığını kontrol ederken, doğruluk keşfi bilgilerin gerçekliğini doğrulamaktadır.

Doğrulanmış ve veri kalite ölçütlerini sağlayan yüksek kaliteli veriler, gelişmiş karar alma süreçlerinin önünü açar. Gerek stratejilerin belirlenmesi ya da yönlendirilmesi gerekse bu verileri kullanan bir işletme, gerekse sağlık sonuçlarını tahmin eden bir yapay zekâ modeli veya bilgi arayan bir tüketici için alınan kararların daha bilinçli ve daha güvenilir olmasını mümkün kılmaktadır.

Web üzerinde heterojen veri kaynaklarından elde edilen veriler, zamanla güncelliğini yitirebilir, alakasız hale gelebilir ve hatta bozulabilir. Veri kalitesi ve doğrulamaya odaklanmak, verilerin taze, ilgili ve hatasız veya bozulmamış elde edilmesini sağlamaktadır.

İnternet üzerindeki bilgilerin doğruluğunun ve dinamiklerinin önemi, sadece akademi ve web endüstrisinden değil, aynı zamanda doğrudan uygulanması için devlet kurumları ve haber ajansları için de büyük ilgi uyandıran araştırmaların yapılmasına yol açmıştır [31].

Büyük verinin doğruluğu, genel olarak veri bilimi topluluğunun özel ilgisini çeken güncel bir konudur. Aynı zamanda "gerçeğin keşfi" sorunu, entelektüel ve teknik olarak, yapay zekâ ve veri yönetimi alanlarında çeşitli araştırma topluluklarından çok sayıda önceki çalışmaya ait yöntemleri de çekecek kadar popüler hale gelmiştir. Söz konusu çalışmaların bazı araştırmalarda, bilgi güvenilirliği, güven yönetimi, bilgi doğrulama, veri birleştirme ve bilgi birleştirme olarak yerini aldığı görülmektedir.

Web üzerinde yer alan varlık ve varlıklar arası ilişkilerin tekdüze yapıda olmaması, İnternet'in dinamik yapısı, veri ve veri kaynaklarının doğruluğunun sağlanması ve bilgi tabanı yapısında yer alan olguların ya da bilgilerin doğruluğunun sağlanması, güvenilir kaynakların güncelliğinin sağlanmasında verimli ve etkin yeni algoritmalar geliştirilmesine ihtiyaç bulunmaktadır. İstatistiksel analiz, makine öğrenimi yöntemleri ve olasılıksal modelleme gibi çeşitli yaklaşımlar, bilginin doğruluğunun değerlendirilmesinde kullanılmaktadır [31].

Güven değerinin hesaplanmasına yönelik olarak literatürde yer alan öne çıkan çalışmalar incelendiğinde, bilinen referans gerçek (ground truth) değeri üzerinden güven değeri hesaplamalarının yapıldığı, bu bağlamda Web ölçekli geniş kapsamlı bir bilgi tabanı oluşturmanın gerekliliği görülmektedir [31,32,33].

Bu tez kapsamında "İnternette Heterojen Veri Kaynaklarından Verinin Toplanması, Doğrulanması ve Sorgulanmasını" Türkçe dilinde gerçekleştirmek amacıyla özgün bir bilgi grafı modeli ve buna yönelik uygulama platformu geliştirilmiştir. Geliştirilen özgün model Türkçe diline yönelik olarak, genel amaçlı, uçtan uca, web ölçekli, doğal dil işleme, bilgi çıkarımı ve güven değeri hesaplamalarını içermektedir.

Doğal dil işleme ve bilgi çıkarımı teknikleri kullanılarak genel amaçlı bir bilgi tabanı oluşturulması, bilginin üçlülere indirgenmesi ve bilgi grafı mimarisinde gösteriminin

gerçekleştirilmesi amacıyla Türkçe konu başlıklarına sahip yaklaşık 530.000 Wikipedia sayfasından oluşan veri kümesi kullanılmıştır.

Wikipedia aşağıdaki özelliklerinden dolayı literatürde yaygın kullanılmaktadır [32]:

- Wikipedia'daki metinler öncelikle gerçeklere dayanmaktadır.
- Wikipedia, çeşitli alanlarda gerçek dünyaya ait varlıklar hakkında belgeler içeren geniş bir kapsama sahiptir.
- Wikipedia'daki makaleler, DBpedia, Freebase, Wikidata, YAGO vb. dahil olmak üzere çeşitli bilgi tabanlarında tanımladıkları varlıklara doğrudan bağlanabilir.
- Wikipedia'da varlıklardan bahsedilmesi genellikle o varlıkla ilgili makaleye bir bağlantı sağlar, böylece çeşitli bağlamlarda etiketlenmiş varlık bahsi örnekleri ve ilişkili bağlantı metni örneklerini sağlar.
- Wikipedia metinsel özelliklerin yanı sıra terim frekansları ve eş-oluşumları, çeşitli daha zengin özelliklere ulaşma imkânı sağlamaktadır.

Tez kapsamında,

- Genel amaçlı, uçtan uca, web ölçekli, doğal dil işleme, bilgi çıkarımı süreçlerini içerecek şekilde referans bilgi tabanı oluşturulması,
- Web üzerinden elde edilen verilerin güven değerinin karşılaştırmalı yöntemler ile hesaplanması,
- Web referans bilgi tabanında yer almayan bilgi ve kaynaklar ile karşılaşılması durumunda, kaynağın güven değerini sağlamaya yönelik olarak alan adının sınıflandırılması ve sınıf değeri dikkate alınarak kaynağa güven değeri atanması,
- Oluşturulan referans bilgi tabanının arama motorları ile entegrasyonu durumunda arama motorlarına özgün unutulma hakkı alt yapısının sağlanması,
- Oluşturulan referans bilgi tabanının ve alan adı sınıflandırması yönteminin arama motorları ile entegrasyonu durumunda, arama motorlarına özgün güvenli arama gösteriminin alt yapısının sağlanması, bu kapsamda alan adlarının yarı danışmanlı veriler üzerinden sınıflandırılması,

problemleri adreslenerek çözümler geliştirilmiştir.

Trpedia.org/search2 adresinden erişim sağlanabilen geliştirilen özgün model ve yazılım platformu, genel amaçlı, Web ölçekli, doğal dil işleme ve bilgi çıkarımı süreçlerini içeren Türkçe diline yönelik literatürdeki ilk bilgi grafıdır. Farklı problemleri çözümüne yönelik olarak bilgi grafı alanında birçok yenilikçi özelliği bünyesinde barındırmaktadır.

Oluşturulan modelin özgün mimarisi sayesinde farklı problemlere adaptasyonu olanaklı kılmakta olup; ML algoritmaları uygulanarak graf üzerinde topluluk keşfi ve düğüm sınıflandırmasının sağlanması gibi pek çok yenilikçi özelliği barındırmaktadır. Çalışmamızın literatüre katkısı:

i. Heterojen veri kaynaklarından Türkçe diline ait metinlerden bilgi çıkarımı süreçlerini gerçekleştirerek, bilgilerin makine okunabilir formatta bilgi tabanında saklanmasını ve graf formatında gösteriminin sağlanması,
ii. Genel amaçlı, Web ölçekli, doğal dil işleme ve bilgi çıkarımı süreçlerini içeren Türkçe diline yönelik literatürdeki ilk bilgi grafı olması,
iii. Oluşturulan modelin özgün mimarisi sayesinde farklı problemlere adaptasyonu olanaklı kılması,
iv. Makine öğrenimi (ML-Machine Learning) algoritmaları uygulanarak graf üzerinde topluluk keşfi ve düğüm sınıflandırmasının sağlanması gibi pek çok yenilikçi özelliğin gerçekleştirilmesi,
v. Wikipedia'da yer alan Türkçe konu başlıkları üzerinden karşılaştırıldığında geliştirilen özgün model ve uygulama platformu kapsamında yapılan üçlü çıkarımlarının, DBPedia ve Yago bilgi tabanları üzerinde yer alan triple çıkarımından çok daha fazla sonuç üretilmesi
vi. Önerilen graf tabanlı özgün çözüm mimarisinin arama motorları ile entegre edilmesi durumunda, arama motorlarında unutulma hakkının uygulanmasına yönelik etkili mekanizmaların geliştirilmesine katkıda bulunması ve nihayetinde bireylerin mahremiyet haklarını koruması,
vii. Önerilen graf tabanlı özgün çözüm mimarisinin telekom operatörleri ve arama motorları seviyesinde entegre edilmesi durumunda alan adlarının sınıflandırılması ve güvenli arama (safe search) özelliklerini abonelere başarılı bir şekilde sağlayabileceği,

şeklinde gerçekleşmiştir.

Tez çalışmasının ikinci bölümünde Web veri kaynakları, verinin toplanması, temsili ve sorgulanması için kullanılan yöntemlere yer verilmiştir. Üçüncü bölümde güven değerinin hesaplanmasına yönelik kullanılan yöntemlere, problemin çözümünü adreslemek amacı ile geliştirilen özgün modele yer verilmiştir. Dördüncü bölümde geliştirilen özgün modele ilişkin olarak deneysel çalışmalara, veri kümesi özelliklerine, güven değeri hesaplaması,

değerlendirme ölçütleri ve deneysel sonuçlara yer verilmiştir. Beşinci ve son bölümde ise sonuçlar değerlendirilmiş ve gelecekte yapılabilecek çalışmalar sunulmuştur.

2. WEB VERİ KAYNAKLARI ve VERİ YÖNETİMİ TEKNİKLERİ

Web sınırsız bir bilgi kaynağı olarak günlük yaşantımızda kullanılmaktadır. Ancak, Web'de bulunan veri kaynakları üzerinde yer alan verilerin homojen olmaması; farklı türde içerik, yapı ve formatların bir karışımı olarak çeşitlilik göstermesi nedeni ile heterojenliği, fırsat ve zorlukları beraberinde getirmektedir. Farklı formatta, farklı yapılarda, farklı kalitelerde, erişim ve güncelleme sıklıkları gibi özellikleri, entegrasyon, kalite, güvenlik ve gizlilik, ölçeklenebilirlik gibi konular zorlukları olarak karşımıza çıkmaktadır. Bununla birlikte kapsamlı analiz, gerçek zamanlı karar vermeyi mümkün kılması, farklı alanlarda inovasyonlara neden olması ise fırsatlar olarak değerlendirmek mümkündür. Bu bölümde Web veri kaynakları ile çalışmanın bileşenlerine ve zorluklarına yer verilmiştir.

2.1. Web Verileri

Bulunduğumuz bilgi çağında; İnternet ve Web, her biri benzersiz ve değerli veri kaynakları içermesi nedeniyle insanlar ve uygulamaların bilgiye erişiminde ilk başvuru noktası haline gelmiştir. Web bilgiyi toplamanın, yorumlamanın ve kullanmanın şeklini işletmeler, araştırmacılar ve bireyler için vazgeçilmez öneme sahip duruma getirmiştir. Web, eğitim kaynaklarından haber makalelerine, e-ticaret platformlarından sosyal medya ağlarına kadar; keşfedilip analiz edilmeyi bekleyen değerli verileri barındırmaktadır.

World Wide Web, doğuşundan bu yana bilgiye erişim şeklimizi dönüştürmüş olup, insanların bilgi paylaşımı yaparak, iletişim kurabileceği bir bilgi alanı olarak 1989 yılında Tim Bernars Lee tarafından tasarlanmıştır. 2002'den itibaren Web 2.0 ile birlikte kullanıcıların daha aktif katılımını kolaylaştırarak sadece tüketici olmaktan çıkıp, içerik yaratıcısı ve yayıncısı olmalarını sağlamıştır. Web 2.0 döneminin ilk yılları, Facebook, YouTube ve Twitter gibi günümüzün en önde gelen sosyal medya sitelerinden bazılarının piyasaya sürülmesiyle iletişim daha dinamik ve geniş kapsamlı hale gelmiştir. Diğer taraftan Web, küresel ekonomide itici bir güç olmuştur. Amazon ve Alibaba gibi e-ticaret platformları, işletmelerin küresel pazarlara kolay bir şekilde ulaşmalarını sağlamıştır. Uber ve Freelancer gibi Web platformları tarafından desteklenen serbest çalışan ekonomisi, milyonlarca insan için iş fırsatları yaratmış ve tüketicilere daha esnek ve çeşitli hizmetler sunmuştur [34].

Web tartışmasız faydalar sunmuş olsa da belgelerin Web'i ortaya çıkmasını ve başarılı olmasını sağlayan aynı prensipler son zamanlara kadar Semantik Web'e uygulanmamıştır. Web'de yönlendirilen iki HTML belge içinde yer alan nesneler arasındaki ilişki veyahut iki belge arasındaki ilişki, doğası gereği aralarındaki anlamsal ilişkiyi ifade edemediği sürece örtük kalacaktır [34,35].

Web 3.0 olarak ta ifade edilen Semantik Web, Web'in bir uzantısı olarak önerilmiştir. Semantik Web, Web'in bir sonraki büyük evrimini temsil etmekte olup, insanlar tarafından okunabilir Web verilerinin, makineler tarafından okunabilir ve anlaşılabilir olmasını amaçlamaktadır. Semantik Web, bilgiyi geniş çapta erişilebilir kılmak amacı ile yapılandırılmış içeriğin artan kullanılabilirliğinin daha yüksek otomasyon seviyelerine olanak tanıdığı bir Web vizyonunu amaçlamaktadır [34,35].

Semantik Web genellikle çeşitli standart ve teknolojilere atıfta bulunmak için kullanılırken, Semantik Web standartları kullanılarak yayınlanan veriler bağlantılı veri (linked data) veya veri ağı (Web of Data) olarak adlandırılmaktadır. Web veri kaynaklarına erişim sağlamak veri hacmi ve çeşitliliği, veri kalitesi ve güvenilirliği gibi beraberinde bazı zorlukları getirir. Zorluklarına rağmen Web, büyük bilgi rezervleri olarak; çeşitli alanlarda büyük potansiyele sahiptir [35].

Semantik Web, makinelerin Web üzerindeki yapılandırılmış içeriği otomatik olarak işleyeceği fikri üzerine kurulmuştur. Gürültülü veriler, güvenilmez yayıncılar ve kasıtlı doğru olmayan veri üretimleri ise bu fikrin işleyişinde önemli sorunları beraberinde getirmektedir.

Dünyanın en büyük heterojen bilgi ağı Web üzerinde yer alan veri ve bilgilerin kalitesi, doğruluğu ve eksiksiz olarak kullanılabilirliğini tanımlamak ve sağlanmasında temel bir faktör haline gelmiştir. Web'de bulunan yapılandırılmış veri miktarı geçtiğimiz yıllarda önemli ölçüde artmış olsa da Web üzerinde bulunan yapılandırılmış ve yapılandırılmamış verilerin kapsamı arasında hâlâ önemli bir boşluk bulunması nedeni ile çeşitli derlemlerin yapısını otomatik olarak çıkarma veya geliştirme yöntemleri önemini korumaktadır [34,35,36].

Öte yandan, arama motorları genel amaçlı bilgilere ulaşmak ya da Web sayfalarını bulmak için mükemmel olsa da Semantik Web vizyonu, Web'deki birden fazla kaynaktan bilgi çekmeyi gerektiren daha karmaşık sorgulara hitap etmektedir. Mevcut aramalar genellikle

hızlı bir şekilde çözülüyor gibi görünse de kullanıcılar, arama motorlarının değerli sonuçlar sunamayacağını bildikleri için şu anda daha karmaşık aramalar yapamamaktadır [37].

Semantik Web, mevcut Web'in makine tarafından okunabilirliğinin zayıf olduğunu varsaymaktadır. Semantik Web alanındaki çalışmalar yeterli olgunluğa ulaştığında; makine öğrenimi karmaşık sonuçlara ulaşabilmek ve makine okunabilir içeriklere ulaşmak mümkün olacaktır [37].

Web ölçekli heterojen veri kaynaklarından elde edilecek verilere yönelik yapılacak çalışma aynı zamanda büyük veri özelliğini taşıdığı değerlendirilmekte olup, büyük verinin genel karakteristik özelliklerine Çizelge 2.1'de yer verilmiştir [36].

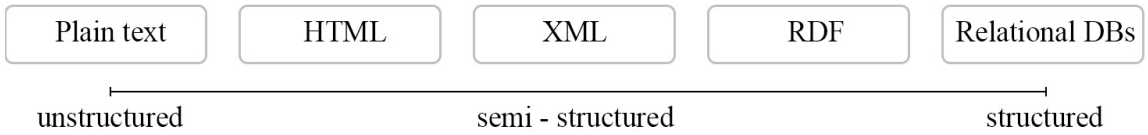
Çizelge 2.1. Büyük veri karakteristik özellikleri

3 Vs	Volume	Veri kapasitesi
	Velocity	Verilerin üretilme veya analiz edilme hızı
	Variety	Veri yapısı/ veri kaynaklarındaki farklılıklar (metin, görüntü, ses, coğrafi veri)
5 Vs	Veracity	Verilerin doğruluğu (belirsizlik içermemesi), özgünlüğü, kaynağı
	Validity	Seçilen veri setinin kullanım amacına uygunluğu ve doğruluğu
7 Vs	Volatility	Verilerin zamansal geçerliliği, güncelliği ve kullanılabilirliği
	Value	Verilerin karar almada kullanışlılığı ve uygunluğu ve bilgiye dönüştürme kapasitesi
10 Vs	Visualization	Veri gösterimi ve yöntemlerin anlaşılabilirliği
	Vulnerability	Veri işleme ile ilgili güvenlik ve gizlilik endişeleri
	Variability	Verilerin değişen anlamı, verilerdeki tutarsızlıklar, önyargılar, belirsizlikler ve verilerdeki gürültü

Heterojen Web veri kaynakları üzerinden yer alan veri türleri, yapısal olmayan (unstructured), yarı yapısal (semi-structured) ve yapısal (structured) olmak üzere genel olarak üç farklı formatta karşımıza çıkmaktadır.

Yapılandırılmamış veriler (plain text), yarı yapılandırılmış veriler (HTML, XML, RDF ve JSON formatında yer alan veriler), yapılandırılmış veriler ilişkisel veritabanlarında yer alan verilerdir. Farklı veri kaynakları üzerinde yer alan veriler üzerinden bilgilerin çıkarımının sağlanması, farklı bilgi çıkarma yöntemleri ve stratejilerini gerektirmektedir. Çıkarılan bilgiler çoğunlukla RDF gibi makineler tarafından okunabilir bir formatta temsil edilmektedir.

Web üzerinde yer alan veri kategorilerine yönelik olarak veri spektrumuna Şekil 2.1'de yer verilmiştir [36].



Şekil 2.1. Veri spektrumu

Web üzerinde yaygın olarak kullanılan veri kaynakları ve kullanım türlerine aşağıda kısaca yer verilmiştir [37]:

- Google, Bing ve Yahoo gibi arama motorları, İnternet üzerinde yer alan çeşitli kaynaklardan içeriklerin keşfedilmesini sağlayarak, bilgiye ulaşma imkânı sunmaktadır.
- Facebook, Twitter, Instagram ve LinkedIn gibi sosyal medya platformları, kullanıcılar tarafından oluşturulan içeriklerin yer aldığı zengin veri kaynaklarıdır. Bu platformlar, bireylerin ve toplulukların eğilimleri, görüşleri ve duyguları hakkında değerli iç görüler sunmaktadır.
- Amazon, eBay ve Alibaba gibi e-ticaret platformları, geniş ürün katalogları ve müşteri yorumları içermektedir. Bu durum, pazar araştırmaları ve tüketici davranışı analizi için değerli bilgiler sağlamaktadır.
- Kamu ve özel şirketlere ait Web siteleri, demografi, ekonomik göstergeler, farklı istatistikler gibi çeşitli konularda veriler sağlamaktadır.
- Forumlar ve bloglar genellikle uzmanların ve meraklıların değerli görüşlerini içermektedir.

Bununla birlikte bazı veri setleri ve web siteleri, Web ölçekli çalışmalarda kullanılmak üzere önemli kaynak işlevi görmektedir. Wikipedia dünyanın en popüler web sitelerinden birisi olarak, gerçek dünyaya ait varlıklar ve konu başlıklarını içermesi nedeniyle güvenilir bir bilgi kaynağı olarak kullanılmaktadır [32,38].

Wikipedia, açık bir şekilde düzenlenebilir içerik modeline dayanan çoklu dil desteğine sahip, web tabanlı, özgür içerikli bir ansiklopedi projesi olarak tanımlanmaktadır. 300'den fazla dil desteği ve yüksek kaliteli içeriği korumak için doğrulanabilirlik politikası da dahil olmak üzere çeşitli mekanizmalar içermektedir. Wikipedia üzerine yer alan konu başlıkları ve

içerikler, iş birliğine dayalı düzenlemeyi, güncel içeriğin oluşturulmasını ve sürdürülmesini mümkün kılmaktadır.

Wikipedia'nın varlık odaklı arama için önem arz etmesinin nedeni gerçek dünyaya ait varlıkları ve varlıklar arası ilişkileri içermesi, bilgi içeriklerini yarı yapılandırılmış olarak bilgi kutuları (infoboxes) aracılığı ile sunabilmesidir [38].

Web ölçekli veri toplanmasına yönelik çalışmalarda Wikipedia ile birlikte DBPedia, Freebase ve Wikidata gibi açık erişimli bilgi tabanları da kullanılmaktadır. Wikipedia zengin içerikler barındırmış olmasına rağmen makine okunabilir formatta yapılandırılmış bir bilgi modeli değildir. Bir bilgi tabanı, gerçek dünyaya ait varlıklar ve varlıklar arası iddialar kümesinden oluşmakta olup [38,39], bununla birlikte genel amaçlı bilgi tabanları Dbpedia, Wikidata varlıklar düzeyinde birbirine bağlanarak, yüz milyonlarca ilişkisel gerçek bilgiyi içeren bağlantılı Web'i oluşturmaktadır [40].

Web üzerinden verilerin elde edilmesi ve kullanımında karşılaşılan ana zorluklara kısaca yer verilmiştir [41-47]:

a) Heterojenlik: Heterojenlik, heterojen Web verisinin ve büyük verinin temel özelliklerinden birisi olup, çeşitlilik olarak ifade edilmektedir. Veri heterojenliğinin başlıca nedeni farklı veri kaynaklarından farklı biçimlerde verinin üretilmesidir. Heterojenlik, veri entegrasyonunda istenen değeri elde etmeyi zorlaştırmaktadır. Web siteleri, IoT veri kaynakları, sosyal ağ platformları başlıca heterojen veri kaynakları olarak karşımıza çıkmaktadır.
b) Verinin belirsizliği: Heterojen veri kaynaklarından elde edilen veriler doğası gereği, gürültü ve tutarsızlıkları içermesi nedeni ile belirsizlikleri beraberinde getirmektedir. Verinin hacmi, çeşitliliği ve hızı arttıkça, veriler üzerindeki doğal olarak belirsizliğin artmasına, söz konusu veriler üzerinden verilen kararlarda güvensizliğe yol açacaktır. Dolayısıyla web ölçekli veriler üzerinden bilginin keşfi ve analitiği için etkili analiz yöntemlerine ihtiyaç duyulmaktadır.
c) Ölçeklenebilirlik: Artan veri hacmi nedeni ile verinin ölçeklenebilirliği ile başa çıkmak önemli bir zorluk olarak karşımıza çıkmaktadır. Aynı zamanda, artan veri miktarı ile birlikte doğruluk, güven ve verimliliği birlikte sağlayabilen etkili analitik çözümler ve mimariler geliştirmek önem arz etmektedir.

d) İşlenebilirlik: Veri hacmi büyüdükçe, ham verileri depolamanın ekonomik olarak uygun olmadığı birçok durumda verinin özetlenmesi ve filtrelenmesi gerekmektedir.

Mevcut veri analitiği yöntemleri ile Web ölçeğinde heterojen veri kaynaklarının analizinde problem karmaşıklığı hızla artarken, hesaplama yeteneği ve çözüm kapasitesinin aynı hızda artmaması önemli bir zorluk olarak karşımıza çıkmaktadır.

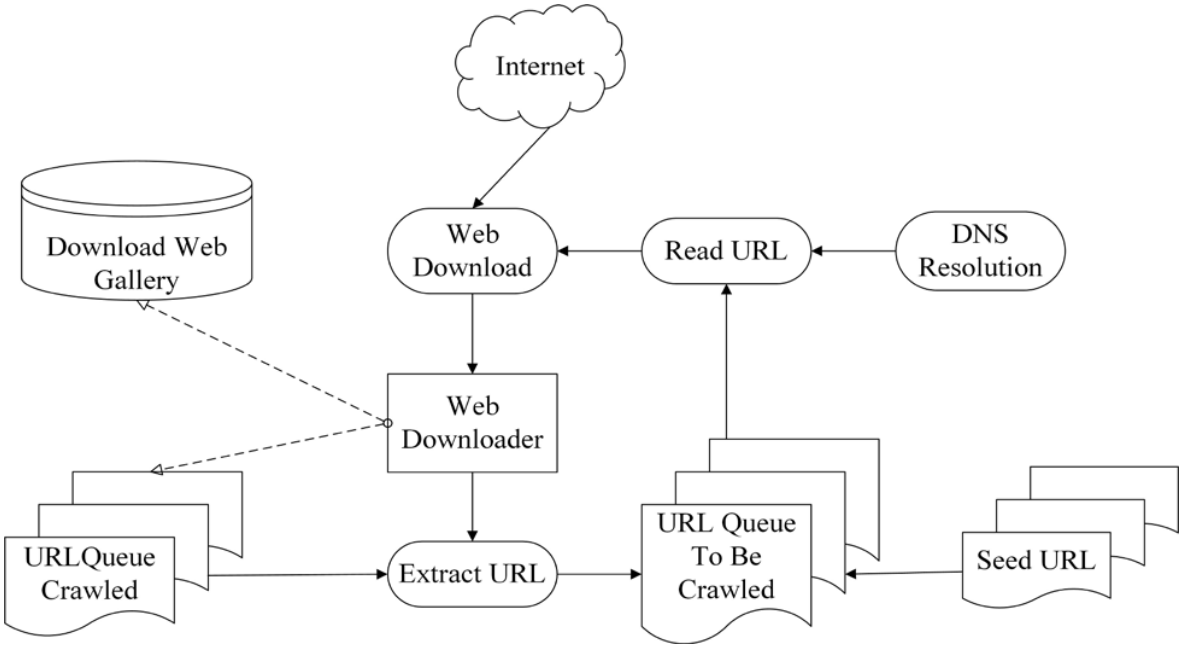
2.2. Web Veri Kaynaklarından Veri Toplanmasında Kullanılan Yöntemler

Web üzerindeki verinin toplanmasında başta Web tarayıcıları olmak üzere farklı yöntemler kullanılmaktadır. Web üzerinde heterojen veri kaynakları üzerinde yer alan verilerin analizi edilebilmesi için verilerin toplanması, temizlenmesi, saklanması ve analizi için gerekli alt yapının oluşturulması olarak karşımıza çıkan bu süreç esnek ve dinamik heterojen verileri destekleyebilmelidir. Bu süreç, iş zekâsından akademik araştırmalara kadar bir dizi uygulama için hayati öneme sahiptir.

Web üzerinden verilerin toplanmasında kullanılan öne çıkan yöntemlere aşağıda yer verilmiştir [48-51]:

- Web kazıma (Web Scraping): Belirli web sayfalarına erişmeyi ve kullanışlı verileri çıkarma işlemidir.
    - HTML ayrıştırma (HTML parsing),
    - DOM ayrıştırma (DOM parsing),
    - Düzenli ifadeler (Regular Expressions- Regex)
- Web Crawler: Genellikle arama motorları için geniş ölçekte indeksleme ve keşfetme amaçlamakta olup, web crawler türleri aşağıdaki gibidir:
    - Odaklanmış Web tarayıcı (Focused Web crawler),
    - Artırılmış tarayıcı (Incremental Web crawler)
    - Dağıtılmış mimaride çalışan tarayıcı (Distributed Web crawler)
    - Gizli tarayıcı (Hidden Web crawler)
- API (Uygulama Programlama Arayüzleri): Geliştiricilere bir platformdan doğrudan yapılandırılmış verileri almasına izin veren web hizmetleri olarak karşımıza çıkmaktadır.
- RSS/Atom Beslemeleri: Sıkça güncellenen içeriği yayınlamak için kullanılan web besleme formatları üzerinden verilerin elde edilmesi işlemidir.

Web tarayıcıları genellikle belirli kurallar veya amaçlara göre web üzerindeki veri kaynaklarında yer alan çeşitli bilgi ve veriyi elde eder ve depolar. Bu, arama motorlarının bilgi elde etmek için kullandığı kanallardan biridir. Web sayfalarına ait ilk URL'den itibaren tarama işlemine başlanır, tüm URL'ler taranmak üzere sıraya konur, belirli bir arama stratejisine göre örülen URL'lere erişilir, web sayfası içeriği veri tabanına kaydedilir. Web tarayıcı uygulamasının prensibi Şekil 2.2'de verilmiştir [49].



Şekil 2.2. Web tarayıcısı blok diyagramı.

Bununla birlikte Web üzerinden veri çıkarımı beraberinde zorlukları getirmekte olup, karşılaşılan zorluklara aşağıda yer verilmiştir [49,50,51]:

- İlk zorluk, insan çabasını olabildiğince azaltarak yüksek bir otomasyon derecesi sağlama konusundadır. Ancak, insan geri bildirimi, bir Web veri çıkarma sisteminin ulaştığı doğruluk seviyesini artırmada temel bir rol oynayabilir. Bu nedenle yüksek otomatik Web veri çıkarımı prosedürleri oluşturma ihtiyacı ile doğru performansı elde etme gereksinimi arasında makul bir denge bulmayı gerektirmektedir.
- Web veri çıkarımı tekniklerinin, sistemin ihtiyacına göre göreceli olarak kısa bir süre içinde büyük miktarda veriyi işleyebilmesi gerekmektedir.

- Sosyal ağ alanındaki uygulamalar ya da daha genel olarak kişisel verilerle ilgilenen uygulamalar üzerinden yapılacak veri çıkarımlarının kişisel verilerin korunmasına yönelik gizlilik dengesini koruması önem arz etmektedir.
- Makine Öğrenimi'ne dayanan yaklaşımlar genellikle manuel olarak etiketlenmiş büyük bir web sayfası eğitim setini gerektirir. Genel olarak, sayfaları etiketleme görevi zaman alıcı ve hataları barındırmaktadır.
- Web veri çıkarım aracı, zamanla değişebilen veri kaynağından rutin olarak veri çıkarması beklenmektedir. Bununla birlikte Web kaynaklarının yapısal değişime uğraması beraberinde zorlukları getirmektedir. Gerçek dünya senaryolarında, Web çıkarım aracının, ilgili Web kaynaklarının yapısal değişikliklerini algılama ve karşılama esnekliğinden yoksun kaldığında doğru bir şekilde çalışmayı durdurabilecek bu sistemlerin bakımının yapılmasına gerek duyulacaktır.

Diğer taraftan Web veri kaynaklarını kullanırken etik düşünceler son derece önemlidir. Bireylerin gizliliğine saygı göstermek, veri anonimliğini sağlamak ve veri koruma düzenlemelerine uyum sağlamak veri toplamasına yönelik bileşenin en önemli ve temel unsurlarıdır.

Sonuç olarak, çeşitli derlemlerin yapısını otomatik olarak çıkarma veya geliştirme yöntemleri Semantik Web bağlamında ve çalışmalarında temel bir konu olmuştur. Bu tür süreçler genellikle bilgi çıkarımı yöntemlerine dayanmakta olup, bu yöntemler de doğal dil işleme, makine öğrenimi alanlarındaki tekniklere dayanmaktadır.

Semantik Web ve bilgi çıkarımı tekniklerinin kombinasyonu iki perspektiften değerlendirilmesi mümkündür. Bir yandan bilgi çıkarımı teknikleri Semantik Web'i doldurmak için uygulanabilirken, diğer yandan Semantik Web teknikleri bilgi çıkarımı sürecini yönlendirmek için uygulanabilmektedir. Bazı durumlarda problemin çözümünü sağlamak amacıyla, mevcut bir Semantik Web ontolojisinin ya da bilgi tabanının bilgi çıkarımına rehberlik etmek için kullanıldığı ve bu ontolojinin ve/veya bilgi tabanının doldurulduğu durumlarda her iki bakış açısı birlikte ele alınmaktadır.

2.3. Web Ölçekli Verilerinin Saklanması ve Sorgulanması

Veri temsilleri sürekli olarak giderek daha karmaşık ve bazen birbirinden bağımsız talepleri karşılamaya yönelik olarak evrim geçirmiştir. Bu evrimin kilometre taşlarından biri, veriyi temsil etmek, yönetmek ve sorgulamak için ilişkisel veri tabanlarının ortaya çıkmasına yönelik olarak 1970'li yıllarında başlarında formüle edilen veritabanları, sürekli olarak incelenmiş ve aşamalı olarak geliştirilmiştir.

Bu süreç içerisinde SQL ve çeşitli optimize edilmiş ilişkisel veritabanları, kullanım öncesi iyi bir şema tasarımı yapılması durumunda verilerin temsili ve yönetiminde başarılı sonuçlar vermektedir.

Graflar, yapay zekâ alanında temel veri modellerinden biri olmuş, graf veri modelleri ve veri tabanları her zaman veri tabanı ve veri madenciliği topluluklarında ilgi çekici bir araştırma alanı olmuştur. Ancak Web'in etkileyici biçimde büyümesi ve bilgi graflarının Google tarafından lanse edilmesi ile birlikte, uzun süredir Semantik Web'de temel bir bileşen olan araştırmalarla ilgili doğrudan Web ölçekli akademik ve endüstriyel ticari uygulamalarda kullanımını yaygınlaştırmıştır. Bilgi graflarının farklı veri odaklı endüstrilerde geniş çapta kullanımı, bilginin kullanımına yönelik değişiklikleri de beraberinde getirmiştir.

Semantik Web çalışmaları ile ontolojiler, bağlantılı veriler ve bilgi graflarının birlikte çalışırlığının sağlanması, teknik değişim formatlarının oluşturulması gerekmekte olup; bu kapsamda veri paylaşımı, keşfi, entegrasyonu ve yeniden kullanımı için verimli yöntemler oluşturmak önem arz etmektedir [52].

Semantik Web aracılığı ile bilginin makineler tarafından anlaşılabilir kılınması, verinin anlamlı bir üst veri ile donatılmasıyla gerçekleştirilir. Semantik Web'de bu üst veri genellikle ontolojiler ya da en azından verinin anlamı üzerinde akıl yürütmeyi kabul eden mantık tabanlı bir semantiğe sahip biçimsel bir dil şeklindedir. Bilgi grafları, Web üzerindeki heterojen veri kaynakları üzerinde yer alan zengin veri içeriklerini, bağlantılı veriler ile anlamsal olarak eşleştirerek omurga görevini görmeye başlamıştır [53].

Bilgi grafları, Web'de birbirinden bağımsız yer alan verileri, bilgileri ve belgeleri birbirine bağlı olduğu bir alana bir araya getirmekte büyük katkı sağlamıştır. Semantik Web çalışmalarında, Web ölçeğinde çeşitli ve çoklu kaynaklardan gelen bağlantılı verileri, bilgi

grafları ile entegre edilmesi, zenginleştirilmesi, yorumlanabilmesi ve sorgulanması önem arz etmektedir [52,53].

2.3.1. Bağlantılı veri

Bağlantılı veri (Linked Data) alanındaki çalışmalar, kavramlar, varlıklar ve özellikler gibi farklı kaynaklar ve farklı veri türleri arasında bağlantılar oluşturmak için Web'in yaygın ve etkin kullanımını ifade etmektedir.

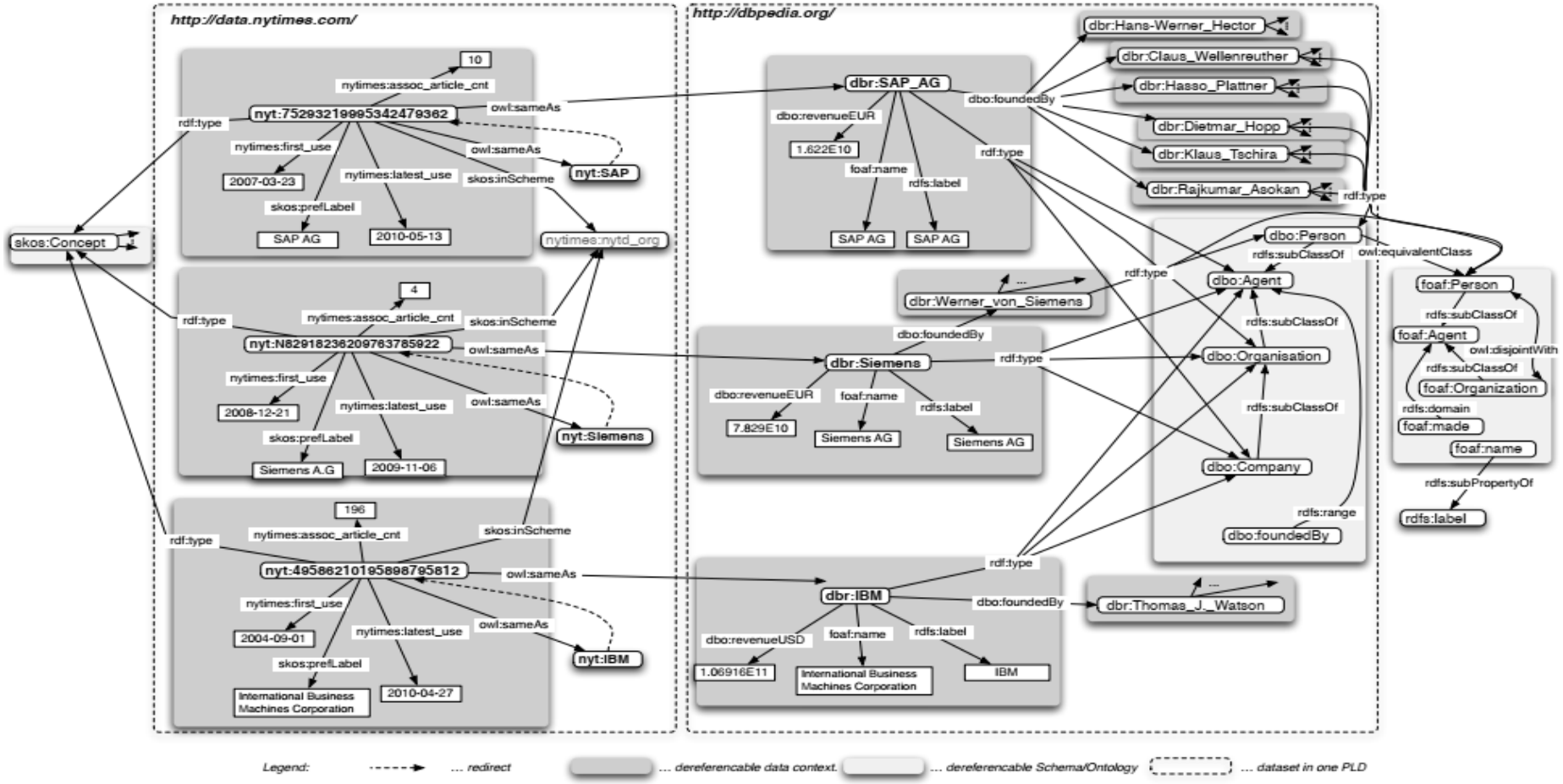
Farklı kaynaklar ve farklı veri türleri, başta Web üzerinde yer alan veri kaynakları olabileceği gibi, farklı coğrafi konumlarda bulunan aynı organizasyonun yönettiği farklı veri tabanları olabileceği gibi aynı çatı altındaki veri düzeyinde etkileşim sağlayamayan heterojen sistemler olabilecektir.

Web 2.0'ın temel yapı taşı birbirine bağlantı türü bilinmeyen HTML dökümanları iken, bağlantılı verinin çalışma mimarisi RDF belgelerine dayanmaktadır. Birbirlerine bağlı varlıklar ve nesnelerden oluşan bir Web ağı, Semantik Web ve Web of Data kavramını ortaya koymaktadır. Bağlantılı veri genel prensiplerine aşağıda yer verilmiştir [53,54,55,56,57]:

- Nesnelerin tanımlanmasında URI kullanımı: En geniş anlamda, bir URI, insanların Web üzerinde yayınlamak istedikleri varlıkları ve veri birimlerini adlandırma konusunda kısıtlamalar getirir ve bir standart oluşturur.
- Http protokolü ile URI'ların desteklenmesi: URI'ların aynı zamanda URL'ler olmasını sağlanması gerekmektedir.
- URI arama işlemlerinin RDF ve SPARQL ile standart hale getirilmesi: Resource Description Framework (RDF) ve SPARQL gibi kurulmuş standartları kullanarak bilgiye erişim sağlanması ve yayınlanması gerekmektedir. Resource Description Framework (RDF) ve SPARQL sorgu dili, Anlamsal Web ve Tanım Lojikleri topluluklarında uzun yıllar boyunca araştırmacılar tarafından geliştirilen önemli standartlardır.
- Diğer URI'lara bağlantılar eklenerek diğer URI'lara erişimin olanaklı kılınması gerekmektedir.

Şeklindedir.

Şekil 2.3'te DBPedia bağlantılı veri şemasına yer verilmiştir [56].



Şekil 2.3. DBPedia bağlantılı veri şeması

Diğer taraftan bağlantılı veriler ile çalışma beraberinde zorlukları getirmekte olup, söz konusu zorluklarına aşağıda yer verilmiştir [56]:

- Bağlantılı veri hacmi çok büyük olması nedeni ile son derece ölçeklenebilir, modüler muhakeme tekniklerine ihtiyaç duymaktadır.
- Bağlantılı veri olarak yayınlanan birçok RDF verisi, OWL yapısından farklılık göstermesi nedeni ile OWL altında doğrudan yorumlanamaz.
- Bağlantılı veriler tutarsızlık içerebilir, veri kümeleri yayıncılar tarafından çelişkili bilgiler sunabilir. Söz konusu tutarsızlıkların önlenmesi amacı ile farklı strateji ve yaklaşımların geliştirilmesini gerektirmektedir.
- Bağlantılı veri içeriklerinin değişimi zorunlulukları beraberinde getirecektir.
- Bağlantılı veri, RDFS ve OWL'den daha fazlasına ihtiyaç duymaktadır.

2.3.2. RDF

Kaynak tanımlama çerçevesi RDF (Resource Description Framework), yapılandırılmış bilgileri tanımlamak için kullanılan W3C tarafından tanımlanmış, Web üzerinde veri değişimini gerçekleştirmek amacı ile kullanılan standart bir dil modelidir.

RDF'nin amacı, uygulamaların orijinal anlamlarını koruyarak Web üzerinde veri alışverişi yapabilmelerini sağlamaktır. HTML ve XML'in aksine, artık asıl amaç belgeleri doğru bir şekilde görüntülemek değil daha ziyade bunların içerdiği bilgilerin daha fazla işlenmesine ve yeniden birleştirilmesine izin vermektir. RDF'nin gelişimi 1990'larda başlamış ve çeşitli öncül diller RDF'nin oluşturulma sürecini etkilemiştir. İlk resmi spesifikasyon 1999 yılında W3C tarafından yayınlanmıştır. Semantik Web vizyonu, 2004 yılında yeniden işlenmiş ve genişletilmiş bir RDF spesifikasyonunun yayınlanmasına neden olmuştur [55-57].

RDF, Web'in bağlantılı yapısını, nesneler arasındaki ilişkiyi ve bağlantının iki yönünü adlandırmak amacı ile URI'leri kullanacak şekilde genişletmesi, farklı uygulamalar arasında bilgilerin paylaşılmasını olanaklı kılmaktadır.

RDF, $t = (s, p, o)$- (subject; predicate, object) formunda üçlüleri kullanarak bir bir bilgi grafının düğüm ve kenarları arasındaki bağlantıyı temsil eden graf tabanlı bir veri modelidir. RDF ayrıca Web üzerinde bilgi graflarını yayınlamak için popüler bir formata dönüşmüştür. Bunların en büyüğü Bio2RDF, DBpedia, PubChemRDF, UniProt ve Wikidata gibi kaynaklar olup; milyarlarca üçlü içermektedir [58,59].

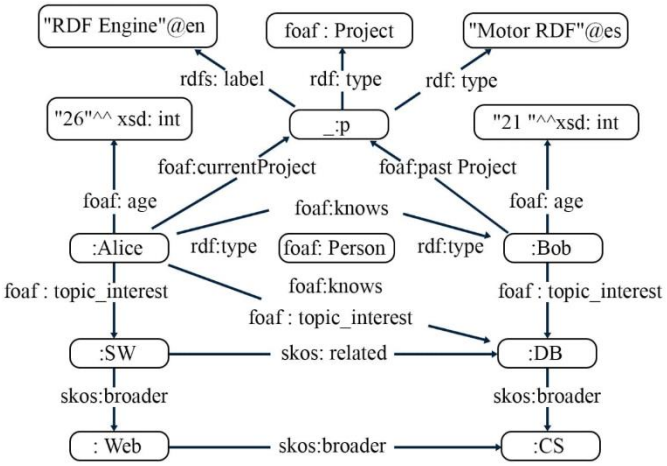
Bir RDF üçlüsü, bir dünya olgusuna ait gerçek (fact) bilgiya veya bir iddiayı (claim) temsil etmektedir. RDF üçlülerinin sabit slotlara (konu, özne, nesne) sahip olması etkileşim yeteneğini sağlayan genel bir ortak çerçeveyi oluşturmaktadır [57,58,59,60]. Bununla birlikte, artan veri miktarı ve RDF formatında temsili, büyük RDF graflarını sorgulamak için optimize edilmiş tekniklere olan ihtiyacı doğurmuştur. RDF depolama, dizinleme ve sorgu işleme olanağı sağlayan motorlara RDF stores olarak atıfta bulunduğu görülmektedir.

RDF verilerinin gösteriminde farklı seçenekler mevcuttur [58-60]:

- RDF/XML: RDF için en eski ve en köklü söz dizim temsillerinden birisi olup, erken dönemde standartlaştırılmıştır. RDF/XML'nin temel mantığı, XML araçlarının RDF formatına getirilmesi ve ayrıştırılması için kullanmaktır. RDF/XML, günümüzde hala en yaygın kullanılan RDF söz dizimlerinden biridir.
- Turtle, RDF'yi temsil etmek için kısa ve sezgisel bir sözdizimi hedefler ve yaygın olarak kullanılan RDF özellikleri için kısayollar içermektedir. Turtle'ın kullanımı, RDF/XML kadar popüler değildir. Bunun temel nedeni, Turtle'ın standartlaşmamış olmasıdır.
- N-Triples, RDF için basit bir söz dizimi ve Turtle'ın uygun bir alt kümesidir. Tüm N-Triples dosyaları geçerli Turtle'dır. N-Triples, Turtle kısayollarının tüm biçimlerini yasaklar ve tüm RDF terimlerinin tam biçimde yazılmasını gerektirir. Örneğin, URI'ler yerine tam olarak <> ayraçları kullanılarak yazılmalıdır ve ön ekler kullanılmaz. Üçlüler tam biçimde yazılmalı ve yeni satırlarla sınırlanmalıdır. Bu nedenle, N-Triples'da her bir bireysel satır, üçlünün o satırdaki ayrıntılarını belirlemek için gerekli olan tüm bilgileri belirtir ve belgenin geri kalanından bağımsız olarak üçlünün çözümlenmesi için kullanılabilir. Bu, N-Triples'ı akış uygulamaları ve hata toleranslı satır satır işleme için popüler kılar.
- JSON LD [56], veriyi büyük ölçüde RDF'ye eşdeğer bir formatta temsil etmek için JSON tabanlı bir söz dizimidir. JSON, birçok Web uygulaması tarafından bir serileştirme formatı olarak yaygın şekilde kullanıldığından, JSON LD, Web

geliştiricilerinin seçtikleri betik dillerinin mevcut JSON ayrıştırma mekanizmalarını kullanarak RDF'yi ayrıştırmasına izin verir. RDF grafları JavaScript nesneleri olarak işlemelerine olanak tanır.

Örnek RDF graf ve üçlü gösterimine Şekil 2.4'te yer verilmiştir [58].



Subject	Predicate	Object
:Alice	rdf: type	foaf : Person
:Alice	foaf: age	"26"^^xsd: int
:Alice	foaf: topic_interest	:DB
:Alice	foaf: topic_interest	:SW
:Alice	foaf : knows	: Bob
:Alice	foaf : current Project	-:p
:Bob	rdf: type	foaf : Person
:Bob	foaf: age	"21"^^xsd: int
:Bob	foaf : topic_interest	:DB
:Bob	foaf : knows	: Alice
:Bob	foaf:past Project	-:p
-: P	rdf: type	foaf : Project
-: P	rdfs: label	"RDF Engine"@en
-: P	rdfs: label	"Motor RDF" les
: SW	skos: broader	:Web
: SW	skos: related	:DB
:Web	skos: broader	:CS
:DB	skos: broader	:CS

Şekil 2.4. RDF graf ve üçlü gösterimi.

Şekil 2.4'ü incelediğimizde, şeklin sol tarafında varlıklar ve varlıklar arası iişkileri gösteren RDF graf gösterimin yer aldığı, şeklin sağ tarafında ise soldaki grafı oluşturan üçlülerin yer aldığı görülmektedir.

Bununla birlikte sorgu sonuçlarının tüm formatları makine tarafından okunabilir nitelikte olması nedeni ile insanlar için anlaşılması zordur. Bu nedenle, sonuçların görselleştirilmesi, popüler araştırma konularından biridir. Tarayıcıları kullanarak yapılan görselleştirmeler; IsaViz, RDF Gravity, DBpedia Mobile, Fenfire ve OpenLink Data Explorer gibi en yaygın yaklaşımlardır. Bu iki yaklaşımın kendi artıları ve eksileri bulunmaktadır. Metin tabanlı sorgu sonucu formatları, bilgi grafiğinin detaylı analizini sağlarken, grafiksel olarak görselleştirilmesi, daha büyük bir resmi görmeyi ve ilgili bilgiyi keşfetmeyi sağlamaktadır. Bilgi grafları, varlıklar ve varlıklar arasındaki ilişkilerin grafiksel olarak temsili ile bilgiye erişimi ve anomalilerin tespitini olanaklı kılmıştır [58].

Verimli veri depolama, indeksleme ve birleştirme işlemleri RDF depoları (ve dolayısıyla SPARQL motorları) için kilit öneme sahiptir. RDF depolarının taşıması gereken temel özelliklere aşağıda yer verilmiştir [58-60]:

- Depolama: Farklı motorlar RDF verilerini farklı yapılar (tablolar, grafikler, vb.), kodlamalar (tamsayı kimlikleri, dize sıkıştırma, vb.) ve ortamlar (ana bellek, disk, vb.) kullanarak depolar. Hangi depolamanın kullanılacağı verilerin ölçeğine, desteklenen sorgu özelliklerinin türlerine vb. bağlı olabilir.
- İndeksleme. Dizinler, hızlı arama ve sorgu yürütme için RDF depolarında kullanılır. Farklı indeksleme türleri verinin türüne bağlı olarak önemli olabilir.
- Join işlemleri*:* Sorguları değerlendirmenin temelinde, birleştirme işlemlerini işlemek için verimli yöntemler yatmaktadır. Geleneksel ikili birleştirmelerin yanı sıra, son yıllarda çok yönlü ve en kötü durum optimum birleştirmelerin yanı sıra GPU tabanlı birleştirme işleme gibi yeni teknikler ortaya çıkmıştır. Birleştirmelerin değerlendirme sırasını optimize etmek de verimli işleme sağlamak için önemli olabilir.

2.3.3. SPARQL

RDF için çeşitli sorgu dilleri önerilmiş olsa da SPARQL protokolü, RDF sorguları için standart haline gelmiştir. SPARQL'in ilk sürümü 2008 yılında standartlaştırılmış, SPARQL 1.1 ise 2013 yılında W3C tarafından yayınlanmıştır. SPARQL, yalnızca birleştirmeleri değil aynı zamanda daha geniş ilişkisel cebirin varyantlarını destekleyen etkileyici bir dil olarak karşımıza çıkmaktadır.

Bununla birlikte Web üzerinde uç nokta adı verilen bilgi tabanları üzerinden SPARQL sorgu hizmeti ortaya çıkmış, gün içerisinde milyonlarca sorguyu olanaklı kılmıştır. SPARQL protokolü üzerinden RDF verilerinin sorgularının depolanmasını, dizinlenmesini ve işlenmesini destekleyen RDF depoları, aynı zamanda SPARQL motorları olarak adlandırılmaktadır [58].

SPARQL, bilgi graflarının standart sorgu dili olarak geniş bir şekilde kullanılmakta olup hemen hemen tüm büyük ölçekli bilgi grafları, SPARQL sorgu desteği sağlamaktadır. SPARQL sorgu sonucu JSON, JSON-LD, XML, RDF/XML, RDF/N3, CSV, TSV ve

HTML vb pek çok formatta veri çıktısı sunabilmektedir. SPARQL, genişletilmiş ilişkisel cebirin çeşitli varyasyonlarını (projection, selection, union, difference vb.) destekleyen ifade gücü yüksek bir dildir. Web üzerinde, günde milyonlarca sorgu almakta olan endpoint olarak adlandırılan yüzlerce SPARQL sorgu hizmeti ortaya çıkmıştır. RDF üzerinde SPARQL sorgularını depolama, dizinleme ve işleme yeteneği olan motorlara SPARQL engine ve RDF store ismi verilmektedir [58-60].

Yüksek düzeyde bir SPARQL sorgusu beş ana bölümden oluşmaktadır [60,61]:

- Verimli veri depolama, indeksleme ve birleştirme işlemi, RDF depoları için (ve dolayısıyla SPARQL motorları için) temel öneme sahiptir.
- Depolama. Farklı motorlar, RDF verilerini farklı yapılar (tablolar, grafikler, vb.), kodlamalar (tam sayılar, tanımlayıcılar, dize sıkıştırma, vb.) ve ortamlar (ana bellek, disk, vb.) kullanarak depolar. Hangi depolamanın kullanılacağı, veri ölçeğine, desteklenen sorgu özelliklerine vb. bağlı olabilir.
- İndeksleme. İndeksler, hızlı aramalar ve sorgu yürütme için RDF depolarında kullanılır. Farklı indeks türleri, değişen zaman-mekân alışverişleri ile farklı işlemleri desteklemesi beklenmektedir.
- Birleştirme İşlemi. Sorguların değerlendirilmesinin temelinde, birleştirmeleri işleme yönelik verimli yöntemler bulunur.
- Sorgu İşlemi. SPARQL, birleştirmelerin ötesinde etkili bir şekilde desteklenmesi gereken çeşitli sorgu özellikleri içeren ifadeci bir dildir; bu özellikler filtre ifadeleri, isteğe bağlı ifadeler vd. sorguları desteklemesi beklenmektedir.

Yerel depolar yönetimi kolay olmakla birlikte, tek bir makinenin kaynakları ölçeklenebilirliği sınırlamaktadır. Bu nedenle, tipik olarak paylaşımsız makinelerden oluşan kümeler üzerinde çalışan çeşitli türlerde dağıtılmış RDF depoları önerilmektedir.

$t = (\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o})$ RDF terimlerini ve değişkenlerini göstermek üzere; $(\boldsymbol{s}, p, o)$, $(s, \boldsymbol{p}, o)$, $(s, p, \boldsymbol{o})$, $(\boldsymbol{s}, \boldsymbol{p}, o)$, $(\boldsymbol{s}, p, \boldsymbol{o})$, $(s, \boldsymbol{p}, \boldsymbol{o})$, $(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o})$, (s, p, o) şeklinde olmak üzere RDF verilerinin genel olarak 2^3 farklı kombinasyonda indekslenmesi mümkündür. Genellikle bir ilişkinin yalnızca birincil anahtarının varsayılan olarak indeksleneceği ve diğer indekslerin manuel olarak belirtilmesi gereken ilişkisel veritabanlarının aksine, çoğu RDF deposu varsayılan olarak sekiz olası üçlü modelin tümünü kapsayan eksiksiz bir indekse sahip olmayı amaçlar. Ancak, seçilen depolama türüne bağlı olarak bu her zaman mümkün olmayabilir.

RDF depoları, RDF verilerini tek bir makinede yöneten yerel depolar (tek düğümlü depolar olarak da adlandırılır), RDF verilerini birden fazla makineye bölen dağıtılmış depolar olmak üzere iki kategoride yapılandırıldığı görülmektedir.

Çizelge 2.2'de lokalde mimaride çalışan RDF depoları ve karşılaştırmalı özelliklerine yer verilmiştir [58].

Söz konusu karşılaştırmalı çizelgelerde RDF verilerinin saklanması, indekslenmesi, birleştirme operasyonları ve sorgu dili özellikleri yönleri ile RDF depolarının ele alındığı görülmektedir.

Çizelge 2.2. Lokal RDF depoları ve karşılaştırmalı özellikleri

Depolama (Storage): T = Üçlü Tablo, Q = Dörtlü Tablo, V = Dikey Bölümleme, P = Özellik tablosu, G = Grafik tabanlı, E = Matris/Tensör tabanlı, M = Çeşitli

İndeksleme (Indexing): T = Üçlü, Q = Dörtlü, E = Varlık, P = Mülk, N = Yol/Navigasyon, J = Birleştirme, S = Yapısal, M = Çeşitli

Birleştirme (Join): P = İkili, M = Çok yönlü, W = En kötü durum optimal, L = Doğrusal cebir

Sorgu (Query) P: R = İlişkisel, N = Yollar/Navigasyonel, Q = Sorgu yeniden yazma

Engine	Year	Storage							Indexing								Join P.				Query P.		
		T	Q	V	P	G	E	M	T	Q	E	P	N	J	S	M	P	M	W	L	R	N	Q
Redland	2001	✓							✓														
Jena	2002	✓														✓	✓						✓
RDF4J	2002	✓		✓												✓	✓						✓
Rssdb	2002	✓														✓	✓						✓
3store	2003	✓														✓	✓						✓
AllegroGraph	2003		✓							✓							✓				✓	II	
Jena2	2003	✓			II											✓	✓	✓					✓
Corese	2004	✓							✓								✓				✓		
JenaTDB	2004	✓	✓						✓	✓							✓				✓	II	
RStar	2004	✓														✓	✓						✓
BRAHMS	2005	✓							✓													II	
GraphDB	2005	✓	✓						✓	✓							✓				✓	II	
Mulgara	2005		✓							✓							✓				✓	II	
RAP	2005	✓							✓							✓	✓				✓		✓
RDF Match	2005		✓							✓	✓			✓		✓	✓	✓			✓		
YARS	2005		✓							✓							✓				✓		
Arc	2006	✓	✓						✓	✓							✓				✓		✓
RDF Broker	2006															✓	✓	✓			✓		
Virtuoso	2006	✓	✓													✓	✓				✓	II	
GRIN	2007					✓									✓		✓						
SW-Store	2007			✓																			
Blazegraph	2008	✓	✓						✓	✓							✓	✓			✓	II	
Hexastore	2008					✓			✓								✓				✓		
Rdf-3X	2008	✓							✓								✓	✓			✓		
BitMat	2009						✓		✓		✓									✓			
Dogma	2009					✓					✓				✓			✓					
LuposDate	2009	✓							✓								✓						
Parliament	2009	✓	✓						✓	✓													
RDF Join	2009	✓							✓					✓			✓	✓			✓		✓
SystemH	2009					✓			✓				✓				✓				✓		
HPRD	2010	✓	✓						✓	✓		✓					✓	✓					
Stardog	2010		✓							✓							✓				✓	II	
StrixDB	2010	✓							✓														✓
dipLODocus	2011					✓			✓		✓						✓				✓		
gStore	2011					✓			✓		✓							✓					
SpiderStore	2011					✓					✓												
SAINT-DB	2012					✓									✓		✓						
Strabon	2012			✓				✓	✓							✓	✓				✓		
BrightstarDB	2013	✓	✓						✓	✓							✓				✓	II	

Engine	Year	Storage							Indexing								Join P.			Query P.		
		T	Q	V	P	G	E	M	T	Q	E	P	N	J	S	M	P	M	W	L	R	N	Q
DB2RDF	2013										✓					✓	✓				✓		✓
OntoQuad	2013		✓							✓							✓	✓			✓		
OSQP	2013					✓									✓		✓				✓		
Triplebit	2013						✓		✓								✓	✓					
R3F	2014	✓				✓			✓			✓	✓				✓						
RQ-RDF-3X	2014		✓							✓							✓				✓		
SQBC	2014					✓					✓							✓					
WaterFow	2014	✓							✓								✓				✓		
GraSS	2015					✓			✓		✓							✓					
k^2-triples	2015			✓					✓								✓						
RDFCSA	2015	✓							✓								✓						
RDFox	2015	✓							✓								✓				✓	II	
TurboHOM++	2015					✓					✓							✓					
ClioPatria	2016		✓							✓											✓	II	✓
LevelGraph	2016	✓							✓								✓	✓					
RIQ	2016		✓			✓			✓						✓			✓					
axonDB	2017	✓							✓			✓					✓	✓					
HTStore	2017					✓					✓												
Ontop	2017							✓								✓					✓		✓
Quadstore	2017		✓							✓						✓	✓				✓	II	
AMBER	2018					✓												✓					
TripleID-Q	2018	✓					✓													✓	✓		
Jena-LTJ	2019	✓							✓										II				
MAGiQ	2019						✓													✓			
BMatrix	2020						✓		✓								✓						
Tentris	2020						✓		✓										II	✓	✓		
Ring	2021	✓							✓										✓				

Çizelge 2.3.'te ise dağıtık mimaride çalışan RDF depoları ve karşılaştırmalı özelliklerine yer verilmiştir. Söz konusu çizelge incelendiğinde dağıtık mimaride çalışan RDF depolarının depolama, indeksleme, birleştirme, sorgu işlemleri, bölümleme ve verilerin saklandığı mimari/teknoloji yönü ile ürünlerin karşılaştırmalı olarak ele alındığı görülmektedir [58].

Çizelge 2.3. Dağıtık mimaride çalışan rdf depoları ve karşılaştırmalı özellikleri

Depolama (Storage): T = Üçlü Tablo, Q = Dörtlü Tablo, V = Dikey Bölümleme, P = Özellik tablosu, G = Grafik tabanlı, E = Matris/Tensör tabanlı, M = Çeşitli,

İndeksleme (Indexing): T = Üçlü, Q = Dörtlü, E = Varlık, P = Mülk, N = Yol/Navigasyon, J = Birleştirme, S = Yapısal, M = Çeşitli

Birleştirme (Join) : P = İkili, M = Çok yönlü, W = En kötü durum optimal, L = Doğrusal cebir

Sorgu (Query) P: R = İlişkisel, N = Yollar/Navigasyonel, Q = Sorgu yeniden yazma

Bölümleme (Partitioning): S = Deyim (Üçlü / Dörtlü) tabanlı, G = Grafik tabanlı, Q = Sorgu tabanlı, R = Çoğaltma

Engine	Year	Storage							Indexing								JoinP.				Query P.			Partitioning				
		T	Q	V	P	G	E	M	T	Q	E	P	N	J	S	M	P	M	W	L	R	N	Q	S	G	Q	R	Store
YARS2 [94]	2007		✓							✓							✓				✓			✓				Custom
Clustered TDB [173]	2008	✓							✓								✓							✓				*Jena TDB*
Virtuoso EE [69]	2008	✓	✓							✓						✓	✓				✓	✓	✓	✓			✓	Custom
4store [91]	2009		✓							✓							✓				✓			✓				Custom
Blazegraph	2009	✓	✓						✓	✓							✓	✓			✓	✓		✓			✓	Custom
SHARD [195]	2009					✓											✓	✓			✓			✓				HDFS
Allegrograph	2010		✓						✓								✓				✓	✓		✓			✓	Custom
GraphDB [125,33]	2010	✓	✓						✓	✓							✓				✓	✓					✓	Custom
AnzoGraph	2011		✓														✓				✓	✓		✓				Custom
CumulusRDF [131]	2011	✓	✓						✓	✓						✓	✓				✓			✓				Cassandra
H-RDF-3X [104]	2011	✓							✓								✓	✓			✓			✓			✓	*RDF-3X*
PigSPARQL [202]	2011	✓															✓	✓			✓		✓	✓				HDFS
Rapid+ [193]	2011			✓													✓	✓			✓		✓	✓				HDFS
AMADA [17]	2012	✓							✓								✓				✓			✓				SimpleDB
H2RDF(+) [178]	2012	✓							✓							✓	✓							✓				HBase
Jena-HBase [121]	2012	✓		✓					✓							✓	✓				✓			✓				HBase
Rya [189]	2012	✓							✓							✓	✓				✓			✓				Accumulo
Sedge [254]	2012					✓											✓								✓	✓	✓	Pregel
chameleon-db [11]	2013					✓			✓					✓			✓				✓				✓	✓		Custom

Engine	Year	Storage							Indexing								JoinP.				Query P.				Partitioning				
		T	Q	V	P	G	E	M	T	Q	E	P	N	J	S	M	P	M	W	L	R	N	Q	S	G	Q	R	Store
D-SPARQ	2013	✓							✓							✓	✓	✓						✓			✓	MongoDB
EAGRE	2013	✓							✓		✓						✓	✓			✓				✓			HDFS
MR-RDF]	2013	✓							✓								✓	✓			✓				✓			*RDF-3X*
SHAPE	2013	✓							✓								✓	✓						✓				*RDF-3X*
Trinity.RDF	2013	✓							✓								✓	✓			✓				✓			Trinity
TripleRush	2013	✓							✓								✓							✓				Signal/Collect
WARP	2013	✓							✓								✓	✓								✓	✓	*RDF-3X*
AKZ14	2014	✓							✓							✓	✓	✓			✓		✓	✓	✓			HBase
Partout	2014	✓							✓								✓				✓					✓		*RDF-3X*
P-LUPOSDATE	2014	✓							✓							✓	✓				✓		✓	✓				HBase
RDF-3X-MPI	2014	✓							✓								✓							✓	✓		✓	*RDF-3X*
Sempala	2014				✓											✓	✓	✓			✓			✓				HDFS
SemStore	2014			✓		✓			✓								✓	✓							✓			*TripleBit*
SparkRDF]	2014			✓					✓					✓			✓				✓			✓				HDFS/Spark
TriAD	2014	✓							✓						✓		✓	✓							✓			Custom
CK15	2015			✓					✓					✓			✓							✓		✓		Custom
CliqueSquare	2015	✓															✓	✓						✓			✓	HDFS
DREAM	2015	✓							✓								✓				✓						✓	*RDF-3X*
AdPart	2016	✓							✓								✓							✓			✓	Custom
DiploCloud	2016					✓			✓		✓						✓				✓				✓	✓		Custom
Dydra	2016		✓							✓							✓				✓						✓	Custom
gStore-D	2016					✓			✓		✓							✓			✓				✓			gStore
Halyard [213]	2016	✓	✓						✓	✓							✓				✓			✓			✓	HBase
JARS [192]	2016	✓							✓								✓	✓						✓				Custom
S2RDF [204]	2016			✓												✓	✓				✓		✓		✓			HDFS
S2X [201]	2016	✓															✓				✓		✓	✓				Spark
SPARQLGX [76]	2016			✓													✓				✓			✓				HDFS
Wukong [211]	2016	✓							✓							✓	✓											DrTM-KV

Engine	Year	Storage							Indexing								JoinP.				Query P.			Partitioning				
		T	Q	V	P	G	E	M	T	Q	E	P	N	J	S	M	P	M	W	L	R	N	Q	S	G	Q	R	Store
CM-Well	2017	✓				✓					✓					✓	✓				✓	✓		✓				Cassandra/Elas
Koral	2017	✓							✓								✓							✓	✓			Custom
MarkLogic	2017	✓							✓								✓				✓	✓		✓				Custom
SANSA	2017	✓		✓													✓						✓	✓				HDFS
Spartex	2017					✓			✓									✓						✓				GSP/Custom
Stylus	2017					✓			✓		✓	✓						✓			✓			✓			✓	Trinity
Neptune	2018		✓							✓							✓				✓	✓					✓	Custom
PRoST	2018			✓	✓											✓	✓	✓					✓	✓				HDFS
RDFox-D	2018	✓							✓								✓								✓			*RDFox*
WORQ	2018			✓										✓			✓	✓										Spark
Wukong+G	2018	✓				✓			✓								✓								✓			*Wukong*
Akutan	2019	✓							✓								✓				✓	✓		✓			✓	RocksDB
DiStRDF	2019	✓			✓											✓	✓				✓			✓				HDFS
gStore-D2	2019					✓								✓			✓	✓							✓	✓		Custom
Leon	2019	✓							✓			✓					✓	✓							✓			Custom
SPT+VP	2019			✓	✓												✓	✓			✓		✓					Spark
StarMR	2019					✓											✓	✓						✓				HDFS
DISE	2020	✓					✓		✓								✓						✓	✓				Spark
DP2RPQ	2020					✓																✓		✓				Spark
Triag	2020					✓								✓			✓	✓					✓		✓			Spark
WISE	2020	✓							✓			✓					✓	✓					✓		✓	✓		Leon
gSmart	2021						✓		✓									✓		✓				✓		✓		Custom

2.4. Web Ölçekli Veri Çalışmalarında Veri ve Bilgi Kalitesi

İçinde bulunduğumuz büyük veri analitiği yaşam döngüsü içerisinde, veri kalitesi ve güvenilirliğinin önemi önemli ölçüde artmıştır. Hacim, hız ve çeşitliliğin özellikleri sıklıkla büyük verinin temel yönlerini tanımlamak için kullanılırken, verilerin doğruluğunu ifade eden, dördüncü "V" olan doğruluk (veracity) kavramı ortaya çıkmış ve önemi artmıştır.

Veri tabanı ve Web teknolojilerinin olgunluğu, kullanıcıları büyük miktarlarda ve çeşitli formatlarda bilgiyi kamuya açmaya teşvik etmiştir. Söz konusu süreç çoklu kaynak verileri üzerinde büyük ölçekli aramalar ve karşılaştırmalı analizler yapma ihtiyacını beraberinde getirmiştir. Ancak heterojen veri yapılarından farklı tür veri içerikleri üzerinden yapılacak veri entegrasyonları; ön yargılı, gürültülü, güncelliğini yitirmiş, yanlış, yanıltıcı ve dolayısıyla güvenilmez verilerin sorunlarını da beraberinde getirmektedir.

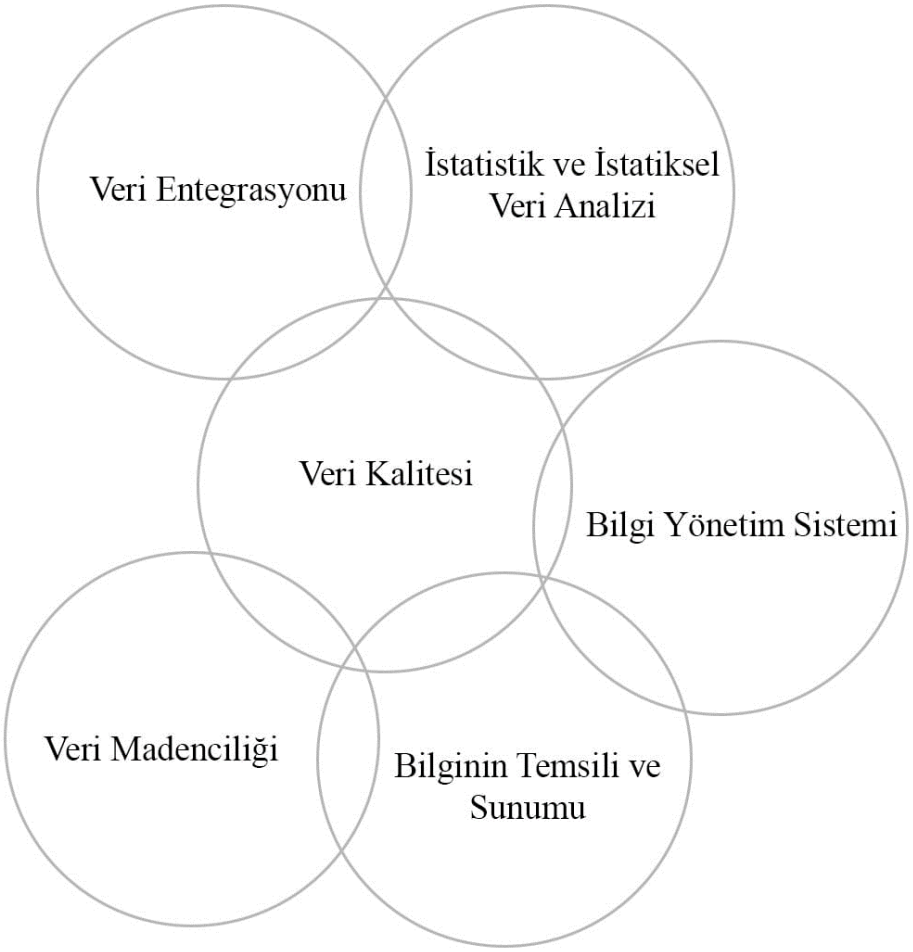
Kopyalama, yazım hataları, aykırı değerler, çelişkiler, tutarsızlıklar, eksik veriler gibi veri kalitesi sorunları, bilgi sistemlerinde yaygın olarak görülmektedir. Düşük kaliteli bilgilere dayalı kararların önlenmesi amacı ile teorik ve pragmatik yaklaşımlara ihtiyaç duyulduğu görülmektedir.

Bununla birlikte geçtiğimiz son on yılda, veri kalitesini ölçmeye ve artırmaya yönelik çalışmalar giderek artan bir ilgi görmüştür. Bilgisayar bilimleri alanında akademik ve endüstriyel topluluğun ilgisini yoğunlaştırdığı önemli bir araştırma konusu hâline gelmiştir.

Veri doğrulama; verinin tutarlılık, doğruluk ve önceden belirlenmiş kalite parametrelerine ve standartlara uygunluk açısından değerlendirilmesi ve doğrulaması için yapılan süreçler bütünü olarak tanımlanmaktadır. Veri doğrulama işlemleri ile amaç, veri seti içindeki herhangi bir tutarsızlık veya anormalliği tespit etmek ve düzeltmek, böylece yanıltıcı sonuçların ve hatalı karar vermenin riskini azaltmaktır [64].

Her organizasyon ve bireyin web sitesi oluşturabiliyor olması, bilginin kontrolsüz ya da dezenformasyon gibi kötü niyetli bir amaçla Web'e klaylıkla yükleme imkânı ve yüklenen bilgilerin geniş kitlelere hızlı bir şekilde ve farklı formatlarda ulaşabiliyor olması Web üzerindeki veri doğrulama ve kalitesine yönelik en önemli sorun olarak karşımıza çıkmaktadır.

Web üzerindeki heterojen veri kaynaklarının, bilgiyi kaynaklarından elde edilebildiği anda kısa sürede yayınlaması, bilginin doğruluğu, güncelliği ve kaynaklarının güvenilirliği açısından kontrol edilmesi gerekliliği önemli zorluklardandır. Web siteleri üzerinde yer alan bilgilerin doğrulanması maliyetli ve uzun süreçler olup, şekil 2.5'te bilgilerin doğruluğunu sağlamaya yönelik araştırma alanları diyagramına yer verilmiştir [65].



Şekil 2.5.Bilgilerin doğruluğunu sağlamaya yönelik araştırma alanları.

Şekil 2.5 incelendiğinde veri kalitesinin sağlanmasına yönelik sürecin, farklı disiplinler arası alanları ilgilendirdiği görülmektedir.

ISO 2008 yılında ISO/IEC 25012:2008 standardını yayınlamıştır. Söz konusu standart kapsamında yapısal bir formatta saklanan veriler için genel bir veri kalitesi modeli sunulması amaçlanmıştır. ISO veri kalite parametrelerine Çizelge 2.4'te yer verilmiştir [65,66].

Çizelge 2.4. ISO veri kalite standartları parametreleri.

DQ Characteristic	Definition
Doğruluk	Kavram veya olayın belirli bir niteliğinin gerçek önemini doğru bir şekilde ifade edilirliği.
Tamamlanmışlık	Varlığa bağlı verilerin, beklenen tüm öznitelikler ve ilişkili varlık örnekleri ile temsil edilmesi.
Tutarlılık	Herhangi bir çelişki olmaksızın diğer verilerle tutarlı ve uyumluluğun sağlanması.
Güvenilebilirlik	Verilerin kullanıcılar tarafından doğru ve güvenilir olduğu kabul edilmesi.
Güncellik	Verilerin güncelliği.
Erişilebilirlik	Verilere özellikle destekleyici teknolojiye veya engelleri nedeniyle özel yapılandırmalara ihtiyaç duyan bireyler tarafından erişilebilirliğin sağlanması.
Uyumluluk	Verilerin, yürürlükte olan standartlara, sözleşmelere, yönetmeliklere ve veri kalitesiyle ilgili diğer kurallara uygunluğunun sağlanması.
Gizlilik	Verilerin yalnızca yetkili kullanıcılar tarafından erişilebilir ve yorumlanabilir olmasının sağlanması.
Verimlilik	Veriler işlenebilir ve uygun miktarda ve türde kaynaklar kullanılarak beklenen performans düzeylerini sağlayabilir olması.
Hassasiyet	Alakalı verilerin tespiti.
İzlenebilirlik	Verilere erişimi ve bunlarda yapılan değişiklikleri kaydeden bir denetim izi mevcut olması, veri faaliyetlerinin izlenebilirliğini ve hesap verebilirliğinin sağlanması.
Anlaşılabilirlik	Verilerin kolaylıkla anlaşılırlığının sağlanması.
Kullanılabilirlik	Verilerin yetkili kullanıcılar tarafından ihtiyaç duydukları her an erişilebilir ve kullanılabilir olmalarını sağlayacak şekilde tasarlanması
Taşınabilirlik	Verilerin, bütünlüğünden ödün vermeden mevcut kaliteleri korunarak kurulacak, değiştirilecek veya bir sistemden diğerine taşınacak şekilde tasarlanması
Geri Kazanabilirlik	Verilerin, öngörülemeyen koşullar veya kesintiler durumunda bile belirli bir operasyon ve kalite seviyesini korumak ve sürdürmek üzere tasarlanması.

Bağlantılı verilerin Web üzerinde bilgi paylaşımının standart bir yolu olarak artan yaygınlığı, kullanıcıların ve organizasyonların geçmişte mevcut olmayan çok büyük veri kümelerinden yapısal verileri kullanmalarına olanak tanımaktadır.

Bununla birlikte diğer veri türleri gibi, bağlantılı verilerde de tutarsızlık, hatalı, güncellik, eksiklik gibi kalite problemleri ile karşılaşılmaktadır. Bu nedenle, bağlantılı veri uygulamalarında kullanılan veri kümelerinin kalitesini değerlendirmek önem arz etmektedir.

Kalite değerlendirmesi, kullanıcıların veya uygulamaların verinin amacına uygun olup olmadığını sağlaması açısından önem arz etmektedir. Web ölçekli verinin doğrulaması ve kalitesine yönelik çalışmalarda veri kalite metriklerinin tanımlanması bir dizi benzersiz zorluğu beraberinde getirmektedir. Söz konusu zorluklar [65,67,68]:

- Bağlantılı veri, çok sayıda heterojen veri kaynağından gelen birbiriyle bağlantılı yayınlanmış verilerden oluşan Web ölçeğinde bir bilgi tabanını ifade etmektedir. Sağlanan bilginin kalitesi, veri sağlayıcının gerek kullanmış olduğu teknik yöntemlere gerekse dezenformasyon vd. niyetine bağlı olarak eksik, gürültülü ya da hatalı veri kümelerini barındırarak yayılmasına neden olacaktır.
- Bağlantılı veri paradigmasının giderek yaygınlaşması, tüketicilerin geçmişte mevcut olmayan büyük miktarda veriden tam olarak yararlanmasına olanak tanımaktadır. Bununla birlikte gerek kaynak gerekse veri türleri açısından heterojenliğin ve veri miktarının artması değerlendirme sürecini zorlaştırmaktadır.
- Bağlantılı veriler tarafından kullanılan veri kümeleri, genellikle orijinal veri kümesinin yaratıcıları tarafından beklenmeyen şekillerde üçüncü taraf uygulamalar tarafından kullanılabilir.
- Bağlantılı veri, heterojen veri kaynakları arasındaki verilerin birbirine bağlanması yoluyla veri entegrasyonu sağlamaktadır, bununla birlikte entegre verinin kalitesi, orijinal veri kaynaklarının kalitesi ile ilişkili olup, zorlukları barındırmaktadır.
- Bağlantılı veri kaynaklarındaki değişiklikler, gerçek dünyadaki değişiklikleri yansıtmalıdır; aksi takdirde veri hızla güncelliğini yitirebilir. Güncelliğini yitirmiş bilgi, veri doğruluk sorunlarını yansıtabilir ve geçersiz bilgi sunabilir.

Bağlantılı verinin kalitesi, dış veri kümesine bağlantılar aracılığıyla tutarlılık, veri temsili kalitesi veya örtülü bilgiye göre tutarlılık gibi bir dizi yeni yönü içerir. İlişkisel veriler için kapalı dünya varsayımı (Close World Assumptions-CWA) tutulması gereken olağan

varsayım iken bağlantılı verilerin birbirine bağlı doğası açık dünya varsayımını (Open World Assumptions-OWA) doğal varsayım haline getirmektedir. OWA'nın veri ve şemalar arasındaki uyumu tanımlama ve değerlendirme zorluğu üzerinde etkisi vardır.

İki örnek arasındaki bir ilişki, şema örneklerin ait olduğu kavramlar arasında böyle bir ilişkiyi modellemese bile geçerli olabilir. Tersine, farklı şemaların iki kavramı arasındaki bir ilişkinin veri örneklerinde temsil edilmediği için geçerli olmadığı sonucuna varamayız.

Bununla birlikte, genellikle bağlantılı veri literatüründe, CWA'nın tamlık ve tutarlılık gibi kalite boyutlarının tanımlanması ve değerlendirilmesi için geçerli olduğu dolaylı olarak varsayılmaktadır [70].

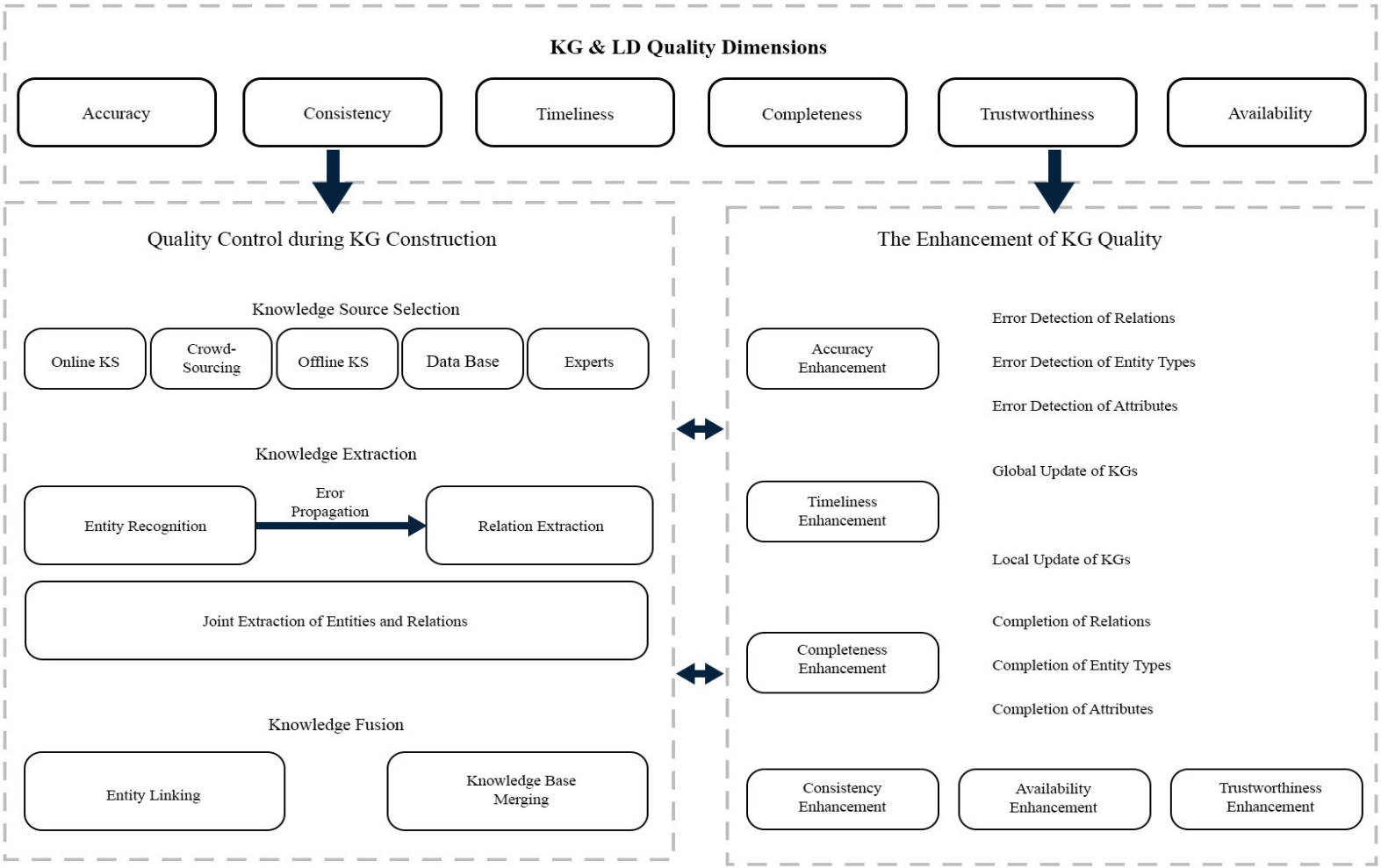
Web üzerinde yayınlanan bağlantılı verinin artan miktarıyla birlikte araştırma toplulukları verinin kalitesinin önemini fark etmiş ve bağlantılı veri kalitesini belirlemek ve değerlendirmeye yönelik araştırmaları devam ettirmektedir. Ancak, bu girişimler, ele aldıkları ve ölçtükleri kalite yönleri açısından farklılıklar göstermektedir. Bu, değerlendirme sonuçlarını karşılaştırmak ve benchmark yapmakta zorluklar yaşanmasına, ayrıca belirli kalite gereksinimlerine göre doğru veri kaynağını seçmekte zorlanmaya neden olmaktadır [71].

Şekil 2.6.'da bilgi grafları ve bağlantılı veriler için genel kalite parametrelerine yer verilmiştir [75]. Şekil 2.6'da görselde yer alan bileşenler değerlendirildiğinde;

- Bilgi grafları ve bağlantılı veri alanındaki çalışmaların kalite bileşenlerinin; doğruluk, tutarlılık tam değerlerin temsili, güncellik, güvenilirlik ve erişilebilirlik olduğu,
- Kalite bileşenlerini sağlamaya yönelik sürecin bilgi grafı inşa süreci ile başladığı; kaynak seçimi, bilgi çıkarımı ve bilgilerin birleşimi aşamaları ile sağlandığı

şeklinde olduğu görülmektedir.

Diğer taraftan Huaman tarafından yapılan çalışmada [76], literatürde veri, bilgi, bilgi graflarının kalite parametrelerinin incelenmesine yönelik öne çıkan araştırmaların kalite metrikleri açısından ele alındığı görülmekte olup, söz konusu değerlendirmelere Çizelge 2.5'te yer verilmiştir.



Şekil 2.6. KG ve LD kalite parametreleri

Çizelge 2.5. Kalite parametrelerinin karşılaştırılması

İlgili Çalışma	Kalite Boyutları	Detaylar
Ref. Wang ve Strong, 1996	4 kategoride gruplandırılmış 15 kalite boyutu bileşeni: Erişilebilirlik (2), Bağlamsal (5), İçsel (4), Temsil (4)	Erişilebilirlik, Doğruluk, Uygun miktar, İnanılırlık, Tamlık, Kısa gösterim, Tutarlı gösterim, Anlaşılma kolaylığı, Yorumlanabilirlik, Objektiflik, İlgililik, İtibar, Güvenlik, Güncellik, Katma değer
Naumann ve Rolker, 2000	3 kategoride gruplandırılmış 22 15 kalite boyutu bileşeni: nesne kriterleri (9), Süreç kriterleri (6), Konu kriterleri (7)	Doğruluk, Veri miktarı, Kullanılabilirlik, İnanılırlık, Tamlık, Kısa temsil, Tutarlı temsil, Müşteri desteği, Dokümantasyon, Yorumlanabilirlik, Gecikme, Nesnellik, Fiyat, İlgililik, Güvenilirlik, İtibar, Tepki süresi, Güvenlik, Zamanlılık, Anlaşılabilirlik, Katma değer, Doğrulanabilirlik
Bizer, 2007	4 kategoride gruplandırılmış 16 kalite boyutu bileşeni: Erişilebilirlik (2), Bağlamsal (7), İçsel (4), Temsili (3)	Erişilebilirlik, Doğruluk, Uygun miktar, İnanılırlık, Tamlık, Kısa temsil, Tutarlılık, Tutarlı temsil, Yorumlanabilirlik, Objektiflik, Saldırganlık, İlgililik, Tepki süresi, Güncellik, Anlaşılabilirlik, Doğrulanabilirlik
Zaveri ve ark. 2016	4 kategoride gruplandırılmış 18 kalite boyutu bileşeni: Erişilebilirlik (5), Bağlamsal (4), İçsel (5), Temsili (4)	Doğruluk, Kullanılabilirlik, Tamlık, Kısa gösterim, Kısalık, Tutarlılık, Birbirine Bağlanma, Birlikte Çalışabilirlik, Yorumlanabilirlik, Lisanslama, Performans, İlgililik, Güvenlik, Sözdizimsel geçerlilik, Güncellik, Güvenilirlik, Anlaşılabilirlik, Çok Yönlülük
Färber ve ark. 2018,	4 kategoride 11 kalite boyutu bileşeni: Erişilebilirlik (3), Bağlamsal (3), İçsel (3), Temsili (2)	Erişilebilirlik, Doğruluk, Tamlık, Tutarlılık, Ara Bağlantı, Birlikte Çalışabilirlik, Lisanslama, İlgililik, Güncellik, Güvenilirlik, Anlaşılabilirlik
Fensel ve ark. 2020	23 kalite boyutu bileşeni	Erişilebilirlik, Doğruluk, Uygun miktar, İnanılabilirlik, Tamlık, Özlü temsil, Tutarlı temsil, Maliyet etkinliği, Manipülasyon kolaylığı, İşlem kolaylığı, Anlama kolaylığı, Esneklik, Hatasızlık, Birlikte çalışabilirlik, Nesnellik, İlgililik, İtibar, Güvenlik, Zamanlılık, İzlenebilirlik, Anlaşılabilirlik, Katma değer, Çeşitlilik
Hogan ve ark. 2021	4 kategoride 10 kalite boyutu bileşeni Doğruluk (3), Kapsam (2), Tutarlılık (2), Kısa ve öz (3)	Doğruluk, Kapsam, Tutarlılık, Özlülük, Birlikte Çalışabilirlik, Lisanslama, Uygunluk, Zamanlılık, Güvenilirlik, Anlaşılabilirlik

Diğer taraftan Web üzerinden elde edilen bağlantılı verilerin kalitelerinin ölçülmesine yönelik literatürde “Linked Data Quality “, “Information Quality”, “KG Quality”, “KG Refinement” gibi farklı isimler farklı kalite ölçütleri ve parametreleri ile konunun ele alındığı [64-75] ve temel olarak doğruluk, tutarlılık, tamlık, güncellik, güvenilirlik ve erişilebilirlik boyutlarını içeren altı (6) ölçütün kabul gördüğü görülmektedir.

Literatürde öne çıkan çalışmalarda kullanılan genel kabul gören veri kalite parametreleri açıklamalarına ve değerlendirme ölçütlerine aşağıda yer verilmiştir [65,72,74,75,76].

2.4.1. Doğruluk

Bilgi graflarının doğruluğu, içerisinde yer alan üçlülerin doğruluk derecesini ifade etmektedir. Doğruluk (accuracy) diğer çalışmalarda sıklıkla tartışılmıştır [75-78]. Wang göre [75], KG doğruluğu,

"*bilginin ne ölçüde doğru, güvenilir ve hatasız olduğunun belgelenmesi*" olarak tanımlanmakta iken,

Mendes [78] tarafından, KG doğruluğu,

“bilginin doğruluğu ve bilgi kalitesi ile eşanlamlı” olarak ele almıştır.

Bununla birlikte doğruluk parametresinin değerlendirilmesi veri setinin kalitesi ve uzman değerlendirmesi ile dikkate almak doğru olacaktır. Bazı araştırmalarda varlıkların doğru olarak sınıflandırılmasının, ilişkilerin tespitinin ve belirsizliğin giderilmesinin de dikkate alındığı görülmektedir.

Doğruluk parametresine ilişkin değerlendirme ölçütleri aşağıda yer verilmiştir [72]:

- Ortalama otomatik doğrulama hataları (Average automatic validation errors)
- Ortalama kaynak doğrulama hataları (Average crowdsourcing validation errors)
- Ortalama veri tipi söz dizimi hataları (Average datatype syntax errors)
- Ortalama söz dizimi kuralları söz dizimi hataları (Average syntactic rules syntax errors)
- Ortalama RDF deseni hataları (Average RDF pattern errors)
- Ortalama yanlış yazılmış literaller (Average ill-typed literals)

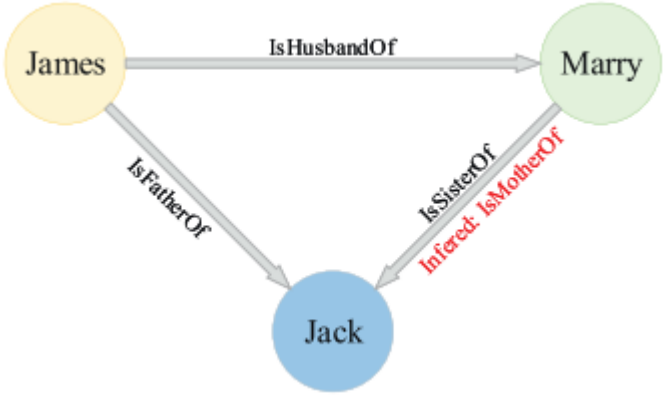
- Ortalama veri tipi uyumluluğu (Average datatype compatibility)
- Ortalama mesafe tabanlı aykırı değerler (Average distance-based outliers)
- Ortalama sapma tabanlı aykırı değerler (Average deviation-based outliers)
- Ortalama dağılım tabanlı aykırı değerler (Average distribution-based outliers)
- Ortalama üçlü doğruluğu (Average triple correctness)
- Ortalama kalabalık kaynak yanlış üçlüler (Average crowdsourced incorrect triples)
- Ortalama yanlış yazılmış literaller (Average misspelled literals)
- Ortalama yanlış etiketler (Average inaccurate labels)
- Ortalama doğru sınıflandırma (Average correct classification)
- Ortalama özellik uyumsuzluğu (Average property discordance)
- Ortalama geçersiz kurallar (Average invalid rules)
- Ortalama varlık yanlış eşleştirmeler (Average entity mismatch).

2.4.2. Tutarlılık

Bağlantılı veri ve bilgi graflarının tutarlılık (consistency) değerlendirmesi, KG'nın üzerinde yer alan bilgilerin kendi üzerinde yer alan diğer bilgiler (üçlüler) ile çelişmeme derecesi olarak tanımlanmaktadır [75,76,78]. Örnek olarak, bir KG aşağıdaki üçlüleri içeriyorsa:

(James Smith, IsFatherOf, Jack Smith), (James Smith, IsHusbandOf, Marry Smith), (Marry Smith, IsSisterOf, Jack Smith).

"Marry Smith, Jack Smith'in annesidir" bilgisi, üçüncü üçlü anlamsal olarak çelişir ve tutarsızlığı oluşturur. Şekil 2.7'de söz konusu tutarsızlığı içeren subgrafa yer verilmiştir [75].



Şekil 2.7. Tutarsızlığı gösteren subgraf örneği

KG’da üçlüler yayınlanmadan önce bilgi muhakemesi yoluyla tutarlılığı kontrol edilerek, yanlış bilginin keşfedilmesi ve düzeltmenin yapılması, tutarlılığın sağlanmasına ve KG doğruluğunun artmasına yol açacaktır. Tutarsızlık tespitine yönelik parametreler:

- Ortalama sabit IRI'lar, (Average stable IRIs)
- Ortalama tutarsız işlevsel bağımlılık konuları, (Average number of inconsistent functional dependence subjects)
- Ortalama ayrık sınıflar, (Average disjoint classes)
- Ortalama yanlış yerleşik sınıflar, (Average misplaced classes)
- Ortalama yanlış yerleşik özellikler, (Average misplaced properties)
- Ortalama yanlış kullanılmış veri tipi özellikleri, (Average misused datatype properties)
- Ortalama yanlış kullanılmış nesne özellikleri, (Average misused object properties)
- Ortalama kullanımdan kaldırılmış konular, (Average deprecated subjects)
- Ortalama kullanımdan kaldırılmış özellikler, (Average deprecated properties)
- Ortalama geçersiz ters işlevsel değerler, (Average invalid inverse functional values)
- Ortalama ontoloji ele geçirme, (Average ontology hijacking)
- Ortalama negatif bağımlı özellikler, (Average negative dependent properties)
- Ortalama geometrik ihlal, (Average geometric violation)
- Ortalama alan tutarlılığı, (Average domain consistency)
- Ortalama aralık tutarlılığı, (Average range consistency)
- Ortalama aksiyon ihlalleri, (Average axiom violations)
- Şema eksiksizliği, (Schema completeness)
- Özellik eksiksizliği, (Property completeness)
- Popülasyon eksiksizliği, (Population completeness)
- Örnek eşleme. (Instance interlinking)

şeklindedir.

2.4.3. Tam değerlerin temsili

Web ölçekli bir çalışmada KG tamlığı, bir KG'nın ilgilenilen bilgiyi ne ölçüde kapsadığını ifade etmekte olup, KG tamlığı da göreceli bir kalite değerlendirme boyutudur [72]. Genel olarak, bir KG, altta yatan uygulamaların bilgi ihtiyaçlarını karşılıyorsa iyi bir tamlık içerdiği kabul edilir.

Örnek olarak, Google arama motoru sonuçları çok sayıda ortak duyu içerir ve genel aramalar için yüksek tamlığa sahip olduğu ve yaygın olarak tercih edildiği değerlendirilmektedir. Bununla birlikte, Google bilgi grafı bazı özel profesyonel alanlar için eksik olabilir. KG doğruluğu ile karşılaştırıldığında KG tamlığı bilginin doğruluğu için katı gerekliliklere sahip değildir.

Varlıklar arasındaki ilişkinin açıkça ifade edilip edilmediğine ve varlık öz niteliklerinin tamamlanıp tamamlanmadığına odaklanır. KG tamlığını değerlendirmek için farklı yöntemler önerilmiştir. KG tamlığını değerlendirmek için ilgili göstergeler, puanlama fonksiyonları, sınıf, varlık sayısını kullanıldığı görülmektedir [72,75,76,81]

KG tamlığı, KG tabanlı uygulamalar için çok önemlidir. Bu nedenle, son çalışmalar KG'lerin tamlığını iyileştirmeye yönelik tekniklerin geliştirilmesine odaklanmıştır [75,76]. Tamlık değerini ölçmeye yönelik ölçütler aşağıda yer verilmiştir [72,75]:

- İç bağlantı derecesi (Interlinking degree),
- Kümeleme katsayısı (Clustering coefficient),
- Merkezilik (Centrality),
- Bağlantılı veri eşlemeleri (Linked Data mappings),
- İç bağlantılar (In-links),
- Ortalama bağlantı sayısı (Average sameAs linked),
- Ortalama boş düğüm sayısı (Average blank nodes),
- Varlık sayısı (Number of entities),
- Farklı özellik sayısı (Number distinct properties),
- Ortalama tanımsız sınıflar (Average undefined classes),
- Ortalama tanımsız özellikler (Average undefined properties),
- Ortalama tanımsız nesneler (Average undefined objects),

2.4.4. Güncellik

Güncellik (timeliness) parametresi, bilginin güncel olma derecesi olarak tanımlanır. Bununla birlikte güncellik parametresi de problemin uygulandığı alana özgün olarak değişkenlik arz edecektir. Örneğin, günlük olarak güncellenen bir KG'nın yüksek güncelliğe sahip olduğu düşünülebilirken COVID-19 araştırması için modellenen bir KG gerçek zamanlı güncellemelerle sürdürülmesi gerekmektedir. Yapılan araştırmalarda KG'nın güncelliği ile doğruluğu arasında bir ilişki olduğu belirtilmektedir [65,71,72,75,76].

2.4.5. Güvenilirlik

KG güvenilirliği (trustsworthiness), bir KG'nın objektiflik, otorite ve doğrulanabilirlik derecesi olup KG'da yer alan bir bilginin ne kadar inandırıcı kabul edildiğini açıklamaktadır. Güvenilirlik, bilginin otoritesi gibi öznel bir bakış açısına sahip kapsamlı bir boyuttur. Bu nedenle diğer boyutlarla güçlü bir ilişkisi vardır. Örneğin, bir KG'nın doğruluğu, güncelliği ve tutarlılığı gibi diğer ölçütlerdeki iyileşmeler KG'nın güvenilirliğinde de iyileşmeye yol açacaktır.

KG güvenilirliği bilgi kaynağının seçimi ve KG oluşturma yöntemlerinden de etkilenebilir. Genel olarak, KG'na ait kaynaklar güvenilir olduğunda ve oluşturma yöntemi üstün performansa sahip olduğunda, oluşturulan KG'nin güvenilirliği de yüksektir. KG'nın güvenilirliği bilgi kaynaklarının güvenilirliği ile değerlendirilebilir. Bir bilginin güvenilirliğinin değerlendirilebilmesi referans bilgi tabanlarına ve güvenilir bilgi kaynaklarına ihtiyaç bulunmaktadır.

Bu perspektiflere ek olarak, güvenilirlik doğrulanmamış bir KG'nın güvenilirliğine yönelik çalışmalar deneysel bir ölçüt haline getirir.

2.4.6. Erişilebilirlik

Bilgi grafı erişilebilirliği, bilginin kullanımına ne kadar uygun olduğunu belirtmek için kullanılmaktadır. Diğer bir deyişle, KG üzerinde yer alan bilgilere ne kadar kolay ve hızlı bir şekilde erişilebileceğini göstermektedir.

Ancak, KG erişilebilirliğini nicel olarak değerlendirmek zordur. Farklı açılardan erişilebilirlik parametresi dikkate alındığında:

- Makineler için, KG erişilebilirliği, bir KG'nın makine tarafından okunabilir bir formda olup olmadığını veya indirilebilir olup olmadığını,
- Kullanıcılar için KG erişilebilirliği, farklı dilleri destekleme yeteneği, teknik sorguların (örneğin SQL) kullanımı gibi özellikleri destekleyip desteklemediğini,
- Bilginin hızlı bir şekilde alınması için, KG erişilebilirliği, veri yanıt hızı, benzer özelliklere sahip varlıklar arasındaki bağlantı derecesi ve bilgi temsili biçimini

Şeklinde olduğu görülmektedir [67,72,75,76,78].

KG erişilebilirliğinin değerlendirme ve geliştirme yöntemleri diğer değerlendirme boyutlarından farklıdır. KG erişilebilirliğini artırmak için tasarım, mimari ve KG'nin diğer yönleriyle başlamak gereklidir.

Yukarıda yer verilen söz konusu 6 veri kalite parametresine ait özet açıklama ve formülüne Çizelge 2.6’da yer verilmiştir [75].

Çizelge 2.6. Veri kalite parametresine ait açıklama ve formülleri

Boyut	Tanım	Kalite Parametresi Ölçümde Kullanılan Hesaplama Yöntemi
Doğruluk	Sözdizimsel geçerlilik (sözdizimsel doğruluk), bir RDF belgesinin serileştirme formatının spesifikasyonuna uygunluk derecesini tanımlar. Anlamsal doğruluk, veri değerinin bir nesnenin doğru durumunu temsil edip etmediğiyle ilgilidir. Veriler hatasız, doğru, hatasız, güvenilir, hatalar kolayca tespit edilebilir, veri bütünlüğü sağlanmış, kesin olarak onaylanmıştır. Doğruluk boyutu, bir nesnenin veri değeri temsillerinin gerçek dünyadaki durumlarına yakınlığını tanımlar. Doğruluk, bir kaynaktaki doğru değerlerin sayısının kaynaktaki toplam değer sayısına oranıdır.	$1 - \langle \frac{hatalı\ örneklerin\ sayısı}{toplam\ örnek\ sayısı} * \frac{dengeli\ mesafe\ ölçümleri}{toplam\ örnek\ sayısı} \rangle$ $\frac{bir\ kaynaktan\ elde\ edilen\ toplam\ doğru\ değer\ sayısı}{bir\ kaynakta\ yer\ alan\ toplam\ değer\ sayısı}$
Tutarlılık	Tutarlılık, bir bilgi tabanının belirli bilgi temsili ve çıkarım mekanizmalarına göre (mantıksal/biçimsel) çelişkilerden arınmış olması anlamına gelir. Tutarlılık, verilerde çelişki olmamasını tanımlamaktadır.	$\frac{ayrık\ sınıfların\ üyesi\ olarak\ tanımlanan\ varlıkların\ sayısı}{veri\ kümesinde\ tanımlanan\ toplam\ varlık\ sayısı}$
Tamlık	Tamlık, gerekli tüm bilgilerin belirli bir veri setinde bulunma derecesini ifade eder. Verilerde yer alan bilgilerin genişliği, derinliği ve kapsamı	$Popülasyon\ tamlığı\ \frac{belirli\ bir\ özellik\ için\ temsil\ edilen\ değer\ sayısı}{belirli\ bir\ özellik\ için\ toplam\ değer\ sayısı}$ $bağlantı\ tamlığı\ \frac{veri\ kümesinde\ birbiriyle\ bağlantılı\ örneklerin\ sayısı}{veri\ kümesindeki\ toplam\ örnek\ sayısı}$
Güncellik	Güncellik, bir kaynaktaki verilerin ortalama yaşıdır. Güncellik, verilerin belirli bir göreve göre ne kadar güncel olduğunu ölçer.	$\frac{\sum_{kaynakta\ yer\ alan\ veri} verinin\ güncelliği}{kaynakta\ yer\ alan\ veri\ sayısı}$ $1 - \frac{gözlem\ sayısı - son\ değiştirilme\ zamanı}{ideal\ güncellik}$

Boyut	Tanım	Kalite Parametresi Ölçümde Kullanılan Hesaplama Yöntemi
Güvenilirlik	Güvenilirlik, bilginin doğru, gerçek ve inandırıcı olarak kabul edilme derecesini tanımlar. Güvenilirlik, verilerin kullanıcı tarafından doğru olarak kabul edilme derecesidir. Güvenilirlik, doğrudan deneyim veya başkalarının tavsiyeleri sonucunda bir varlığın veya veri kümesinin itibarını temsil eder.	$\frac{bir\ kaynakta\ yer\ alan\ doğru\ değerlerin\ sayısı}{kaynaktaki\ toplam\ değer\ sayısı}$
Erişilebilirlik	Kullanılabilirlik, bilginin ne ölçüde mevcut veya kolay ve hızlı bir şekilde elde edilebilir olduğudur. Kullanılabilirlik, tüm erişim yöntemlerinin düzgün çalışmasını ifade eder. Bir veri setinin kullanılabilirliği, bilginin (veya bir kısmının) mevcut, elde edilebilir ve kullanıma hazır olma derecesidir.	$\frac{sunucunun\ SPARQL\ isteklerine\ yanıt\ verme\ sayısı}{SPARQL\ isteklerinin\ toplam\ sayısı}$ $1 - \frac{başarısız\ erişim\ sayısı}{toplam\ HTTP - GET\ request\ sayısı}$

2.5. Web Analitiğinde Bilgi Grafları

Internet üzerindeki heterojen veri kaynakları üzerinde yer alan gürültülü veriler üzerinden bilginin keşfedilmesi, veri analitiğine ve büyük veri analitiğine katkı sağlayacak şekilde makine bilgisine dönüştürülebilmesi bilgisayar bilimleri alanındaki çalışmalarının uzun süredir devam eden bir hedefidir.

Semantik Web alanındaki çalışmalar ile Web üzerinde yer alan heterojen veriler üzerinden bilginin çıkarılması ve yapılandırılmış içeriğin artan kullanılabilirliğinin daha yüksek otomasyon seviyelerine olanak tanıdığı bir Web vizyonunu amaçlamaktadır.

Web üzerindeki heterojen veri kaynakları ve veri miktarı, büyük veri tanımının daha da ötesine geçmiştir. Bugün, büyük veri terimi potansiyel olarak yanıltıcı olabilir çünkü boyut sadece büyük verinin önemli yönlerinden biridir. Büyük veri kavramındaki "büyük" kelimesi, daha fazla verinin iyi veri ve daha güçlü içgörüler anlamına geldiği şeklinde düşünmemize neden olmaktadır. Ancak veri hacminin yalnızca iyi sonuçlar elde etmek için yeterli olmadığını anlamak önemlidir. Verinin nasıl dağıtıldığı, düzenlendiği, entegre edildiği ve temsil edildiği gibi verinin kalitesi ve tutarlılığına yönelik diğer hususlar veri boyutundan çok daha önemli hâle gelmiştir.

Gerek büyük veri alanındaki çalışmalarda gerekse Semantik Web alanındaki araştırmalarda heterojenlik kavramı, veri kaynaklarının ve veri türlerinin farklılığını ifade etmektedir. Heterojenlik değerlendirmesi farklı boyutlar ile ifade edilmekte olup, heterojenlik kavramına ait boyutlara aşağıda yer verilmiştir [32,36]:

- Yapısal çeşitlilik, veri temsilini ifade etmektedir. Örnek olarak, uydu görüntülerinin formatı Web üzerinde oluşturulan tweet'leri depolamak için kullanılan formattan çok farklıdır.
- Ortam çeşitliliği, verinin iletilmekte olduğu ortamı ifade etmektedir. Örnek olarak, bir konuşmanın ses kaydı ile konuşmanın transkripti aynı bilgiyi iki farklı ortamda temsil edebilir.
- Anlamsal çeşitlilik, veriyi ölçmek veya tanımlamak için kullanılan birimlerin (terimlerin) anlamını ifade etmektedir. Örneğin, Etiyopya'daki bir hizmetten gelen yüksek maaş, Amerika Birleşik Devletleri'ndeki benzer bir hizmetten gelen yüksek maaştan anlamsal olarak farklıdır.

- Erişilebilirlik farklılıkları, verinin farklı sürekliliklerde erişilebilir olabileceği anlamına gelmektedir. Trafik kameralarından aralıklı olarak alınan görüntüler ve uydudan gelen verilerin sadece belirli bölgeye yönelmesinde elde edilen görüntüler gibi değerlendirilmesi mümkündür.

Bununla birlikte Web ölçekli heterojen veriler ile çalışmanın beraberinde getirmiş olduğu "verinin belirsizliği", "ölçeklenebilirlik", "güncellik" "tutarlılık" gibi sorunları üstesinden gelebilmek için oluşturulacak bir veri modelinin;

- Heterojen veri kaynakları ve farklı türlerde veriler ile entegre edilebilmesi,
- Hızlı bir şekilde içgörü üretebilmek için analitik yöntemlere ve makine öğrenimi tekniklerinin uygulanmasına olanak sağlaması,
- Operasyonel ve karar süreçlerine yerleştirilmesi,

desteklemesi beklenmektedir.

Bilgi grafları, gerçek dünya problemlerine ait varlıklar arası ilişkilileri sezgisel ve özlü bir soyutlama olarak sunabilmesi ve 2012 yılında Google tarafından arama sonuçlarının iyileştirilmesine yönelik kullanımının duyurulması ile birlikte popüler hale gelmiştir. Bilgi graflarına yönelik olarak birçok çalışmada genel semantik gösterimi veya temel özellikleri tanımlamayı amaçlayan tanımlar verilmiş olmasına rağmen geniş kabul görmüş resmi bir tanım bulunmamaktadır.

Bir bilgi grafı, gerçek dünya olgularını ve anlamsal ilişkilerini, düğümlerden ve kenarlardan oluşan, düğümlerin varlığı ve düğümler arasındaki kenarın ise varlıklar arasındaki anlamsal ilişkiyi temsil ettiği yönlendirilmiş graf olarak tanımlanmaktadır.

Paulheim [70] bilgi graflarına yönelik olarak dört kriter tanımlamıştır.

- Bilgi grafları, gerçek dünya varlıklarını ve bunların bir grafikte düzenlenmiş karşılıklı ilişkilerini tanımlar.
- Bilgi grafları, varlıkların olası sınıflarını ve ilişkilerini bir şemada tanımlar.
- Bilgi grafları, potansiyel olarak herhangi bir varlığı birbirleriyle ilişkilendirmeye olanak tanır.
- Bilgi grafları, farklı konu alanlarını (farklı domainler) kapsar.

Ehrlinger ve Wöß mevcut birkaç tanımı analiz etmiş [81] ve bilgi graflarının muhakeme gücünü vurgulayan önermiş olup; bilgi grafı tanımını, "bilgiyi bir ontolojiye kazandırır, entegre eder ve yeni bilgi türetmek için bir muhakeme uygular" şeklinde ifade etmiştir.

Wang, çoklu ilişkisel çizge olarak bir tanımlama önermiştir [82]. Söz konusu tanıma göre, "bilgi grafı, sırasıyla düğümler ve farklı kenar türleri olarak kabul edilen varlıklar ve ilişkilerden oluşan çoklu ilişkisel bir graftır" şeklinde ifade etmiştir.

- Bir bilgi grafı $G = \{E, R, F\}$ olarak tanımlanmaktadır.
- $E, R\ ve\ F$ sırasıyla varlıklar, ilişkiler ve gerçekler kümeleridir.
- Bir gerçek F ile temsil edilir, bir üçlüden oluşur ve $(h, r, t) \in F$ olarak gösterilir.

Geçtiğimiz son on yıl içinde, Wikipedia gibi düzenlenmiş bilgi paylaşım kaynaklarının kullanılabilirliği gibi gelişmelerden yararlanarak bilgilerin derlenmesi (knowledge harvesting) alanında, Web ölçeğinde İngilizce dilinde ölçeklenebilir bilgi çıkarımına yönelik olarak büyük ilerlemeler kaydedilmiştir.

Bilgi çıkarımı çalışmalarında semantik, verinin anlamı yani gerçek dünyada bir kavramı veya nesneyi temsil etmek için bir veri nesnesinin etkin kullanımı anlamına gelmektedir. 1980'lerden bu yana AI topluluğu, akıllı sistemleri ve ajanları dünya hakkında genel, formalize edilmiş bilgiyle besleme fikrini desteklemektedir.

Web ölçekli verilerin toplanması, doğrulanması ve sorgulanmasını sağlamaya yönelik bilgi graflarının oluşturulması temel olarak aşağıdaki katkıları sunacaktır [83,84,85]:

- Bilgi grafları, varlıklar arasındaki dolaylı ilişkileri temsil eden yollarla sorgulama ve çalışma için daha sezgisel bir temsil ve soyutlama sunmaktadır. Örnek olarak metro haritaları, uçuş rotaları, sosyal ağlar, protein yolları vb. genellikle graf olarak görselleştirilmektedir.
- Bilgi grafları ilişkisel şemalara kıyasla, farklı, eksik, gelişen veriyi daha basit ve esnek bir şekilde temsil etmek için bağımsız bir şekilde tanımlanabilen bir yol sunabilmektedir.
- Bilgi grafları, farklı graflara ait bileşenleri birleşimini mümkün kılmaktadır.
- Makine okunabilir formatta, yüksek hassasiyetle sorgulanabilen, semantik bilgi tabanı sunabilmektedir.

- Metin ifadelerini hızlı ve doğru bir şekilde adlandırılmış varlıklara eşleme konusunda hızlı ve doğru eşlemeleri desteklemektedir.
- Web'de varlık-ilişki odaklı semantik aramayı olanaklı kılmaktadır.
- Web sayfalarında varlıkları ve ilişkileri tespit etmeyi ve bunlar hakkında ifadelerde bulunmayı ve ifadeleri olasılıksal değerlendirmeyi sağlamaktadır.
- Kim/nerede/ne zaman vb. sorularını yanıtlamada varlıklar ve ilişkilerle ilgilenmekte yardımcı olan doğal dil soruları yanıtlama temelini katkı sunmaktadır.

Bilgi graflarında kanonikleştirme (canonicalization), her bir varlığı benzersiz bir şekilde tanımlayabileceğimiz ve aynı varlığa atıfta bulunan tüm gerçeklerin, varlığın keşfedildiği yüzey adlarından bağımsız olarak temsil edilmesini amaçlamaktadır [86,87]. Varlıklara ait belirsizliğin giderilmesi, refere edilen gerçek varlığı bulunmasını amaçlamaktadır.

2.5.1. Bilgi graflarının oluşturulmasına yönelik süreçler

Heterojen veri kaynaklarından doğal dil işleme ve bilgi çıkarma teknikleriyle bilgi graflarının oluşturulmasına yönelik süreçler, bilginin temsili, sunumu, depolanması, birleştirilmesi, kalite parametrelerinin sağlanması vd birçok teknolojiyi içermektedir.

Bilgi grafları, veri tabanlarındaki yapılandırılmış verilerden veya heterojen veri kaynaklarından doğal dil işleme ve bilgi çıkarımı teknikleri kullanılarak oluşturulması mümkündür. Literatürde öne çıkan çalışmalar incelendiğinde, bilgi graflarının oluşturulmasında farklı yaklaşımların önerildiği, amaçlarına göre bilgi graflarının genel amaçlı veya alana özel olarak sınıflandırıldığı görülmektedir [89,90].

Diğer taraftan ontoloji ve şema oluşturma yöntemine göre ise bilgi graflarının tasarımında top down (yukarıdan aşağıya) ve bottom up (aşağıdan yukarıya) olmak üzere iki temel yaklaşımın olduğu, yukarıdan aşağıya yaklaşımında, önceden bir ontoloji veya şemanın tasarlandığı, sonrasında bilgi grafının verilerle donatılarak oluşturulduğu görülmektedir.

Diğer taraftan aşağıdan yukarıya yaklaşımında, grafın oluşturulmasında gerekli bileşenlerin; ontolojilerin, varlıkları ve üçlülerin bilgi çıkarımı teknikleri kullanılarak elde edildiği ve bilgi grafının güncellendiği görülmektedir [91,92].

Yukarıdan aşağıya yaklaşımı, mevcut şemaları ve ontolojileri kullanarak belirli bir alana odaklanan alana özgü bilgi graflarını oluşturmak için kullanılmaktadır. Bununla birlikte aşağıdan yukarıya yaklaşımı açık veri setlerinden (linked open data sets) ontolojiler ve şemalar üretilmesini kapsamaktadır [92]. İki yaklaşım arasındaki temel farkın; alan odağı, şema-ontoloji oluşturulması ve ontoloji birleştirme süreçlerinde oluştuğu görülmektedir [91,92,93].

Literatürde öne çıkan çalışmalar incelendiğinde, bilgi grafı oluşturma süreçlerinin dört ana aşamada olmasının uygun olacağı değerlendirilmektedir [90-94]:

- Bilgilerin temsili: Bu aşama, bilgi grafiği oluşturmak için çeşitli kaynaklardan gelen bilgilerin yapılandırılmış ve makine tarafından okunabilir bir formatta temsil edilmesini içermektedir.
- Bilgi çıkarımı: Yapılandırılmamış veya yarı yapılandırılmış veri kaynaklarından anlamlı bilgi çıkarmak için çeşitli teknikler uygulanır. Doğal dil işleme, adlandırılmış varlık tanıma, ilişki çıkarma ve anlamsal ek açıklama bu aşamada kullanılan bazı yöntemlerdir.
- Veri hizalama ve bilgi füzyonu: Bu aşamada, çıkarılan bilgiler tutarlı bir bilgi grafiği oluşturmak için entegre edilir ve birbirine bağlanır. Bu, varlıkların hizalanmasını ve çizge yapısının düzenlenmesini içermektedir.
- Veri kalitesinin sağlanması: Veri kalitesi güvencesi, bilgi grafiği oluşturmanın çok önemli bir yönüdür. Bu, veri temizleme, veri tekilleştirme, hata işleme, tutarlılık kontrolleri, veri doğrulama, tutarsızlıkları çözme ve bilginin doğruluğunu, eksiksizliğini ve güncelliğini koruma gibi süreçleri içermektedir.

Bu dört aşama, bilgi graflarının oluşturulması için kapsamlı bir çerçeve sunarak çeşitli veri kaynaklarından bilginin toplanması, bilginin çıkarılmasını ve bilginin gösterimini sağlamaktadır. Ancak bilgi graflarının dinamik ve gelişen doğası nedeniyle bu süreçlerin sürekliliğinin sağlanması, bir yaşam döngüsü çerçevesinde ele alınmasını gerektirmektedir [95].

Bilgi grafları üzerindeki verilerin temsilinde varlıkları, varlıklara ait özellikleri ve varlıklar arasındaki ilişkilerin (S-P-O) yada (h,r,t) formatında üçlüler şeklinde temsil edilmesi, bilgi graflarında bilginin temsil edilmesini sağlamaktadır. Bir üçlüden oluşan bilgi yapısı, üç unsurdan oluşan basit bir yapıdır:

-Konu/özne (S): Hakkında beyanda bulunulan varlık veya kavramı temsil eder.

-Yüklem/İlişki (P): Özne ve nesne arasındaki ilişkiyi veya niteliği tanımlar.

-Nesne (O): Yüklem aracılığıyla özne ile ilişkilendirilen değer veya varlığı temsil eder.

Örnek olarak, "Albert Einstein" "ExpertIn" "Physics" üçlü ifadesinde "Albert Einstein" konu başlığını, "ExpertIn" ilişki ya da ontolojiyi ve "Physics" nesneyi temsil etmektedir.

Bu üçlü veri formatı, karmaşık ilişkilerin ve bilgilerin yapılandırılmış bir şekilde temsil edilmesini sağlamaktadır.

2.5.2. Bilginin Temsili

Bilginin temsili (KR), bilgilerin modellenmesini ve temsil edilmesini amaçlayan yapay zekânın alt alanlarından biridir. KR, bilginin daha iyi yorumlanmasını ve anlaşılmasını sağlamak amacıyla hem insan kullanıcılar hem de otomatik işleme için bilginin temsiline odaklanır. KR, varlıklar, onların anlamsal sınıfları, varlıkların nitelikleri ve varlıklar arasındaki ilişkiler hakkındaki bilgileri temsil etmeyi amaçlanmaktadır [91,92].

Bilgi grafı, düğümler ve kenarlardan oluşan yönlendirilmiş bir graf olarak tanımlanımaktadır. Burada düğümler gerçek dünyadaki varlıkları, kenarlar ise varlıklar arasındaki anlamsal ilişkileri temsil etmektedir.

RDF (Resource Description Framework), OWL (Web Ontology Language) ve SPARQL (SPARQL Protocol and RDF Query Language) Semantik Web standartlarında yaygın olarak kullanılan bilginin temsililinde kullanılan dilleridir.

Bu diller, bilgi grafları kavramı ile birlikte, varlıkların ve ilişkilerin temsil edilmesi ve yönetilmesi için temel bileşenleri oluşturur. Bilgi grafları ile ilgili bilgiler tipik olarak bu diller kullanılarak temsil edilir ve yönetilir. Bilgiler RDF store'da RDF (Kaynak Tanımlama Çerçevesi), RDFS (Kaynak Tanımlama Çerçevesi Şeması) veya JSON (JavaScript Nesne Gösterimi) gibi formatlarda temsil edilir [52-63,91-93].

2.5.3. Bilgi Çıkarımı

Bu aşama, bilgi graflarının oluşturulmasının temel ve en önemli yapı taşlarından birini oluşturmakta olup; yapılandırılmamış, yapılandırılmış ve yarı yapılandırılmış verileri içeren homojen ve heterojen veri kaynaklarından varlık ve ilişkileri keşfetmeyi içermektedir [91]. Bu, varlıkları ve ilişkileri keşfetme, varlıklarla ilgili belirsizlikleri çözme ve grafta eksik ilişkileri tamamlama gibi süreçleri içerir [92]. Bu aşamada, farklı veri kaynaklarından ilgili bilgi çıkarımı için doğal dil işleme ve makine öğrenimi gibi çeşitli teknikler kullanılmaktadır.

Bu aşamanın amacı, yapılandırılmış ve yapılandırılmamış veri kaynaklarından yararlanarak bilgi graflarındaki varlıkların ve ilişkilerin kapsamlı bir şekilde ele alınmasını, belirsizliklerin giderilmesini ve eksik bağlantıların tamamlanmasını sağlanması olarak karşımıza çıkmaktadır. Bu süreç, bilgi graflarının zenginleştirilmesini ve daha gelişmiş bilgi tabanlı uygulamalar ve analizlerin yapılmasını mümkün kılmayı da aynı zamanda amaçlamaktadır [95,97,98].

Bilgi çıkarımı aşamasında, doğal dil işleme teknikleri metin tabanlı bilgileri analiz etmek için kullanılır. Bu aşama, varlıkların tespiti, belirsizliklerin giderilmesi ve birbirleri ile ilişkilendirilmesini ve ontoloji ve ilişkilerin çıkarılması gibi alt adımları içerirmektedir. Bu aşama kapsamında işlenmiş verilerden yapılandırılmış verilerin ortaya çıkarılması amaçlanmaktadır [97,98,99].

İsimlendirilmiş varlıkların tespiti (Named Entity Recognition – NER), metinlerden isimlendirilmiş varlıkların (örneğin kişi adları, konumlar, organizasyonlar veya diğer belirli ilgi alanlarındaki varlıklar) tanımlanmasını ve çıkarımını içerir. İsimlendirilmiş varlıkların belirsizliklerinin giderimi (Named Entity Disambugution – NED), varlık adlarında herhangi bir belirsizliğin çözümlenmesini, tespit edilen varlığın farklı bilgi tabanlarından doğru varlıkla ilişkilendirilmesini amaçlamaktadır [100,101].

Varlıkların ve ilişkilerin harici bilgi tabanlarına bağlanması süreci, NEL (Named Entity Linking – NEL) olarak tanımlanmakta olup, gerçek dünya varlıklarının ve ontolojilerinin, diğer bilgi kaynaklarına bağlanmasını, zenginleştirilmesini ve bağlamsal ortak anlayış sağlamayı içermektedir [102].

NERC, NED ve NEL, otomatik çeviri, text özetleme, doküman sınıflandırma, bilgi tabanı ve bilgi graflarının oluşturulması, anlamsal arama gibi uygulamaların geliştirilmesindeki temel yapı taşlarından birisi olarak yer almaktadır. Farklı dillerdeki metin türlerinden varlıkların çeşitliliğini çıkarmak için bir dizi teknik geliştirilmiş olmasına rağmen, doğal dil işleme ve bilgi çıkarımında isimlendirilmiş varlıkların tespiti, sınıflandırılması, belirsizliklerin giderilmesi ve linklendirilmesine yönelik olarak farklı disiplinlerdeki araştırmalar devam etmektedir. Varlıkların belirsizliğinin giderilmesinde farklı dış kaynaklar üzerinden linklendirilmesi belirsizliğinin giderilmesine ve okunabilirliğinin artmasına katkı sağlamaktadır.

Varlıkların tespiti, linklendirilmesi ve belirsizliğinin giderilmesi (NER-NEC-NED), bir metinde geçen varlıkların tanımlanması ve girdi olarak sağlanan bir referans bilgi tabanına bağlanmasını amaçlamaktadır. Varlıkların tespiti, hazır bir NER aracı kullanılarak gerçekleştirilebilineceği gibi; referans sözlük ve bilgi tabanları üzerinden de gerçekleştirilebilir.

İlişkilerin tespiti ve çıkarımı, metinlerde bahsedilen varlıklar arasında anlamlı ilişkileri ve ontolojileri çıkarılması amaçlanmaktadır. Bu aşamada *“PERSON” “lives in” " TURKEY"* gibi doğrudan ilişkileri veya daha derin anlamsal analiz gerektiren daha karmaşık ilişkileri tanımlamaların tespitini amaçlar [103,104].

Genel olarak, bilgi çıkarımı aşaması, yapılandırılmamış kaynaklardan değerli içgörüler ve bilgiyi ortaya çıkarmak için metin verilerini yapılandırılmış temsillere dönüştürmek için doğal dil işleme tekniklerini kullanmaktadır.

2.5.4. Bilgi Füzyonu

Veri entegrasyon alanında devam eden araştırmalar, aynı varlık için farklı verilerin tanımlanması ve varlıklar arasındaki ilişkinin eşleştirilmesine yönelik zorlukları ele almayı amaçlamıştır. Farklı veri kaynaklarından bilgi çıkarma, çeşitli hataları beraberinde getirerek veri birleştirme çabalarını gerektirir. Veri birleştirme, farklı kaynaklardan elde edilen farklı öznitelik değerleri arasındaki tutarsızlıkları çözmeyi, hataları düzeltmeyi ve bilginin doğru temsilini sağlamayı amaçlamaktadır [65,105,106].

Veri entegrasyonu ve bilgi çıkarımı çalışmaları, veri kalitesinin sağlanması, tutarsızlık ve belirsizliklerin giderilmesine yönelik konuları da kapsamaktadır. Birden fazla kaynaktan gelen verilerin hizalanması ve entegre edilmesine yönelik olarak geliştirilen teknikler ile elde edilen bilginin güvenilirliği, tutarlılığı ve güncelliğinin sağlanması amaçlanmaktadır. Bu alandaki araştırmalar, veri heterojenliğinden kaynaklı zorlukları aşmayı, veri doğruluğunu artırmayı ve bilgilerin etkili temsilini amaçlamaktadır.

Veri ve bilgi füzyonu bu endişeleri ele alarak farklı kaynaklardan gelen verilerin entegrasyonu ve birleştirilmesi, veri kalitesinin iyileştirilmesine ve daha anlamlı içgörülerün çıkarılmasına katkıda bulunur [105]. Veri birleştirmenin amacı, birden fazla bilgi kaynağını kullanarak farklı veri kaynaklarından elde edilen değerleri ilişkilendirmek, birleştirmek ve analiz etmek suretiyle bir veri öğesinin doğru değerini çıkarmaktır. Bu işlem, farklı kaynaklardan gelen verileri birleştirerek ve entegre ederek doğru bir değer üretme sürecini içerir.

Daha gelişmiş bir veri birleştirme türü, bilgilerin birleştirilmesidir. Burada odak noktası birden fazla kaynaktan alınan bilginin birleştirilerek gerçek s-p-o (konu-yüklem-nesne) değerlerinin belirlenmesi, çakışmaların giderilmesi ve birden fazla doğru değerin olması durumunda bilginin temsiline yönelik stratejilerin belirlenmesini sağlamaktır [97,106].

Bu süreç, güven değerlerinin hesaplanmasını ve olasılık temelli akıl yürütme tekniklerini kullanmayı içerir. Bilgi birleştirme, farklı kaynaklardan gelen çatışan veya tamamlayıcı bilgileri basit bir birleştirmeden daha fazlasını yaparak uzlaştırmayı, elde edilen bilgilerin doğruluğunu ve güvenilirliğini artırmayı amaçlamaktadır.

Bilgi birleştirme sürecinde sadece veriler birleştirilmez, aynı zamanda varlıkların ve ilişkilerinin gerçek değerlerini belirlemek için çaba gösterilir. Veri kaynaklarının

güvenilirliği, güvenilirliği ve tutarlılığı analiz edilir ve güven değerleri hesaplanarak bilginin güvenilirliği değerlendirilir. Belirsizlikleri ele almak ve bilinçli kararlar vermek için olasılık hesaplamaları kullanılır.

Veri birleştirme ve bilgi birleştirme tekniklerini kullanarak, birden fazla bilgi kaynağının değerlendirilerek, belirsizliklerin ve çakışmaların etkili bir şekilde ele alınarak veri öğeleri için doğru ve güvenilir değerler elde edilmesi hedeflenmektedir [106].

2.5.5. Bilgi Kalitesi

Bilgi grafları, karar verme ve içgörü sağlama amacıyla bir araya getirilmiş bağlantılı varlıklar ve ilişkilerin yapılandırılmış bir çerçevesidir. Bilgi graflarının kalitesi, içerdiği bilgilerin güvenilirliği ve kullanılabilirliği üzerinde doğrudan etkilidir. Kalite metrikleri genellikle grafın verileri ne kadar doğru, kapsamlı ve zamanında yakaladığı, temsil ettiği ve ilettiği üzerine odaklanır.

Bölüm 2.4.'te yer verilen kalite boyutları ve parametreleri, bilgi graflarının oluşturulması içinde geçerli olup, bilgi graflarının yaşam döngüsü içerisinde sürekliliğinin sağlanması önem arz etmektedir [75,76].

Doğruluk ve eksiksizlik temeldir; bunlar verilerin doğruluğunu ve grafiğin çeşitli konuları ve ilişkileri ne kadar kapsamlı bir şekilde ele aldığını ölçer. Tutarlılık, ayrıntı seviyesi ve bağlantılılık, verilerin uniform, yeterince detaylı ve iyi bağlantılı olmasını sağlar, bu da karmaşık sorgulara ve daha iyi iç görülere olanak tanır. Güncellik, güvenilirlik ve alaka düzeyi de kritiktir, verilerin güncel kalmasını, güvenilir kaynaklardan gelmesini ve hedef kullanım durumlarına uygun olmasını sağlar.

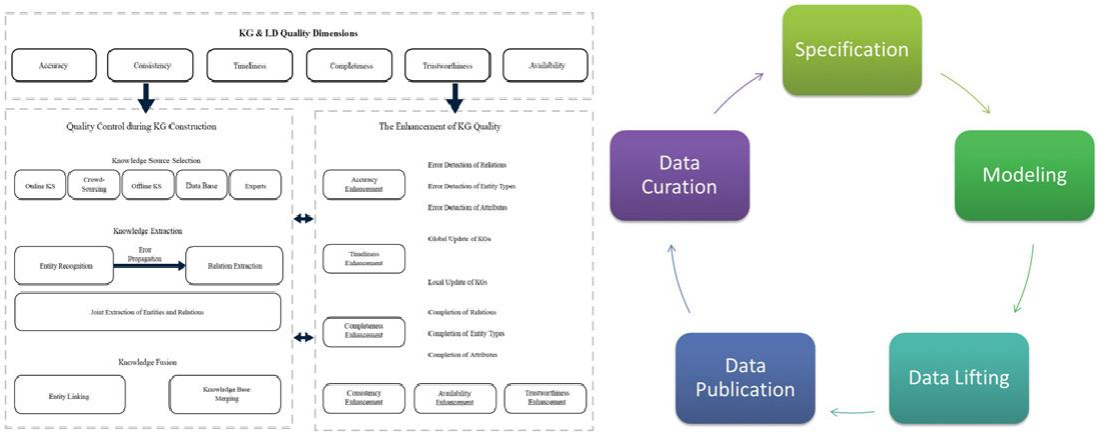
Bu ana parametrelerin yanı sıra, grafiğin uzun vadeli sürdürülebilirliği ve diğer sistemlerle kolay entegrasyonu için ölçeklenebilirlik ve birlikte çalışabilirlik gereklidir.

Erişilebilirlik ve veri kaynağı da eşit derecede önemlidir, kullanıcıların grafiği ne kadar kolay bir şekilde etkileşime girebildikleri ve verilerin kaynağının ne kadar şeffaf bir şekilde belgelendiği üzerinde etkilidir. Genel olarak, kalite parametreleri, bir bilgi grafının etkililiğini değerlendirmek ve iyileştirmek için çok boyutlu bir yaklaşım sunar.

Bilgi graflarının kalite yönü ile değerlendirilmesi, grafta sunulan bilgilerin doğruluğunu, eksiksizliğini ve kullanışlılığını sağlamayı amaçlamaktadır.

Bölüm 2.4.'te yer verilen kalite boyutları ve parametreleri, bilgi graflarının oluşturulması içinde geçerli olup, bilgi graflarının yaşam döngüsü içerisinde sürekliliğinin sağlanması önem arz etmektedir [75,76].

Şekil 2.8'de bilgi grafları ve bağlantılı veri çalışmalarında kalite bileşenleri ve bilgi graflarının yaşam döngüsüne birlikte yer verilmiştir [71,75].



Şekil 2.8. Bilgi grafları ve bağlantılı veri çalışmalarında kalite bileşenleri

3. WEB VERİ KAYNAKLARININ GÜVEN DEĞERİNİN HESAPLANMASI İÇİN YENİ BİR MODEL

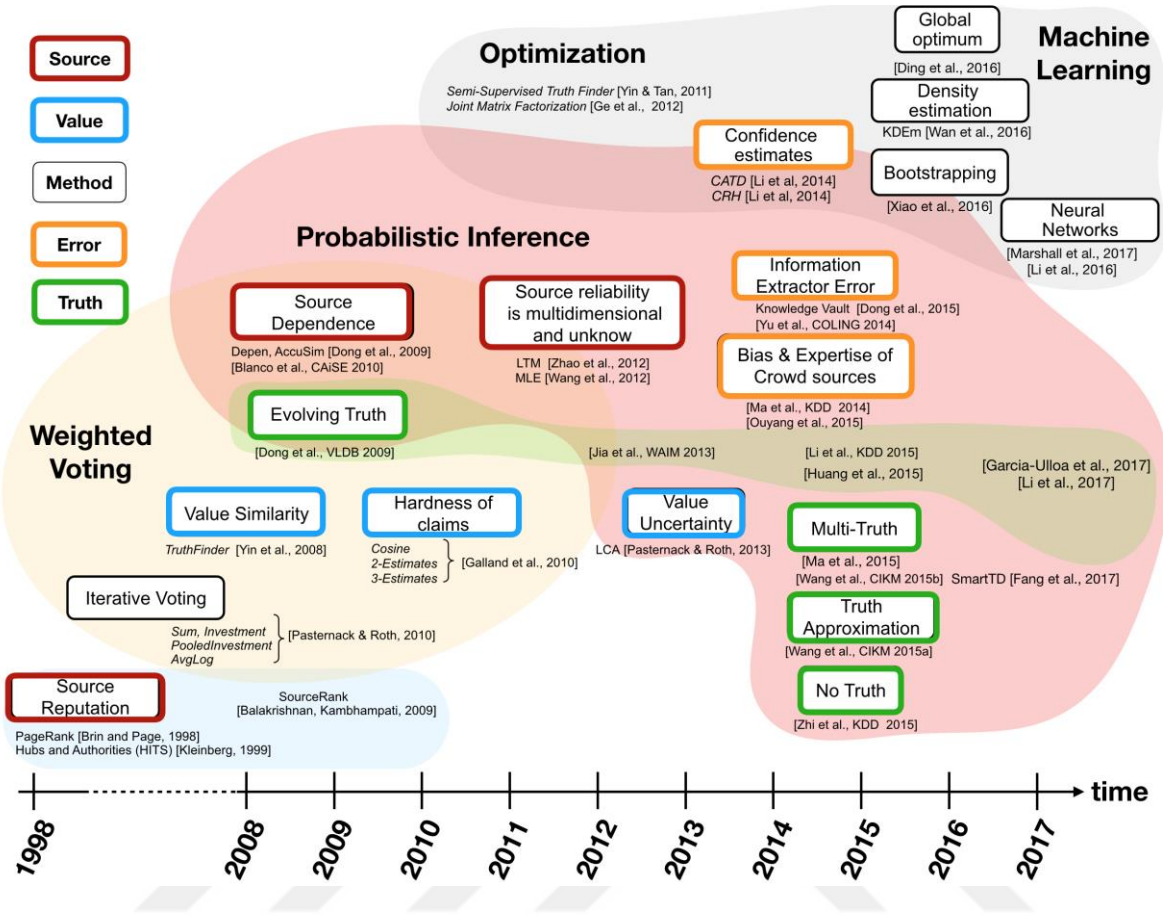
İnternet'in daha geniş kaynaklara sosyal medya platformları aracılığı ile erişilebilir olması; bilgi dolaşımının hızla yaygınlaşmasına olanak sağlamıştır. Bununla birlikte çeşitli algoritmaların, otomasyon sistemleri vasıtası ile verilerin şekillendirilerek yayılması yolu ile kamusal yaşamın şekillendirilmesini ifade eden bilişimsel propaganda kavramının günümüzde oldukça yaygınlaşmış durumda olduğu görülmektedir.

Gerçekliğin keşfi, yapay zekâ, veri tabanları ve karmaşık sistemler alanındaki birçok araştırma topluluğu tarafından incelenen uzun süredir devam eden zorlu bir sorun olarak karşımıza çıkmaktadır. Gerçekliğin keşfi problemi literatürde "fact-checking", "data and knowledge fusion", "information trustworthiness", "credibility", "information corroboration", "truth discovery" olarak farklı adlar altında incelendiği görülmektedir [31].

Gerçekliğin keşfi, veri tabanı disiplinindeki ve Web ölçekli bilgi tabanı oluşturmaya yönelik araştırmalarda bilgilerin entegrasyonunda yaşanan sorunların çözümü amacı ile ortaya çıktığı, heterojen veri kaynaklarından elde edilen gürültülü bilgiler arasındaki çatışmaları çözümünün öncelikli olarak amaçlandığı görülmektedir [107].

Gerçekliğin keşfi çalışmaları ile birden fazla bilgi kaynağı tarafından iddia edilen bir dizi iddianın doğruluk etiketini tahmin etmek ve kaynakların güvenilirliğini önceden bilgi olmadan veya çok az bilgi ile çıkarmaktır. Literatürde gerçekliğin keşfine yönelik ilk çalışma Yin ve arkadaşları tarafından yapılmış, iteratif hesaplamalı yöntem önerilmiştir [108].

Gerçekliğin keşfine yönelik olarak literatürde sınırlı sayıda öne çıkan araştırmacı tarafından sınırlı sayıda yöntemin önerildiği görülmektedir. Literatürde gerçekliğin keşfine yönelik yapılan çalışmalarda kullanılan yöntemlere, gerçekliğin keşfi konusundaki araştırmaların katkısına ve kronolojik olarak ilerlemesine Şekil 3.1'de yer verilmiştir [106].



Şekil 3.1. Gerçekliğin keşfine yönelik yapılan çalışmalar

Şekil 3.1'de yer alan bilgiler incelendiğinde:

- Gerçekliğin keşfine yönelik yeni buluşların ve katkıların yer aldığı çalışmaların 2008-2017 yılında arasında yapıldığı,
- Source, Value, Method, Error ve Truth parametreleri açısından söz konusu çalışmaların güçlü yanlarının vurgulandığı,
- Gerçekliğin keşfine yönelik genel yöntemlerin ağırlıklı oylama, optimizasyon tabanlı ve olasılıksal yaklaşımları içerdiği,
- Son çalışmalarda makine öğrenimi yöntemlerinin uygulandığı,

görülmektedir.

3.1. Gerçekliğin Keşfi Çalışmalarında Kullanılan Yöntemler

Günümüzün veri odaklı dünyasında, çok sayıda kaynaktan gelen bilgilerin doğruluğunu belirlemek önem arz etmektedir. Birden fazla kaynağı analiz ederek ve karşılıklı referans alarak bilgilerin güvenilirliğini belirlemeyi amaçlayan gerçekliğin keşfine yönelik çalışmalar bu çabada kritik bir rol oynamaktadır. Heterojen veri kaynaklarından elde edilen veriler üzerinden yapılan bilgi çıkarımlarında, aynı bilgiye yönelik çatışmalar, tutarsızlıklar, gürültü ve yanıltıcı bilgiler ile karşılaşılması gelmesiyle gerçek keşfi için çeşitli yöntemler geliştirilmiştir.

Literatürde gerçekliğin keşfine yönelik çalışmalar incelendiğinde genel olarak üç yöntemin olduğu görülmektedir. Söz konusu yöntemler:

- Oylama tabanlı yöntemler
- İstatistiksel tabanlı yöntemler
- Optimizasyon tabanlı yöntemler

şeklinde olup detaylarına ilerleyen bölümlerde yer verilmiştir.

3.1.1. Ağırlıklı oylama tabanlı yöntemler

Ağırlıklı oylama tabanlı yöntemler (MV-Majority Voting), değer etiketlerini (doğru veya yanlış) ve güven puanlarını çoğunluk oylamasının (MV) bazı varyantlarını kullanarak hesaplar. Geleneksel MV, kaynakların çoğunluğu tarafından iddia edilen değeri doğru değer olarak kabul eder ve eşitlik durumunda rastgele bir değer seçer [107,108,109,110]. Tablo 3.1'de çoğunluk oylamasının uygulanmasına yönelik örneğe yer verilmiştir [108].

Çizelge 3.1. Çoğunluk oylamasının uygulanmasına ilişkin örnek

Source/Result	George Washington	Abraham Lincoln	Mahatma Gandhi	John Kennedy	Barack Obama	Franklin Roosevelt
Source 1	Virginia	Illinois	Delhi	Texas	Kenya	Georgia
Source 2	Virginia	Kentucky	Porbandar	Massachusetts	Hawaii	New York
Source 3	Maryland	Kentucky	Mumbai	Massachusetts	Kenya	New York
Majority Voting	Virginia	Kentucky	Delhi	Massachusetts	Kenya	New York

MV yönteminin her kaynağın aynı kaliteye sahip olduğunu varsayması, ancak gerçekte kaynaklar farklı kalitelere, kapsama alanına ve kapsama alanına sahip olması nedeni ile tek başına bu şekilde değerlendirme olası hatalı sonuçlara neden olacaktır.

3.1.2. Gerçekliğin keşfi çalışmalarında istatistiksel tabanlı yöntemler

Bayesian ve olasılıksal tabanlı yöntemler MV'nin eksikliklerini gidermek, kaynakların ve olguların bilinmeyen özellikleriyle ilgili sorunlarını tam olarak ele almak için önerilmiştir. Truth Finder algoritması bu yöntem üzerine üretilmiş, kaynak güvenilirliğini tahmin etmek ve değer benzerliğini dikkate alarak doğru iddiaları belirlemek için Bayes analizini uygulayan ilk yöntemdir [107,108,111].

Depen ve Accu modelleri [107] aynı zamanda kaynak kopyalama tespit tekniklerini içeren ilk Bayes modelleridir ve kaynak bağımlılığının doğruluk keşfindeki önemini vurgulamaktadır.

Olgulara ait değerler belirlenen gerçeğe yakınsa bunları destekleyen kaynaklar ağırlık kazanacaktır. Amaç, olasılığı en üst düzeye çıkarmak ve iddialar gerçeğe yakın olduğunda yüksek kaynak ağırlıklarını tahmin etmektir.

Olasılıksal ve istatistiksel modeller, kaynak güvenilirliğindeki belirsizlikleri ele almak için tasarlanmıştır. Bayesian tabanlı modeller, kaynakların güvenilirliğini değerlendirmek ve etkilerini buna göre ayarlamak için Bayesian çıkarımının gücünden yararlanmaktadır. Ancak, bu modeller genellikle bağımsız kaynakları varsayar, bu da gerçek dünya senaryolarında, bilgi kaynaklarının birbirine bağımlı olabileceği durumlarda geçerli olmayabilir.

Eşitlik 3.1'de genel formüle yer verilmiştir [107].

$$\prod_{s\epsilon S} p(w_s|\beta) \prod_{o\epsilon\theta}(p(v_o^s|\alpha) \prod_{s\epsilon S} p(v_o^s| v_o^*, w_s)) \quad (3.1)$$

3.1.3. Gerçekliğin keşfi çalışmalarında optimizasyon tabanlı yöntemler

Doğruluk keşfinin iteratif yöntemler ile keşfedilmesi yönteminde doğruluk hesaplaması ve kaynak güvenilirliği tahmini birbirlerine bağlıdır. Bu nedenle bazı doğruluk bulma yöntemleri doğruluk hesaplama adımı ve kaynak ağırlık tahmini adımının yakınsama sağlanana kadar iteratif olarak yürütüldüğü iteratif prosedürler olarak tasarlanmıştır.

Gerçek keşfi için kullanılan iteratif yöntemler, bir çözüme yaklaşmak için sürekli olarak bilgiyi günceller. Bu yöntemler, başlangıçta varsayılan bir gerçeği veya kaynak güvenilirliğini iteratif olarak rafine ederek doğruya yaklaşmaya çalışır. Ancak, bu yöntemler bazen belirli bir eşiğe ulaşıncaya kadar veya belirli bir durma kriteri karşılanıncaya kadar birden çok iterasyon gerektirebilir.

Optimizasyon tabanlı yöntemler, kaynak kaliteleri ile iddiaların doğruluğu arasındaki ilişkiyi yakalayabilen bir optimizasyon fonksiyonu belirlemeye dayanır. Bu iki parametre kümesini birlikte hesaplamak için iteratif bir yöntem kullanılır. İteratif yöntemler ile kaynak güvenilirliği ve fact doğruluğu adım adım iteratif olarak ve istenen değere yakınsama sağlanıncaya kadar hesaplama işlemlerine devam edilmektedir [107,108,109,110,111].

Bir hedef fonksiyonunu en iyileştirerek (genellikle bir kayıp fonksiyonunu minimize ederek) gerçeği keşfetmeyi amaçlarlar. Optimizasyon tabanlı yöntemler, genellikle büyük veri setlerinde hızlı ve etkili olabilirler, ancak doğru bir hedef fonksiyonunun seçilmesi kritiktir.

Optimizasyon tabanlı yöntemlerde, v_O^s kaynak verisi ile karşılaştırılan ile v_O^* verileri arasındaki uzaklık ölçümleri gerçeklenir, min uzaklığı sağlayan değerlerin güvenilirliği yüksek kabul edilir.

Genellikle optimizasyon fonksiyonu Eş 3.2'de yer verildiği gibi formüle edilir [107,108]:

$$arg_{\{w_s\}} min_{\{v_o^*\}} \quad \sum_{o \in O} \sum_{s \in S} w_s \; . \; d\,(v_o^s, v_o^*) \tag{3.2}$$

Çizelge 3.2'de güven değeri hesaplamasında kullanılan yöntemler ve özelliklerine yer verilmiştir.

Çizelge 3.2. Güven değeri hesaplamasında kullanılan yöntemler ve özellikleri

		Yöntemler ve Algoritmalar						
Boyutlar	**Alt Değerlendirme Bileşenleri**	**Truth Finder**	**3 - Estimates**	**MLE**	**SimpleLCA**	**LTM**	**DEPEN**	**SSTF**
Yöntem	Karmaşık parametre ayarı	Hayır	Hayır	Hayır	Hayır	Evet	Evet	Evet
	Tekrarlanabilirlik	Evet	Evet	Evet	Evet	Hayır	Evet	Evet
	Yakınsama	Evet	Evet	Hayır	Hayır	Hayır	Evet	Evet
	Ölçeklenebilirlik	Evet	Evet	Hayır	Evet	Evet	Hayır	Evet
	Eğitilebilirlik	Hayır	-	Hayır	Hayır	Hayır	No	Yes
Veri Girişi	Veri Tipi (String:S Kategorik:C, Sayısal: N, Boolean: B)	S, C, N	S, C, N	B	S, C, N	S, C, N	S, C, N	S, C, N
	Tekli / Çoklu	Tekli / Çoklu	Tekli / Çoklu	Tekli	Tekli / Çoklu	Tekli	Tekli / Çoklu	Tekli / Çoklu
	Benzerlik	Evet	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır
	Korelasyonlar	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır
Önceki Bilgi	Kaynak Kalitesi	Sabit, Uniform	Sabit, Uniform	Sabit, Kaynak türüne bağlı	Sabit, Kaynak ve veri türüne bağlı	Kaynak türüne bağlı	Sabit, Kaynak ve veri türüne bağlı	Sabit, Uniform
	Kaynak Bağımlılığı	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır	Hayır
	Talep Zorlukları	Hayır	Evet	Hayır	-	Hayır	Hayır	Hayır
	Tekli / Çoklu	Tek	-	Tek	Tek	Çoklu	Tek	Tek
Çıktı	En az bir gerçek /Gerçek yok	En az bir gerçek	En az bir gerçek	En az bir gerçek	En az bir gerçek	En az bir gerçek	En az bir gerçek	En az bir gerçek
	Açıklama, Zenginleştirme	Yok	Yok	Yok	Yok	Yok	Yok	Yok

3.2. Gerçekliğin Keşfi Çalışmalarının Bileşenleri

Veri biliminde gerçekliğin keşfi, çeşitli endüstriler ve araştırma alanları için giderek daha fazla öneme sahip duruma gelmektedir. Mevcut veri setleri üzerinde anlamlı ve kullanışlı bilgilerin çıkarılması, organizasyonların daha iyi kararlar alabilmesi için kritik bir faktördür.

Ancak bu çalışmalar, bir dizi karmaşık bileşeni içerirmektedir. Gerçekliğin keşfi çalışmalarının ilk adımı, doğru ve yüksek kaliteli veri toplamaktır. Toplanan veriler üzerinde yer alan eksik, aşırı değerler ve gürültülü veri gibi sorunlar, analizin doğruluğunu olumsuz etkilere neden olacağı nedeni ile ön işlemeyi gerektirmektedir. Diğer taraftan uygun model seçimi, gerçekliğin keşfi çalışmalarının bir diğer önemli bileşenidir. Güven değerinin hesaplanmasını etkileyen bileşenlere aşğıda yer verilmiştir [107,108]:

3.2.1. Giriş verisi

Gerçekliğin keşfi çalışmalarında kaynaktan gelen verilerin ön işleme öncesinde farklı yönleri ile değerlendirilmesi gerekmektedir. Bu nedenle giriş verileri farklı yönleri ile değerlendirilmesi esastır. Giriş verilerinde dikkat edilmesi gereken hususlara aşağıda yer verilmiştir [106,107,112]:

- Tekrarlı Veri: Farklı çalışmalarda her bir farklı kaynaktan bir fact'in bir kez geleceği varsayımı değerlendirilmektedir. Bununla birlikte tekrarlayan/güncellenen veriler için sistem üzerinde tekrarlı dolaşımı ya da güncellemeyi sağlayacak bir politikanın belirlenmesi ve hamdatanın değerlendirilmeye alınması önem arz etmektedir.
- Değerlendirilemeyen Nesneler: Herhangi bir çakışmaya uğramayan fakat değerlendirmeler sonucunda hakkında herhangi bir işlem tesis edilemeyen verilere ne yönde işlem yapılacağının belirlenmesi de önem arz etmektedir.
- Giriş Veri Formatı: Farklı kaynaklardan aynı nesneye yönelik gelecek verilerde farklılık arz edebilir. Bu nedenle format farklılıklarını dikkate alarak değerlendirilme yapılması gerekmektedir.
- Yapılandırılmış-Yapılandırılmamış Verilerin Değerlendirilmesi: Literatürde yer alan çalışmaların bir kısmının yapılandırılmış veriler üzerinden gerçeklik tespiti yaptığı, yakın zamanda yapılan çalışmalarda ise yapılandırılmamış ham veri üzerinden hesaplamaların yapıldığı görülmektedir. Yapılandırılmamış veriler üzerinden

yapılan çalışmaların fayda getirebileceği değerlendirilmektedir, bununla birlikte gürültülü ve belirsizlik içeren veriler üzerinden işlem yapılıyor olması nedeni ile beraberinde zorluk getireceği değerlendirilmektedir.

- Akış Verileri: Literatürde yer alan çok sayıda çalışmanın statik veriler üzerinden sonuç ürettiği görülmektedir. Çevrim içi akış verileri üzerinden sonuç üretilmesine yönelik değerlendirmeler önem arz etmektedir.

3.2.2. Kaynak Güvenilirliği

Veri biliminde gerçekliğin keşfi çalışmaları, genellikle büyük veri setlerinin analizini ve modellemesini içerir. Bu tür çalışmaların başarısı, büyük ölçüde kullanılan verinin kalitesine ve kaynak güvenilirliğine bağlıdır. Veri kaynağının güvenilir olmaması, çalışmanın sonuçlarını yanıltıcı ve hatalı hale getirecektir. Kaynak güvenilirliğini belirleyen bileşenlere aşağıda yer verilmiştir [106,107,112]:

- Kaynak tutarlılığı: Bir kaynaktan üretilen fact değerlerinin ve fact'ler ile ilişkili nesnelere aynı olasılık değeri ile doğru bilgi sağladığı yaklaşımıdır.
- Kaynak bağımsızlığı: kaynaktan üretilen fact'lerin birbirinden bağımsız olarak doğru veriler için benzerlik göstereceği varsayımıdır.
- Kaynak bağımsızlığı analizi: Birbirine bağlı olarak veri kopyalayan kaynaklarda aynı hataların yapılacağı bu nedenle veri kopyalamanın olabileceğinin dikkate alınmasıdır. Bununla birlikte güvenilir kaynak ve doğru bilgiler için bu yaklaşımın çözüm sağlamayacağı değerlendirilmektedir.
- Kaynak seçimi: Gerçeklik tespitine yönelik çalışmalarda güvenilirlik değeri düşük kaynaklar üzerinden yapılan inceleme çalışmaları sonucu olumsuz olarak etkileyecektir.

3.2.3. Varlıkların güvenilirliğinin değerlendirilmesi

Veri biliminde "gerçekliğin keşfi" çalışmalarında varlıkların (örneğin, veri kaynakları, algoritmalar veya veri toplama araçları gibi) güvenilirliği büyük önem taşır. Güvenilir olmayan bir varlık, çalışmanın tüm sonuçlarını geçersiz kılabilir, yanıltıcı analizlere yol açabilir veya stratejik kararlar için yanlış yönlendirmelerde bulunabilir. Güvenilirlik, bir varlığın veri toplama, veri işleme ve analiz aşamalarında konsistans ve doğruluk sağlama kapasitesi olarak değerlendirilir. Özellikle, veri toplama araçlarının ve kaynaklarının güvenilir olması, toplanan verinin kalitesini ve dolayısıyla analizin güvenirliğini doğrudan etkileyeceği değerlendirilmektedir.

Varlıkların güvenilirliğini değerlendirmek için çeşitli yöntemler ve metrikler mevcuttur. İlk olarak, varlığın geçmiş performansı ve itibarı incelenebilir. Eğer bir veri kaynağı veya toplama aracı daha önce güvenilir sonuçlar ürettiyse, bu genellikle iyi bir göstergedir. İkinci olarak, veri setinin istatistiksel özellikleri, örneğin, eksik veri oranı, hata payı ve anomaliler gibi faktörler, varlığın güvenilirliği hakkında bilgi sağlar. Üçüncü olarak, bağımsız doğrulama veya çapraz doğrulama yöntemleri, farklı varlıkların ürettiği veri veya sonuçlar arasındaki uyumu değerlendirerek güvenilirlik hakkında daha fazla bilgi verebilir.

Gerçeklik tespitine yönelik az sayıda çalışmada (3-Estimates, Multi-dimensional Truth-Finding Model) nesnelerin güven değerinin hesaplanması ve kaynak güvenilirliği üzerinden nesnelerin güvenilirliğinin hesaplandığı, birlikte artışının yapıldığı ya da cezalandırıldığı görülmektedir [106,107,112,114]:

- Tamamlayıcı oylama yaklaşımı*:* Bu teknik, "gerçek olan her nesne için yalnızca bir gerçek değer" olduğu varsayımına dayanmaktadır. Aynı zamanda Local Closed World Assumption (LCWA) yöntemi ile benzerlik arz etmektedir. Bilgi tabanında yer alan doğruların doğru olduğu, dışarıda yer alan diğer bilgilerin ise yanlış değer içerdiği varsayımı ile hareket edilmektedir.
- Değerler arasındaki bağlantı*:* Bir nesneye yönelik Fact değerlerinin birbirini etkileyebileceği varsayımı ile hareket edilmekte olup yüksek olasılık değerine sahip fact değeri sonuca yansıtılır.
- Bir duruma ilişkin olarak farklı türdeki verilerin değerlendirilmesi: farklı türdeki verilerin işlenmesine benzerliğinin ölçümüne yönelik yaklaşımların sağlanması da önem arz etmektedir.

- Veri türlerindeki belirsizliklerin değerlendirilmesi: verinin sayısal ya da alfanumerik olmasına yönelik gerekli çözüm yaklaşımlarının sağlanması gerekli olduğu değerlendirilmektedir.
- Hiyerarşik verilerin değerlendirilmesi: bir nesye ait birden fazla doğrunun olması durumuna yönelik yaklaşımların da değerlendirilmesi önem arz etmektedir. Bu senaryoya uygun bir kişinin doğum yeri için ilçe/il/ülke fact verileri ile karşılaşılması durumunda sistemin nasıl davranması gerektiğinin yaklaşımlarının tasarıma yansıtılması olarak değerlendirilmelidir.

3.2.4. Çıktı Değerlerinin Değerlendirilmesi

Veri biliminde "gerçekliğin keşfi" çalışmalarında elde edilen sonuçların doğru bir şekilde değerlendirilmesi, projenin başarısı için kritik öneme sahiptir. Sonuçlar, genellikle bir model veya algoritma tarafından üretilir ve bu sonuçların doğruluğu, çalışmanın güvenilirliği, yenilikçiliği ve uygulanabilirliği açısından belirleyicidir. Eğer sonuç değerleri yanlış değerlendirilirse, bu yanıltıcı yorumlara, stratejik hatalara ve hatta olumsuz etik veya yasal sonuçlara yol açabilir. Bu nedenle, sonuç değerlerinin doğru bir şekilde değerlendirilmesi, analizin bütünlüğü ve güvenilirliği için şarttır.

Sonuç değerlerini değerlendirmek için çeşitli yöntemler ve metrikler kullanılır. İstatistiksel testler, modelin performansını ölçmek için sıklıkla başvurulan bir yöntemdir. Modelin uygulanabilirliği ve genelleştirilebilirliği önemli bir değerlendirme faktörüdür. Bunun için çapraz doğrulama, karışıklık matrisi, hassasiyet, özgüllük ve F1 skoru gibi metrikler sıklıkla kullanılır. Diğer taraftan çıktı değerlerinin değerlendirilmesi gereken diğer hususlara aşağıda yer verilmiştir [106,107,112,114]:

- Birden Fazla Gerçek Değer ile Karşılaşılma Durumu (Single truth v.s. multiple truths): Gerçeklik değerinin hesaplanmasına yönelik tasarımda bir nesneye ilişkin birden fazla fact ve farklı doğruluk değerlerinin olması durumunda sistemin nasıl davranacağına yönelik yaklaşımın belirlenmesi işlemidir. Bir kişinin doğum yeri için ilçe/il/ülke üçlü olgu/iddia değerleri ya da doğum tarihi ile birlikte yaş verisinin yer aldığı üçlü olgu/iddia değerleri örnek olarak gösterilebilir.
- Güvenilirliği Ölçülemeyen Doğru Değerlerin Gösterimi*:* Sistemin herhangi bir sonuç üretemediği, güvenilir kaynaklardan elde etmiş olduğu verilerin gösteriminde UNK olarak gösteriminin sağlanması yaklaşımıdır.

- Sonuç Değerlerinin Gösterim Yöntemi (Labeling v.s. scoring): hesaplama işlemi sonucu elde dilen verilerin True/False ya da hesaplanan güven değeri ile gösteriminin sağlanması işlemidir.

3.3. Veri Kaynakları için Güven Değeri

"Heterojen Web Veri Kaynakları için Güven Değeri" kavramı, İnternetten elde edilen farklı türdeki verilerin güvenilirliği ve güvenilirliğini değerlendirmek için kullanılan bir metrik veya metrik setini ifade etmektedir.

Heterojen Web veri kaynakları, sosyal medya gönderileri, haber makaleleri ve akademik dergilerden veritabanlarına, açık kaynak raporlar vd. her şeyi içerebilir. Güven değeri, birden çok kaynaktan veri toplanıp analiz edilerek bilgilendirilmiş kararlar veya tahminler yapılacak senaryolarda çok önemlidir. Bu süreç, gürültülü ve hatalı veri üreten veri kaynaklarının belirlenmesi ve filtrelenmesi ve verinin faydasının daha doğru bir temsilini sağlamayı amaçlamaktadır.

Güven değerini hesaplarken, genellikle kaynağın itibarı, tarihsel doğruluğu, verilerin güncelliği ve hatta bilgilerin eldeki göreve bağlamsal alakası gibi bir dizi faktör göz önünde bulundurulur. Gelişmiş algoritmalar ve makine öğrenimi modelleri, bu güven değerlerini dinamik olarak atamak ve güncellemek için sıkça kullanılır. Söz konusu süreç ve bileşenler, sistemlerin ve karar vericilerin, farklı kaynaklardan gelen verilere uygun şekilde ağırlık vermesini sağlar, böylece analiz veya çıktının kalitesi artırmayı amaçlamaktadır. Böylece, çeşitli amaçlar için Web kaynaklı veri kullanırken güven ve güvenilirlik katmanının eklenmesi amaçlanmaktadır.

Bu bölümde kaynak güvenilirliği değerlendirmesi, doğruluk keşfinin en önemli özelliği olduğundan, burada kaynak güvenilirliği tahmini hakkında bazı varsayımları ele alınmıştır.

3.3.1. Kaynak tutarlılık varsayımı

Çoğu doğruluk keşfi yöntemi kaynak tutarlılık varsayımına sahiptir, bu şu şekilde açıklanabilir. Bir kaynak, tüm nesneler için aynı olasılıkla doğru bilgi sağlama eğilimindedir. Bu varsayım birçok uygulamada makul bir varsayımdır. Kaynak güvenilirliğinin tahmininde en önemli varsayımlardan biridir [110-116].

3.3.2. Kaynak bağımsızlığı varsayımı

Bazı doğruluk keşfi yöntemleri kaynak bağımsızlık varsayımına sahiptir. Bu durum şu şekilde yorumlanabilir. Kaynaklar gözlemlerini birbirinden bağımsız olarak yaparlar, birbirlerinden kopyalama yapmazlar. Bu varsayım şu şekilde ifade edilebilir: Gerçek bilgi farklı kaynaklar arasında benzer veya aynı olma olasılığı daha yüksektir ve farklı kaynaklar tarafından sağlanan yanıltıcı bilgi daha farklı olma olasılığı daha yüksektir [110-116].

3.3.3. Kaynak bağımlılığı analizi

Bu yöntemde kaynaklar arasında kopyalama ilişkisi tespit edilmeye çalışılır ve daha sonra tespit edilen ilişkiye göre kaynakların ağırlıkları ayarlanır. Kopyalama tespiti, bazı kaynakların diğerlerinden bilgi kopyaladığı uygulama senaryolarında faydalıdır. Kopyalama tespiti, bazı kaynakların birbirlerinden çok sayıda ortak hata yaptığı durumlarda etkili olur, çünkü bu kaynaklar birbirleriyle bağımsız değillerdir. Ancak bazı kaynaklar bilgiyi iyi bir kaynaktan kopyaladığında bu prensip etkisiz hâle gelir. Kopyalama tespit işlemi doğruluk keşfiyle sıkı bir şekilde entegre edilir ve tespit edilen kopyalama ilişkisi ile keşfedilen gerçek değerler iteratif olarak güncellenir [108,110].

3.4. Web Veri Kaynakları İçin Güven Değeri Hesaplama Modeli

Heterojen Web veri kaynakları üzerinden verilerin toplanması, doğrulanması ve güven değerinin hesaplanmasına yönelik problemin çözümünde tüm veri kaynaklarının eşit derecede güvenilir olmaması nedeni ile güven değeri hesaplama modeli oluşturulması önem arz etmektedir.

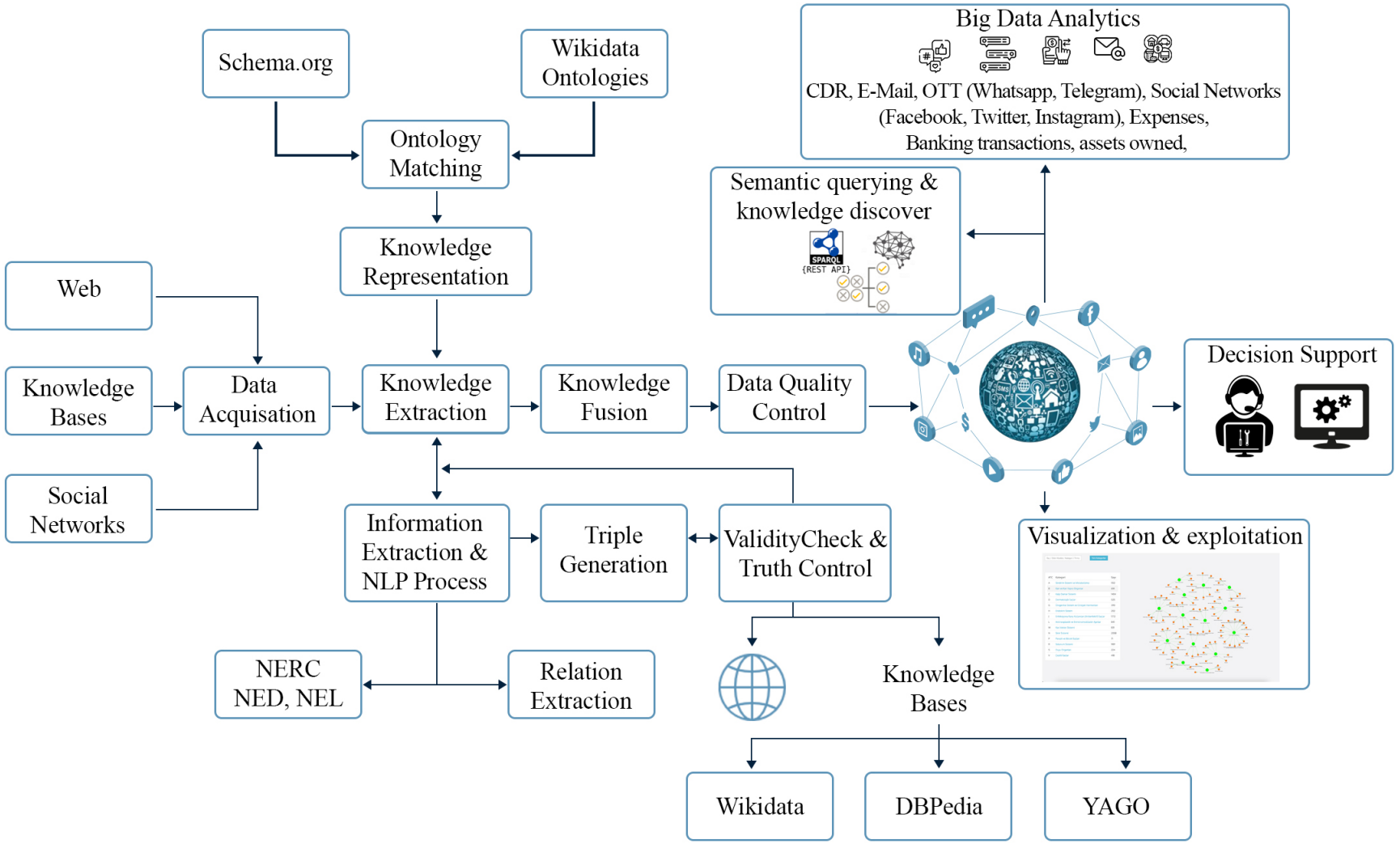
Bu bölümde tez kapsamında önerilen özgün bilgi grafı modeli ve güven değeri hesaplama modeli ve mimarisi ana hatlarıyla ele alınmıştır. Bu kapsamda literatürde öne çıkan çalışmalarda yer alan yöntemler ve gelişmeler metodolojik olarak incelenerek özgün bir bilgi grafı modeli tasarlanmış ve yazılım platformu geliştirilmiştir.

3.4.1. Geliştirilen Mimari

Yapılan literatür taraması ve örnek incelemeler sonucunda güven değerinin hesaplanmasına yönelik çalışmalar incelendiğinde, temel gerçek (ground truth) değeri üzerinden güven değeri hesaplamalarının yapıldığı, bu bağlamda Web ölçekli geniş kapsamlı bir bilgi tabanı oluşturmanın gerekliliği görülmüştür [31,32,33].

Bu kapsamda özgün bir bilgi tabanı ve bilgi grafı modeli tasarlanmış ve buna yönelik uygulama platformu geliştirilmiştir. Geliştirilen özgün model Türkçe diline yönelik olarak, genel amaçlı, uçtan uca, Web ölçekli, doğal dil işleme, bilgi çıkarımı ve güven değeri hesaplamalarını içermektedir.

Yukarıda yer verilen süreçleri gerçekleştirmek amacı ile geliştirilen yazılım platformu TRPedia genel mimarisine Şekil 3.2’de yer verilmiştir.



Şekil 3.2. TRpedia genel mimarisi

Şekil 3.2'de geliştirilen TRPedia yazılım platformu özgün mimarisine verilmiştir. Söz konusu genel mimari kapsamında bileşenlere aşağıda yer verilmiştir.

- Data Acquisation bloğunda sistem, heterojen Web veri kaynakları, bilgi tabanlarından ve sosyal medya platformlarında veri toplamaktadır.
- Knowledge Representation bloğunda sistem schema.org ve Wikidata.org ontolojileri arasında eşleştirme yaparak Türkçe diline özgü temel ontolojilerin uluslararası çalışmalarda kabul görmüş ontolojilerini oluşturmaktadır.
- Knowledge Extraction bloğu, doğal dil işleme ve bilgi çıkarımı (IE) tekniklerini kullanarak varlıkların tespiti, sınıflandırılması, belirsizliğinin giderilmesi ve tespit edilen ontolojiler kapsamında üçlülerin oluşturulmasını sağlamaktadır. Oluşturulan üçlüler aynı zamanda "validity check" aşamasında olması gereken paternlere uyumluluğu kontrol edilmektedir.
- Knowledge Fusion bloğu kapsamında farklı kaynaklardan gelen bilgiler çakışma-tutarsızlık gibi durumlar karşısında stratejilerin uygulandığı aşamadır.
- Gerçekliğin keşfi (Truth Discovery) ile elde edilen üçlülerin güven değeri hesaplamaları, verilerin elde edildiği kaynakların güven değeri hesaplamaları ve diğer bilgi tabanları (YAGO ve DBPedia) ile çapraz kontroller sağlanmaktadır.
- Yukarıda yer verilen tüm aşamaları geçen üçlüler, bilgi tabanında diğer varlıklar ile ilişkilendirilerek saklanmaktadır.
- Görselleştirme aşamasında ise bilgi tabanında yer alan üçlülerin ilgili varlığa ait sorgusu gerçeklendiğinde; bilgi grafı formatında diğer varlıklar ile bağlantılarını içerecek şekilde tarayıcıda gösterimi ve bilgi kutusu gösterimi sağlanmaktadır.

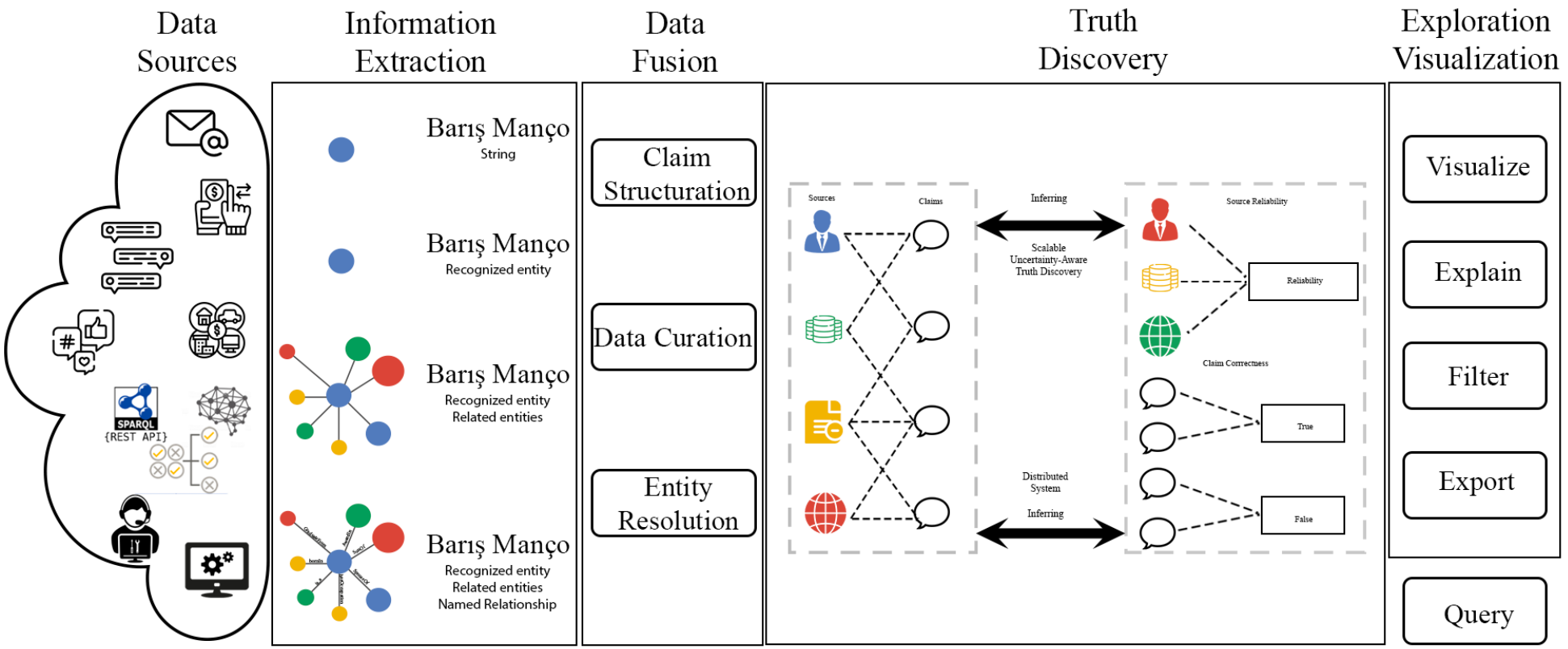
3.4.2. Önerilen Model

Doğal dil işleme ve bilgi çıkarımı teknikleri kullanılarak genel amaçlı bir bilgi tabanı oluşturulması, bilginin üçlülere indirgenmesi ve bilgi grafı mimarisinde gösteriminin gerçekleştirilmesi amacıyla Türkçe konu başlıklarına sahip yaklaşık 530.000 Wikipedia sayfasından oluşan veri kümesi kullanılmıştır.

Geliştirilen özgün model ve yazılım platformu kapsamında aşağıda yer verilen süreçleri gerçekleştirmek amacı ile genel mimariye uygun biçimde kademeli olarak gösteren TRPedia blok diyagramına Şekil 3.3'te yer verilmiştir. Şekil 3.3'te yer verilen blok diyagram kapsamında aşağıda yer verilen süreçler yerine getirilmiştir:

- İnceleme yapılacak etki alanlarının belirlenerek farklı kaynaklardan verilerin toplanması,
- Web sayfalarında yer alan metinlerin belirlenecek yapıya göre ayrıştırılması (parse edilmesi,)
- Bilgi tabanının oluşturulması,
- Yaygın olarak bilgi tabanlarında kullanılan ontolojilerin tespiti ve söz konusu ontolojiler arasında haritalama/eşleştirmenin yapılması,
- Metinlerden IE ve NLP temel yöntemleri uygulanarak verilerin üçlü formatına indirgenmesini sağlamak amacı ile:
    - Varlıkların tespit edilmesi (NER), Varlıklara yönelik anlamsal belirsizliklerin giderilmesi (NED), Varlıkların yaygın olarak kullanılan bilgi tabanlarında linklendirilmesi (NEL),
    - İlişkilerin çıkarımı (RE),
    - RDF formatında üçlülerin üretilmesi,
    - Üretilen RDF üçlülerinin doğrulanması,
- Farklı bilgi tabanları ile entegrasyonu ve gelen bilgilerin bilgi füzyonu işlemine tabi tutulması,
- Bilgi tabanı güncelliği ve doğruluğu,
- Bilgi tabanında yer alan bilgilerin bilgi grafı formatında gösterimi,
- Platformun sürekli olarak veri akışını, incelenmesini, saklanmasını ve ilgili ortamlarda güvenlik gereksinimleri ile saklanması ve paylaşılmasını sağlayacak uygun yaşam döngüsünün oluşturulmasını,

Sağlamak amacı ile özgün bir bilgi grafı modeli ve buna yönelik uygulama platformu geliştirilmiştir.



Şekil 3.3. TRPedia blok diyagramı

Şekil 3.3'ün bileşenlerine yönelik açıklamalara aşağıda yer verilmiştir.

Geliştirilen TRPedia özgün modeli ve yazılım platformu literatürde öne çıkan çalışma ve yöntemler dikkate alınarak geliştirilmiş olup, 5 aşamalı bloktan oluşmaktadır.

- Data Sources bloğu; sistemin heterojen Web veri kaynakları vd. veri kaynaklarından yapılandırılmamış, yarı yapılandırılmış ve yapılandırılmış veriler ile entegrasyonuna olanak sağlamaktadır. Bu blok içerisinde Web Crawler, Web Scraping ve Web parsing tekniklerini uygulamak amacı ile PHP yazılım dilinde uygulama bileşeni geliştirilmiştir.

- Information Extraction bloğunda temel bilgi çıkarımı teknikleri dikkate alınarak üçlülerin çıkarımı sağlanmıştır. IE bloğunda süreçleri gerçekleştirmek amacı ile PHP yazılım dilinde uygulama bileşeni geliştirilmiştir. Çıkarımı sağlanan üçlüler MYSQL veritabanında yönlendirilmiş ilişkileri gösterimi sağlayacak şekilde saklanmıştır. Diğer taraftan metin işleme işlemlerinde hız gerektiren işlemler için Elastic Search kullanılmıştır.

  Barış Manço örneği üzerinden açıklamalara yer verecek olur isek:

  - Metin içerisinde yer alan "Barış Manço" ifadesi metin ayrıştırma işlemleri sonucunda tespiti yapıldığında öncelikle "string" olarak tanımlanmaktadır.
  - NER tekniğinin uygulanması sonrasında ise "Barış Manço" ifadesi bir varlıktır (Recognized Entity).
  - NERC tekniğinin uygulanması sonrasında ise "Barış Manço" varlıksal olarak "PERSON" insan olarak tanımlanmıştır.
  - NED teknikleri uygulanması sonrasında "Barış Manço" isimlendirilmiş varlığa yönelik bir anlamsal belirsizliğin bulunmadığı görülmüştür.
  - RE tekniklerinin uygulanması sonrasında sistemde tanımlı ontolojiler üzerinden metin içerisinde yer alan "Barış Manço" ya ait ilişkiler ve ilişkili olduğu varlıklar tespit edilmektedir ("Recognized Entity, Related Entities with Named Relationship").
  - Bu bloğun sonucunda varlığa ait geçerli üçlülerin diğer bilgi tabanları ile eşleştirilerek oluşturulma süreci tamamlanmaktadır.

- Data Fusion aşamasında, “Barış Manço” ait üçlüler arasında çatışma olup olmadığı, çatışma olması durumunda data & knowledge fusion stratejileri ile çatışmaların önlenmesi sağlanmaktadır. Bu aşamada söz konusu süreçleri yerine getirmek amacı ile PHP yazılım dilinde uygulama bileşeni geliştirilmiştir.

- Truth Discovery (gerçekliğin keşfi ve güven değeri hesaplamaları) bloğunda önceki bloklardan elde edilen üçlülerin bilgi tabanında saklanması öncesi, üçlülerin ve üçlülerin elde edildiği kaynakların güven değerinin hesaplanması sağlanmaktadır.

- Bu aşamada Web veri kaynaklarının sınıf değerini belirlemek amacı ile TF-IDF ve metin benzerliği algoritmalarını uygulamak amacı ile uygulamak amacı ile PHP yazılım dilinde uygulama bileşeni geliştirilmiş, Elastic Search teknolojisi kullanılmıştır.

- Exploration & Visualization bloğunda ise önceki aşamalardan geçerek bilgi tabanında yer alan bilgilerin, bilgi grafı formatında ilişkileri ve üçlüleri ile birlikte gösterimi sağlanmaktadır. Bu aşamadaki gereksinimleri yerine getirmek amacı ile e PHP yazılım dilinde uygulama bileşeni geliştirilmiştir.

Bununla birlikte bu aşama aynı zamanda varlıklara ait sorgu sonuçlarında ontoloji tabanlı ilişkilerin gösterimi aynı zamanda anomalilerin gösterimini de olanaklı kılmaktadır.

3.4.3. Güven Değeri Hesaplanmasına Yönelik Detaylar

Bu bölümde gelilştirilen özgün model ve uygulama platformuna ait detaylara yer verilmiştir.

Veri kaynakları

Geliştirilen özgün bilgi grafı modeli ve uygulama platformu TRPedia, aşağıda yer verilen farklı türdeki text verisini işleme tabi tutabilmektedir.

- Yapısal olmayan veriler (Unstructured data; Wikipedia ve Web sayfaları)
- Yarı yapısal veriler (Semi Structured data; Wikipedia bilgi kutuları, Yago, DBpedia)
- Yapısal veriler (Structured data; veri tabanları ve bilgi tabanları)

Söz konusu veri kaynakları üzerinde yer alan verilerin çekilmesi amacı ile Web parsing teknikleri kullanılmıştır.

Bilginin temsili

Bilgi temsili, bilgiyi modellemeyi ve temsil etmeyi amaçlayan yapay zekâ alt alanlarından biridir. Bir bilgi grafi, düğüm ve kenarlarından oluşan yönlendirilmiş bir graf olarak tanımlanmaktadır. Geliştirilen modelimizde, oluşturulan üçlüler bilgi tabanı üzerinde saklanmıştır.

Bilgi tabanları, bilgiyi insan kullanıcıları ve otomatik işleme için bilgiyi daha iyi yorumlama ve anlama imkânı sağlamak amacıyla temsil etme üzerine odaklanmaktadır. Bilgi tabanları, varlıklar hakkındaki bilgiyi, anlamsal sınıflarını, varlıkların özelliklerini ve varlıklar arasındaki ilişkileri temsil etmeyi amaçlar.

Geliştirilen modelde, düğümler gerçek dünya varlıklarını temsil ederken, kenarlar varlıklar arasındaki anlamsal ilişkileri temsil etmektedir. Bu kapsamda üretilen bilgilerin bilgi tabanlarında saklanması ve RDF (kaynak tanımlama çerçevesi) formatında temsili sağlanmıştır.

Bilgi grafı parametrelerine aşağıda yer verilmiştir:

- $W = \{W_1, W_1 \ldots . W_N\}$ $Wikipedia$ içerisinde yer alan Türkçe konu başlıkları
- $W_i = \{S_1, S_{2,} \ldots, S_m\}$ her konu başlığı içerisinde yer alan cümleler
- $G = \{E, R, F\}$ etiketli ve yönlü bilgi grafı
- $E = \{e_1, e_2 \ldots . e_n\}$ varlıkları kümesi (düğümler),
- $R = \{r_1, r_2 \ldots . r_m\}$ ilişkilerin kümesi (kenarlar),
- $F_{i,j,k} = \{e_i, r_j, e_k\}$ üretilen üçlülerin kümesi ($facts/triples$),
- $Y_{i,j,k} \in \{0,1\}$ "1" geçerli üçlülerin olması durumunda bilgi tabanında temsili,
- $Y \in \{0,1\}^{N_i R_j N_k}$ geçerli üçlünün, 3 boyutlu matriste temsili,
- $Y_{ijk} = \begin{cases} 1, (e_i, r_j, e_j)\ triple\ exist \\ 0, \quad other \end{cases}$

Bilgi çıkarımı süreçleri

Geliştirmiş olduğumuz özgün bilgi grafının oluşturulmasındaki temel ve en önemli yapı taşlarından olan bu aşamada; heterojen veri kaynakları üzerindeki yapısal olmayan (unstructured), yapısal (structured) ve yarı yapısal (semistructured) veriler üzerinden; varlıkların ve ilişkilerin keşfedilmesi, varlıklara yönelik belirsizliklerin giderilmesi, graf üzerindeki eksik ilişkilerin tamamlanmasına yönelik süreçleri kapsamaktadır.

Bilgi tabanında yer alan bilgilerin gösteriminde bilgi grafı mimarisi gösterimi tercih edilmiştir. Bilgi grafı gösterimi benzerlik, topluluk ya da anomali durumlarının tespitini rahatlıkla ortaya koyabilmektedir.

Bilgi çıkarımı aşamasında, doğal dil işleme teknikleri ile metinsel bilgilerin ayrıştırılarak (parse edilerek), isimlendirilmiş varlıkların tespiti (NER), belirsizliğinin giderilmesi (NED), isimlendirilmiş varlıkların linklendirilmesi (NEL), ilişki çıkarımı (RE) gibi alt aşamalar sonrasında yapısal verilerin ortaya konulmasına yönelik süreçleri içermektedir.

Bilgi grafının oluşturulmasında, bilgi çıkarımı ve doğal dil işleme yöntemleri ile metinlerden isimlendirilmiş varlıkları çıkarmak ve bunları graf üzerinde düğümlerde temsil edilmesini sağlamak aşağıda yer verilen dört ana görevi içermektedir [99,122]:

- İsimlendirilmiş Varlık Tanıma (NER- Named Entity Recognition), metin içerisindeki adlandırılmış bir varlıkların tespitini,
- İsimlendirilmiş Varlıkların Sınıflandırılması (NEL- Named Entity Classification), isimlendirilmiş varlığa ait türün belirlenmesi (PER/LOC/ORG/MISC)
- İsimlendirilmiş varlıkların belirsizliğinin giderilmesi (NED- Named Entity Disambugution),
- İsimlendirilmiş Varlıkların Linklendirilmesi (NEL- Named Entity Linking), belirsizliği giderilmiş her varlık için standart bir IRI sağlanmasını

içermektedir.

NERC-NERD-NEL aşamalarından sonra birbirine bağlı olmayan varlıklar arasındaki ilişkilerin tespiti gerektirmektedir. Yapılandırılmış veriler için ilişkiler açık ve kolayca tanımlanabilmekte iken, yarı yapılandırılmış veriler arasındaki ilişkilerin tespitinde diğer makine öğrenimi tekniklerinin yanı sıra örüntü tabanlı ve kural tabanlı yaklaşımlar kullanılması gerekmektedir.

Tez kapsamında geliştirilen özgün modelde, Wikipedia'da yer alan tüm konu başlıkları hiyerarşik varlık – ilişki çıkarımını sağlayacak şekilde aşağıdaki bilgi çıkarımı bileşenleri uygulanmıştır:

- NERC çözümünde Hidden Markov Models (HMM), Support Vector Machines (SVM), Maximum Entropy models (MaxEnt), Decision Trees (DT), ve Conditional Random Fields (CRFs) uygulandığı görülmektedir. Sözlük/CRF/kural tabanlı hibrit yöntem kullanılmıştır.
- NERD çözümünde: bilgi tabanlı yöntemler, istatistiksel hesaplamalar, benzerlik, TF-IDF, Page-Rank hesaplamlarının kullanıldığı görülmektedir [123,124,125]. Yapılandırılmamış, metinsel veriler söz konusu olduğunda ise anlamsal rol etiketleme (sequential role labeling), öznitelik çıkarımı, kernel tabanlı yaklaşımların kullanıldığı görülmektedir [126-129]. Doküman benzerliği, varlığın bulunduğu paragraf içerisindeki durumunun değerlendirilmesi, Yago/DbPedia üzerinde linklendirilmesi yöntemlerini içeren hibrit yaklaşım kullanılmıştır.

- İlişki Çıkarımında: Boot Strapping, (Person lives LOCATION) benzeri örüntüler ve kural tabanlı (Cümle içerisinde NER ve Ontolojiler) yöntemler kullanılmıştır.

Aşağıda PER/LOC/ORG kategorisinde örnek cümleler ve üçlülere yer verilmiştir.

Person:

1. Pattern: <person> <was born in> <location>
Örnek cümle : Mustafa Kemal Atatürk İzmir'de doğdu.

2. Pattern: <person> <is known for> <achievement/activity>
Örnek cümle : Ahmet Arif, şiirleriyle tanınır.

3. Pattern: <person> <invented/discovered> <invention/discovery>
Örnek cümle: Türk Hekim Hulusi Behçet, Behçet hastalığını keşfetti.

4. Pattern: <person> <was a leader of> <country>
Örnek cümle : Turgut Özal Türkiye'nin 8. Cumhurbaşkanıydı.

5. Pattern: <person> <won> <award/year>
Örnek cümle : Aziz Sancar 2015 Nobel Kimya Ödülü'nü kazandı.

Location:

1. Pattern: <location> <is in> <country>
Örnek cümle: Ankara Türkiye'de bulunan bir şehirdir.

2. Pattern: <location> <is known for> <landmark/attraction>
Örnek cümle: Kapadokya peribacalarıyla ünlüdür.

3. Pattern: <location> <was founded in> <year>
Örnek cümle: İstanbul 330 yılında kuruldu.

4. Pattern: <location> <is a part of> <region>
Örnek cümle: Ege Bölgesi Türkiye'nin bir parçasıdır.

5. Pattern: <location> <is famous for> <product/resource>

Örnek cümle: Çanakkale zeytinyağı üretimiyle ünlüdür.

Organization:

1. Pattern: <organization> <was founded in> <year>

Örnek cümle: Türk Hava Kurumu 5 Şubat 1925'te kuruldu.

2. Pattern: <organization> <is located in> <location>

Örnek cümle: TÜBİTAK Ankara'da bulunan bir kurumdur.

3. Pattern: <organization> <focuses on> <area of expertise>

Örnek cümle: UNICEF çocuk hakları alanında faaliyet gösterir.

4. Pattern: <organization> <was established by> <founder>

Örnek cümle: Anadolu Ajansı İsmet İnönü tarafından kurulmuştur.

5. Pattern: <organization> <provides> <product/service>

Örnek cümle: Türk Telekom iletişim hizmetleri sunar.

Bilgi füzyonu

Bilgi füzyonu kapsamında farklı bilgi tabanları ile entegrasyonu ve gelen bilgilerin bilgi bilgi tabanı güncelliği ve tutarlılığının sağlanması amaçlanmış; TRPedia bilgi tabanı YAGO ve DBPedia bilgi tabanları ile eşleştirilmiş, bilgi çıkarımı aşamasında kullanılan ontolojilerin schema.org ile mümkün olduğunca uyumlu olması amaçlanmıştır.

Schema.org, Google, Yahoo, Bing ve Yandex tarafından anlamsal arama ve gösterime yönelik olarak üzerinde ontolojiler konusunda uzlaşı sağlamış olmasıdır [93,96].

Çizelge 3.3'te varlıkların çıkarımı ve linklendirilmesine (Entity Extraction & Linking systems) yönelik literatürde yer alan çalışmalarda geçen yöntemleri içerir karşılaştırmalı sonuçlara yer verilmiştir. Söz konusu çizelgede, çalışmanın ya da ürünün adı, çalışmanın gerçekleştirildiği sene, varlıkların linklendirilmesinde kullanılan bilgi tabanı, eşleştirme yöntemi, indeklemede kullanılan yöntem, çalışmada kullanılan içerik, metin işlemede kullanılan yöntem ya da araç ve belirsizliğin giderilme yöntemi (M/K/G/C/L) olarak yer verildiği görülmektedir.

Çizelge 3.4'te belirsizliğin giderilmesinde kullanılan özelliklere yönelik literatürde yer alan çalışmalarda geçen yöntemleri içerir karşılaştırmalı sonuçlara yer verilmiştir. Söz konusu çizelgede ürün adları ve kullanılan parametre ve boyutlara yer verildiği görülmektedir [32].

Çizelge 3.5'te kavram çıkarımı ve linklendirilmesine (Concept Extraction & Linking systems) yönelik literatürde yer alan çalışmalarda geçen yöntemleri içerir karşılaştırmalı sonuçlara yer verilmiştir [32]. Söz konusu çizelgede çalışmanın adı, hangi etki alanına yönelik yapıldığı (domain), çalışmanın amacı ya da hedefi, filtreleme metodu, ilişkiler, bağlantı, refererans kaynağı ve kullanılan yönteme yer verildiği görülmektedir [32].

Çizelge 3.3. Varlıkların çıkarımı ve belirsizliğin giderilmesine yönelik çalışmalar

KB, kullanılan ana bilgi tabanını ifade eder;
Eşleştirme ve İndeksleme, KB'deki varlık etiketlerini eşleştirmek/indekslemek için kullanılan yöntemleri ifade eder;
Bağlam, kullanılan bağlamsal bilgi kaynaklarını ifade eder;
Tanıma, varlık bahsini tanımlama sürecini ifade eder;
Anlam ayrımı, kullanılan üst düzey anlam ayrımı özelliklerinin türlerini ifade eder (M: Mention, K: Keyword, G: Graph, C: Category, L: Linguistic);'-' bilgi bulunamadı, kullanılmadı veya uygulanamaz anlamına gelir

Sistem	**Yıl**	**Bilgi tabanı**	**Eşleştirme**	**İndeks**	**İçerik**	**Tespit**	**Belirsizliğin Giderilme Yöntemi**
ADEL	2016	DBpedia	Keyword	Elastic Couchbase	Wikipedia	Tokens Stanford POS Stanford NER	M,G
AGDISTIS	2014	Any	Keyword	Lucene	Wikipedia	Tokens	M,G
AIDA	2011	YAGO2	Keyword	Postgres	Wikipedia	Stanford NER	M,K,G
AIDA-Light	2014	YAGO2	Keyword LSH	Dictionary LSH	Wikipedia	Tokens	M,K,G,C
		Wikipedia					
Babelfy	2014	WordNet	Substring	—	Wikipedia	Stanford POS	M,G,L
		BabelNet					
CohEEL	2016	YAGO2	Keywords	—	Wikipedia	Stanford NER	K, G
DBpedia Spotlight	2011	DBpedia	Substring	Aho-Corasick	Wikipedia	LingPipe POS	M,K
DoSeR	2016	DBpedia YAGO3	Exact Keywords	Custom	Wikipedia	—	M, G
ExPoSe	2014	DBpedia	Substring	Aho-Corasick	Wikipedia	LingPipe POS	M,K
ExtraLink	2003	Custom (Tourism)	—	—	—	SProUT XTDL	[*Manual*]
GianniniCDS	2015	DBpedia	Substring	SPARQL	Wikipedia	—	C
JERL	2015	Freebase Satori	—	—	Wikipedia	Hybrid (CRF)	K,G,C,L
J-NERD	2016	YAGO2	Keyword	Dictionary	Wikipedia	Hybrid (CRF)	M,K,C,L
Kan-Dis	2015	DBpedia Freebase	Keyword	Lucene	Wikipedia	Stanford NER	K,G,L
KIM	2004	KIMO	Keyword	Hashmap	—	GATE JAPE	G
KORE	2012	YAGO2	Keyword LSH	Postgres	Wikipedia	Stanford NER	M,K,G
LINDEN	2012	YAGO1	—	—	Wikipedia	—	C
MAG	2017	Any	Keyword Substring	Lucene	Wikipedia	Stanford POS	M,G
						UIUC NER	
NereL	2013	Freebase	Keyword	Freebase API	Wikipedia	Illinois Chunker Tokens	M,K,G,C, L
NERFGUN	2016	DBpedia	Substring	Dictionary	Wikipedia	—	M, K, G
NERSO	2012	DBpedia	Exact	SPARQL	Wikipedia	Tokens	G
SDA	2011	DBpedia	Keyword	—	Wikipedia	Tokens	K
SemTag	2003	TAP	Keyword	—	Lab. Data	Tokens	K
THD	2012	DBpedia	Keyword	Lucene	Wikipedia	GATE JAPE	K,G
Weasel	2015	DBpedia	Substring	Dictionary	Wikipedia	Stanford NER	K, G

Çizelge 3.4. EEL sistemlerinde belirsizliğin giderilmesinde kullanılan özellikler

EEL sistemleri tarafından kullanılan anlam ayrımı özelliklerine genel bakış

(M:Metrik tabanlı, K:Anahtar kelime tabanlı, G:Grafik tabanlı, C:Kategori tabanlı, L:Dilbilimsel tabanlı)

(mo:mention-only, mm:mention-mention, mc:mention-candidate, co:candidate-only, cc:candidate-candidate; v:various)

Sistem	String benzerliği – [M | mc]	Tür karşılaştırması – [M | mc]	Anahtar ifadeler – [M | mo]	Çakışan başlıklar – [M | mm]	Abbreviations – [M | mm]	• Bahsi geçen aday bağlamları – [K | mc]	Aday-aday bağlamları – [K | cc]	Yaygınlık / Öncelik – [G | mc]	İlişkisellik (Ortak atıf) – [G | cc]	İlişkisellik (KB) – [G | cc]	Merkezilik – [G | co]	Kategoriler – [C | v]	Dilbilimsel – [L | v]
ADEL	✓	✓		✓	✓	✓	✓		✓				
AGDISTIS	✓			✓						✓	✓		
AIDA			✓			✓		✓	✓	✓			
AIDA-Light	✓					✓	✓					✓	
Babelfy											✓		✓
CohEEL						✓		✓		✓			
DBpedia Spotlight	✓		✓			✓							
DoSeR	✓							✓			✓		
ExPoSe [238]	✓	✓	✓			✓							
GianniniCDS												✓	
JERL			✓			✓	✓	✓	✓				✓
J-NERD	✓	✓	✓		✓	✓	✓					✓	✓
Kan-Dis						✓	✓			✓	✓	✓	✓
KIM										✓			
KORE			✓		✓	✓	✓	✓	✓	✓			
LINDEN									✓			✓	
MAG	✓			✓	✓	✓				✓	✓		✓
NereL	✓		✓				✓	✓	✓	✓		✓	✓
NERFGUN	✓		✓				✓			✓	✓		
NERSO											✓		
SDA						✓							
SemTag						✓							
THD						✓					✓		
Weasel						✓			✓		✓		

Çizelge 3.5. Kavram çıkarımı ve linklendirilmesinde kullanılan yöntemler

Ortam, deneylerin gerçekleştirildiği birincil alanı belirtir;
Hedef, sistemin belirtilen amacını gösterir (KE: Anahtar Kelime Çıkarımı, OB: Ontoloji Oluşturma, SA: Anlamsal **Açıklama**, TE: Terminoloji Çıkarımı, TM: Konu Modelleme);
Tanıma, kullanılan terim çıkarma yöntemini özetler;
Filtreleme, uygun alan terimlerini seçmek için kullanılan yöntemleri özetler;
İlişkiler, metindeki terimler arasında çıkarılan anlamsal ilişkileri gösterir (Hyp.: Hyponyms, Syn.: Synonyms, Mer.: Meronyms);
Bağlantı, terimlerin bağlı olduğu KB'leri/ontolojileri gösterir;
Referans, kullanılan diğer kaynakları gösterir;
Yöntem, konuları modellemek (ve etiketlemek) için kullanılan yöntem(ler)i gösterir. '-' bilgi bulunmadığını, kullanılmadığını veya uygulanabilir olmadığını gösterir.

Çalışma	Yıl	Domain	Hedef	Tanıma	Filtreleme	İlişkiler	Bağlantı	Referans	Yöntem
ArllahyariK	2015	Çoklu alan	TM	Token tabanlı	İstatistiksel	—	DBpedia	Wikipedia	LDA / Graph
Canopy	2013	Çoklu alan	TM	-	-	—	DBpedia	Wikipedia	LDA / Graph
CardilloWRJVS	2013	Tıp	TE	TExSIS	Manuel	—	SNOMED, DBpedia	—	—
ChemuduguntaHSS	2008	Bilim	SA	Jeton tabanlı	İstatistiksel	—	—	CIDE, ODP	LDA
ChenJYYZ	2016	Komp. Bilim, Haberler	TM	DBpedia Spotlight	Grafik/İstatistik.	—	DBpedia	—	pLSA / Graph
CimianoHS	2005	Turizm, Finans	OB	POS tabanlı	İstatistiksel	Hyp.	—	—	—
CRCTOL	2010	Terörizm, Spor	OB	POS tabanlı	İstatistiksel	Hyp.	—	WordNet	—
Distiller	2014	-	KE	Stat./Leksik	Hibrit	—	DBpedia	—	—
DolbyFKSS	2009	I.T., Enerji	TE	POS tabanlı	İstatistiksel	—	DBpedia, Freebase	—	—
F-STEP	2013	Haberler	TE	WikiMiner	WikiMiner	Hyp.	DBpedia, Freebase	Wikipedia	—
FGKBTE	2014	Suç, Terörizm	TE	Token tabanlı	İstatistiksel	—	FunGramKB	—	—
GillamTA	2005	Nanoteknoloji	OB	Örüntüler	Hibrit	Hyp.	—	—	—
GullaBI	2006	Petrol	OB	POS tabanlı	İstatistiksel	—	—	—	—
JainP	2010	Comp. Bilim.	TM	POS tabanlı	Stat. / Kılavuz	—	Custom onto.	—	Hierarchy
JanikK	2008	Haberler	TM	-	İstatistiksel	—	Wikipedia	—	Graph
LauscherNRP	2016	Politika	TM	DBpedia Spotlight	İstatistiksel	—	DBpedia	—	L-LDA
LemnitzerVKSECM	2007	E-Öğrenme	OB	POS tabanlı	İstatistiksel	Hyp.	OntoWordNet	Web search	—
LiTeWi	2016	Yazılım, Bilim	OB	Topluluk	WikiMiner	—	Wikipedia	—	—
LossioJRT	2016	Biyomedikal	TE	POS tabanlı	İstatistiksel	—	UMLS, SNOMED	Web search	—
MoriMIF	2004	Sosyal Veri	KE	TermeX	İstatistiksel	—	—	Google	—
MunozGCHN	2011	Telekomünikasyon	KE	POS tabanlı	Hibrit	—	DBpedia	Wikipedia	—
OntoLearn	2001	Turizm	OB	POS tabanlı	İstatistiksel	Various	—	WordNet, Google	—
OntoLT	2004	Nöroloji	OB	POS tabanlı	İstatistiksel	Hyp.	Any	—	—
OSEE	2012	Biyoinformatik	SA	POS tabanlı	KB tabanlı	—	Gene Ontology	Various (OBO)	—
OwlExporter	2010	Yazılım	OB	POS tabanlı	KB tabanlı	—	Any	—	—
PIRATES	2010	Yazılım	SA	KEA	Hibrit	—	SEOntology	—	—
SPRAT	2009	Balıkçılık	OB	Örüntüler	TermRaider	Syn. Hyp.	FAO	WordNet	—
TaxoEmbed	2016	Çoklu alan	OB	-	KB tabanlı	Hyp.	Wikidata, YAGO	Wikipedia, BabelNet	
Text2Onto	2005	-	OB	POS/JAPE	İstatistiksel	Hyp. Mer.	—	WordNet	—
TodorLAP	2016	Haberler	TM	DBpedia Spotlight	-	Hyp.	DBpedia	Wikipedia	LDA
TyDI	2010	Biyoteknoloji	OB	-	YaTeA	Syn. Hyp.	—	—	—
VargaCRCH	2014	Mikropostlar	TM	OpenCalais	-	—	DBpedia, Freebase	—	Graph / ML
ZhangYT	2009	Haberler	TE	Manuel	İstatistiksel	—	—	WordNet	—

3.4.3.1 Güven Değeri Hesaplaması için Geliştirilen Yöntem

Bilgi tabanında saklanan, bilgi grafında temsil edilen üçlülerin güven değerinin hesaplanmasına yönelik olarak çoğunluk oylaması, ağırlıklı çoğunluk oylaması ve kaynak güvenilirliğini dikkate alan hibrit model geliştirilmiştir. Kullanılan yönteme ilişkin olarak algoritmaya aşağıda yer verilmiştir.

Algorithm Ensemble_Truth_Discovery_KG:

Input:
- Triples: A list of triples (S, P, O)
- SourceWeights: A dictionary of initial source weights (reliability scores)

Output:
- TrueTriples: The list of discovered true triples

Initialize:
MajorityVotingResults = {}
WeightedMajorityVotingResults = {}
SourceReliabilityResults = {}
TrueTriples = []

 # Majority Voting
for each triple in Triples:
MajorityVotingResults[triple] += 1
end
TrueTriples_MajorityVoting = argmax(MajorityVotingResults)

 # Weighted Majority Voting
for each triple in Triples:
source = triple.source
WeightedMajorityVotingResults[triple] += SourceWeights[source]
end
TrueTriples_WeightedMajorityVoting = argmax(WeightedMajorityVotingResults)

 # Source Reliability Model
for each source in SourceWeights:

```
    calculate reliability based on various metrics (e.g., accuracy, coverage)
    update SourceWeights[source]
  end
  for each triple in Triples:
    source = triple.source
    SourceReliabilityResults[triple] += SourceWeights[source]
  end
  TrueTriples_SourceReliability = argmax(SourceReliabilityResults)

        # Ensemble
        TrueTriples=Combine(TrueTriples_MajorityVoting,
        TrueTriples_WeightedMajorityVoting, TrueTriples_SourceReliability)

  return TrueTriples

  Function Combine(triples1, triples2, triples3):
    # Combine the true triples obtained from the three methods
    return combined_triples
```

Web kaynak güvenilirliğinin hesaplanmasında URL sınıflandırması doğruluk tahmininin temel taşını oluşturmaktadır. URL sınıflandırması, URL'lerden anlamlı özelliklerin çıkarılmasını ve bu özelliklerin makine öğrenimi modellerini eğitmek için kullanılmasını içerir. URL'den çıkarılan özelliklere dayanarak bir kaynağın güvenilirliğine yönelik sınıflandırmanın gerçekleştirilmesi mümkündür.

Bununla birlikte, URL sınıflandırmasının doğal dil işleme, ağ analizi ve makine öğreniminden ileri teknikler gerektiren karmaşık bir görev olduğunu unutmamak önemlidir. Bu teknikler, kaynağın güvenilirliğinin yalnızca URL'nin ötesinde kapsamlı bir şekilde değerlendirilmesine olanak tanımaktadır.

Kaynak güvenilirliğinin hesaplanmasına yönelik geliştirilen alternatifler topluluğu (ensemble methods) içeren hibrit model kapsamında güvenilir-nötr-güvensiz şeklinde etiketlenmiş, her üç kümeye ait 7.000’er alan adının olduğu 21.000 alan adından oluşan veri seti sınıflandırma işlemleri için kullanılmıştır.

URL sınıflandırma işleminde, TF-IDF (Term Frekans-Ters Belgilendirme Frekansı) ve metin benzerliği yöntemi kullanılmıştır. Söz konusu teknik URL'leri önceden tanımlanmış farklı kategorilere ayırmak için kullanılan bir tekniktir. Bu yaklaşım, URL içeriğindeki kelimelerin frekansını kullanarak sayısal bir temsil oluşturur ve ardından URL'ler arasındaki benzerliği ölçerek kategorilerini belirler.

Çizelge 3.6'da ise gerçekliğin keşfi ve güven değerinin hesaplanmasında kullanılan temel terminolojilere yer verilmiştir [112].

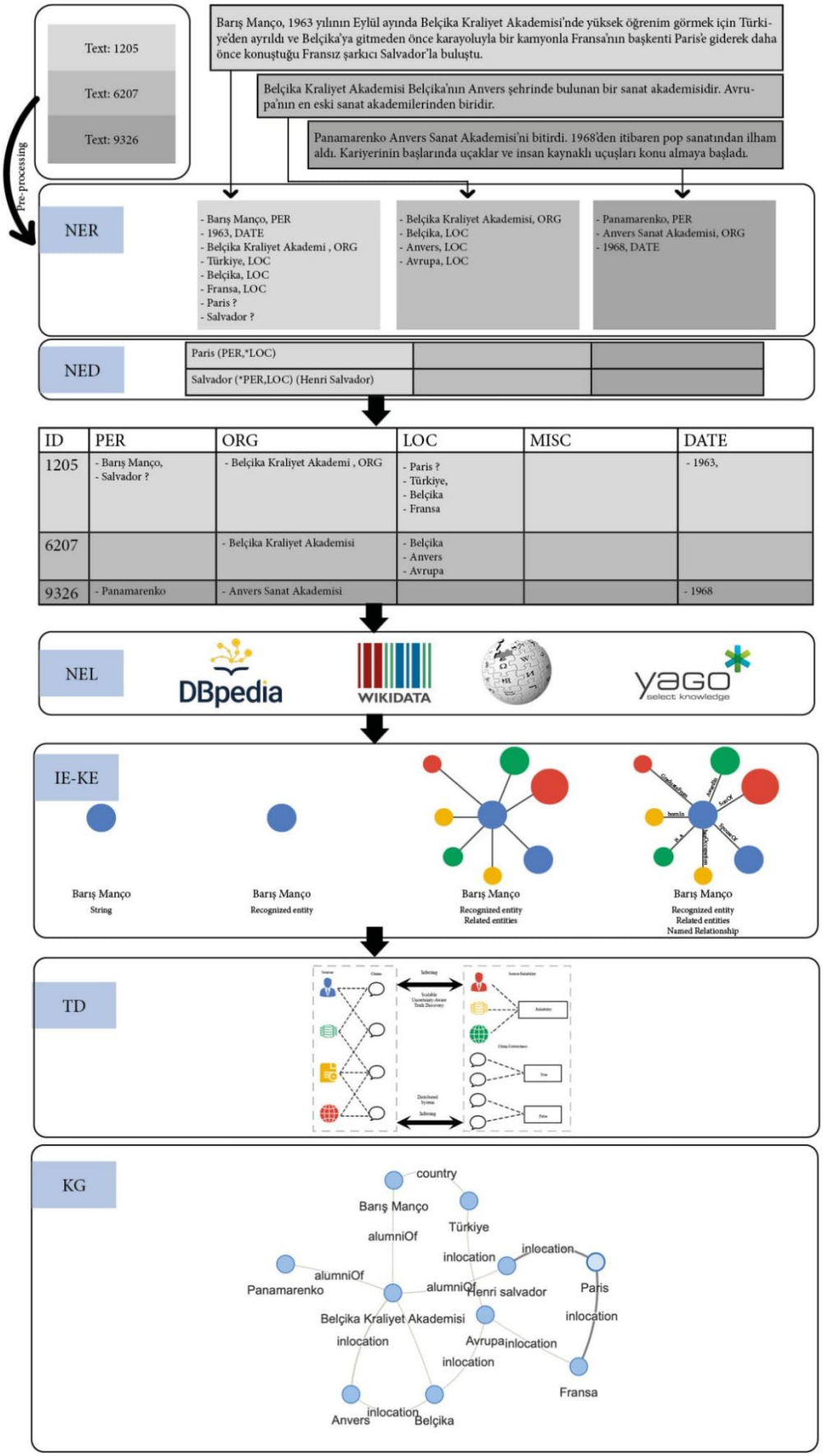
Şekil 3.4'te bilgi çıkarımına yönelik tüm süreçleri içeren Infografiğe yer verilmiştir.

Şekil 3.5'te güven değeri hesaplamasına yönelik genel blok diyagrama [106,112] yer verilmiştir.

Şekil 3.6'da ise iteratif güven değeri hesaplamaya yönelik akış diyagramına yer verilmiştir [106,112].

Çizelge 3.6. Gerçekliğin keşfi hesaplanmasında kullanılan temel terminolojiler

Terminoloji	Açıklama
Gerçek (Fact)	Gerçek, doğru bir iddiadır.
Yanlış iddia (Allegation)	İddia yanlış bir iddiadır.
İddia (Claim)	İddia, tanımlanmış bir kaynak tarafından sağlanan bir değerdir.
Değer (Value)	Değer, gerçek dünyadaki bir varlığın bir özelliğinin değerini ifade eder.
Nesne (Object)	Nesne, gerçek dünyadaki bir varlığı ifade eder. Bir nesne birden fazla özellik, nitelik veya özellik ile tanımlanabilir.
Veri Ögesi (Data item)	Veri öğesi, belirli bir nesne örneği için değerli bir özelliktir.
Kaynak (Source)	Kaynak, veri öğeleri sağlayıcısıdır.
Birbirini Dışlayan Küme (Mutualexclusive set)	Birbirini dışlayan küme, belirli bir nesne özelliği için birden fazla kaynak tarafından talep edilen değerler kümesidir.
Kaynak Güvenilirliği (Source trustworthiness)	Bir kaynağın güvenilirliği (veya doğruluğu veya güvenilirliği), iddialarının güvenirliği göz önüne alındığında kaynağın ne kadar güvenilir olduğunu ölçen bir puandır.
Güven değeri (Value confidence)	Bir değerin güveni, onu iddia eden kaynakların güvenilirliği göz önüne alındığında değerin doğruluğunu ölçen bir puandır.
Doğruluk Etiketi (Truth label)	Doğruluk etiketi, belirli bir nesne özelliği için bir iddianın değerinin doğru veya yanlış olduğunu belirleyen bir Boolean değeridir.
Temel Gerçek (Ground truth)	Temel/Referans doğruluk değeri (bazen altın standart olarak da adlandırılır) gerçekler kümesidir (yani, gerçek dünyada doğru olduğu bilinen nesne özelliklerinin değerleri). Bu küme genellikle manuel olarak doğrulanır/etiketlenir ve doğruluk keşif yöntemlerinin kalite performans değerlendirmesi için kullanılır.

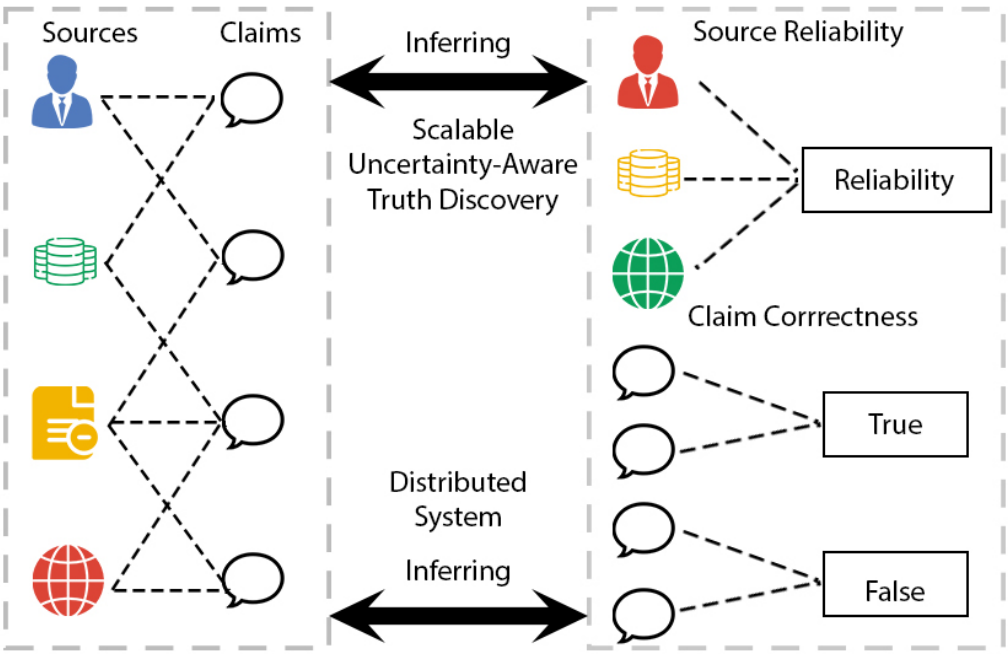


Şekil 3.4. TRPedia IE- KE-TD-KG Infografiği

Şekil 3.4'te Wikipedia üzerinde yer alan Barış Manço konu başlığı üzerindeki metin üzerinde yer alan bir paragraf üzerinden sistemin genel işleyişine yer verilmiştir. Söz konusu infografik kapsamında uçtan uca

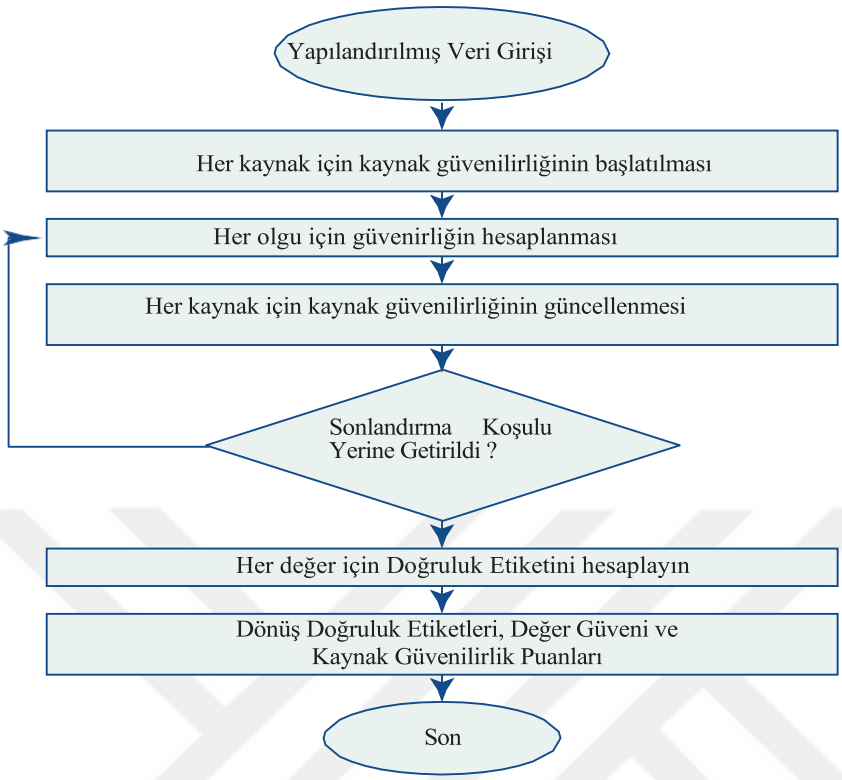
- Metin ayrıştırma işlemini,
- NER süreçlerini (varlıkların tespiti),
- NED süreçlerini (belirsizliklerin giderimi)
- NEL süreçlerini (varlıkların bilgi tabanlarına linklendirilmesi)
- IE (NER-NED-NEL-RE) süreçleri sonrası varlık ve varlığa ait ontolojiler tespit edilerek değerli/anlamlı bilgiler oluşturulmasını,
- Gerçekliğin keşfi ve güven değeri hesaplanmasını,
- Bilgi grafı formatında sub graf gösterimini

bileşenlerini içermektedir.



Şekil 3.5. TRPedia Truth Discovery blok diyagramı

Şekil 3.5'te gerçekliğin keşfi ve güven değeri hesaplamasında kullanılan kavramsal genel metodolojiye yer verilmiştir. Heterojen veri kaynaklarından elde edilen veriler üzerinden çıkarımı yapılan üçlüler farklı iddiaları içerebilir. Söz konusu süreç kaynakların ve iddiaların güvenilirliğini bir bütün olarak ele almayı gerektirmektedir [112].



Şekil 3.6. Güven değeri hesapanmasında iteratif yöntemlere yönelik akış diyagramı

Şekil 3.6'da gerçekliğin keşfi ve güven değeri hesaplamasında kullanılan iteratif yönteme ait akış diyagramına yer verilmiştir [112]. Önermiş olduğumuz modelde her kaynağın güvenilirliğinin hesaplanmasında önceden sınıf değerleri belirlenmiş kaynaklar üzerinden TF-IDF ve benzerlik yaklaşımı kullanılmaktadır.

3.4.4. Geliştirilen Özgün Mimariye Ait Psuedo Code

Şekil 3.2., Şekil 3.3 ve Şekil 3.4'te yer verilen mimariyi gerçekleştirmeye yönelik olarak kaba koda (pseudo code) aşağıda yer verilmiştir. Şekil 3.5'te genel güven değeri keşfi blok diyagramına yer verilmiştir.

KG Construction Pipeline

```
//Initialize an empty Knowledge Graph (KG)
Graph KG

function KnowledgeGraphConstructionPipeline :

    //Initialize a list of URLs (heterogeneous web data sources)
    listURL [] ← null
    For each URL in the list:
    For i to listURL:
    begin
        sourceClassification (TF-IDF):
        //Source Classification Using TF-IDF & Calculate Similarity
            (Black-Untrutsfull, Gray-Nötr, White Trusthfull)
        Identify source type and perform appropriate data extraction:
            dataExtr(url)
            if Source=Webpage then
                webScraping(url)
            else if Source=Database then
                dbScraping(url)
            else if Source=KnowledgeBase then
                KBScraping(url)
            else if Source=socialMedia then
                socialMediaScraping(url)
            else
             Print ("Unknown format")
            Break;
```

If the source is a webpage, perform web scraping to extract data
If the source is a database, query the database to retrieve data
Handle errors and exceptions as needed

Clean and preprocess extracted data:
//Remove irrelevant components (HTML tags, special characters, etc.)
Cleaned_Source ← removeCompenents(sources)
//Tokenize the cleaned text into sentences
listMorfo [] ← morfologicAnalysis(Cleaned_Source)

for i to listMorfo :
begin
//Perform Named Entity Recognition (NER):
$node_i$ ← NERAnalysis(sentences)
//Disambugution entities to known knowledge bases (DBpedia, Wikidata, Yago)
nodei_Disambugution ← NEDAnalysis (paragraph)
// Link entities to knowledge bases (DBpedia, Wikidata, Yago)
nodei_linking ← NELAnalysis(nodei)
relation_i ← REAnalysis (sentences)
fact_i ← nodei, rei, nodej
If PatternCompatible(fact_i) = TRUE then
begin
If TruthDiscoverMajority = TRUE or TruthDiscoverWeightedMajority = TRUE or TruthDiscoverModified = TRUE then
begin
addKG (fact_i)
end
end

end

Post-processing:

Remove isolated nodes

return KG

end

4. DENEYSEL ÇALIŞMALAR

Tezin bu bölümünde yapılan çalışmalarda ele alınan veri kümeleri, bu verilerin saklanma şekli, deneysel çalışmalar için geliştirilmiş yazılım ortamı ve bileşenlerine yer verilmiştir. Ayrıca güven değerlerinin yayılma işlemlerinde tez kapsamında önerilen yaklaşımlar temel alınarak uygulanan deneyler ve elde edilen sonuçlara bu bölüm kapsamında yer verilmiştir.

4.1. Geliştirilen Yazılım Ortamı

Tez kapsamında önerilen modelin deneysel olarak sınanması için yazılım ortamı hazırlanmıştır. Yazılım ortamı hazırlanırken Web ölçekli bir bilgi tabanının otomatik olarak uçtan uca oluşturulması kapsamında yapılması gereken işlemlere yönelik olarak literatürde yer alan yöntemler ve son gelişmeler metodolojik olarak incelenmiştir.

Türkçe diline yönelik olarak yukarıda yer verilen amaçlara uygun olarak kapsamlı ve çoklu etki alanı desteğine sahip çevrim içi, dinamik aşağıda yer verilen katkıları sağlayan özgün bir model ve uygulama platformu geliştirilmiştir.

Heterojen Web veri kaynaklarının doğruluk tespiti bağlamında, bilgi grafları çerçeve işlevi görerek kritik bir rol oynamaktadır. Heterojen web kaynaklarında bulunan varlık ilişkilerini modelleyerek, bağlamı dikkate alan analiz ve varlıklar arası ilişkilerin gösterimini mümkün kılmaktadır. Doğruluk tespiti, verilerin ve kaynakların doğrulanması çelişkili bilgilerin çözülmesi sürecini ifade etmektedir. Bilgi grafları, graf içindeki ilişkileri ve bağlam bilgisini modellemek suretiyle doğruluk tespiti algoritmaları, çelişkileri analizi olanaklı kılmaktadır.

Çalışma kapsamında söz konusu yapıyı oluşturulmasında yazılım dili olarak Php, verilerin saklanması için Mysql kullanılmıştır. Mysql veri tabanı üzerinde veriler üçlü yapısında aralarındaki ilişkileri içerecek şekilde depolanmıştır.

Wikipedia, Wikidata ve Schema.org'un genel ontoloji yapısı arasında ontoloji eşleştirmesi yapılarak, varlıkların düğüm, ontolojilerin ve ilişkilerin kenar görevi gördüğü graf tabanlı bir mimari benimsenmiştir. Üretilen üçlü değerler sistem üzerinde RDF ve JSON formatlarında, her bir varlık ve ilişki ise URI formatında temsil edilmektedir.

NERC aşamasında Sözlük/Kural/CRF/Hibrit yöntemlerinin bir kombinasyonu kullanılmıştır. NED ve Named Entity Linking (NEL) işlemleri için belge benzerliği ve

istatistiksel hesaplamalar dikkate alınmıştır. Varlıklar, yaygın olarak kullanılan Wikipedia/Yago/DbPedia bilgi tabanları ile eşleştirilmiştir.

TRPedia içinde varlıklar ve ontolojiler URI formatında gösterimi sağlanmıştır. İlişki tespiti ve çıkarımı konusunda Bootstrapping, Sequential Role Labeling ve Rule-Based yöntemleri kullanılmıştır.

Tespit edilen varlıklar ve ontolojilere dayalı olarak oluşturulan varlıkların belirsizliğini ve bağlantısını çözme yaklaşımları göz önünde bulundurularak, Üçlü ifade formatına uymayan RDF'ler dikkate alınmamıştır. RDF'lerin anlamlılığı, kalıp ve kural yapılarına uyup uymadıklarına göre kontrol edilmiştir.

Güven değeri hesaplamasında, çoğunluk oylaması, ağırlıklı çoğunluk oylaması ve kaynak güvenilirliğini dikkate alan hibrit model geliştirilerek uygulanmıştır.

4.2. Veri Kümesi

https://dumps.wikimedia.org/trwiki/ adresinde bulunan Türkçe dil konu başlıklarına sahip ~530.000 Wikipedia sayfasından oluşan bir veri kümesi kullanılmış ve orijinal bir bilgi grafiği TRPedia oluşturulmuştur.

Literatürde öne çıkan benzer çalışmalarda Wikipedia'nın yaygın kullanımı sahip olduğu benzersiz avantajlar yönü ile açıklanmaktadır [32]:

- Wikipedia'daki metin öncelikle gerçeklere dayanmaktadır.
- Wikipedia, çeşitli alanlarda gerçek dünyaya ait varlıklar hakkında belgeler içeren geniş bir kapsama sahiptir.
- Wikipedia'daki makaleler, DBpedia, Freebase, Wikidata, YAGO vb. dahil olmak üzere çeşitli KB'lerde tanımladıkları varlıklara doğrudan bağlanabilir.
- Wikipedia'da varlıklardan bahsedilmesi genellikle o varlıkla ilgili makaleye bir bağlantı sağlar, böylece çeşitli bağlamlarda etiketlenmiş varlık bahsi örnekleri ve ilişkili bağlantı metni örnekleri sağlar.
- Wikipedia metinsel özelliklerin yanı sıra terim frekansları ve eş-oluşumları, çeşitli daha zengin özelliklere Wikipedia'dan ulaşma imkânı sağlamaktadır.

Veri setine ait istatistiklere Çizelge-4.1’ de yer verilmiştir.

Çizelge 4.1. Deneysel veri seti özellikleri

Veri Seti İstatistikleri	**Sayı**
Toplam Konu Başlığı Sayısı	529.648
Toplam Cümle Sayısı	4.521.477
Tekil Kelime Sayısı	612.298
Tekil Kök Kelime Sayısı	177.936

4.3. Deney Tasarımı

Geliştirilen özgün modelin başarımını ölçmek amacı ile Farber [69], Pillai [70] ve Wu [71] tarafından yapılan bilgi graflarının karşılaştırmasına yönelik yapılan araştırmalarda kullanılan yöntem dikkate alınarak Wikipedia üzerinde yer alan 529.648 Türkçe diline ait konu başlığı üzerinden bir çalışma yapılmıştır.

Diğer taraftan bilgi grafında üretilen üçlülere yönelik deneysel çalışmaların yapılabilmesi amacı ile geçerleme kümesine yanlış üçlülerin eklenmesi gerekmektedir. Yanlış üçlülerin oluşturulmasında PTrustE’nin [129] ve DSKRL [130] yaklaşımı göz önünde bulundurulmuştur.

Bu kapsamda 100 Wikipedia konu başlığına ait veri seti üzerinden üretilen 1000 adet üçlünün yer aldığı veri kümesinden (h, r, s) üçlülerinden oluşan ilişki kümesinden rastgele r ilişkisi seçilerek (h, r, s’) yanlış (corrupted) üçlüsü oluşturulmuştur. Oluşturulan üçlüler veri kümesine sıra gözetilmeksizin rastgele eklenmiştir. Yanlış üçlü oluşturma oranı farklı deney tasarımları için veri seti kümesi %20, %40 ve %60 oranlarında uygulanarak yanlış verileri içerecek şekilde veri seti kümesi zenginleştirilmiştir.

Farklı hatalı üçlü oranlarının üretilerek hata oranının artışına göre performans sonuçlarının ölçülmesi amaçlanmıştır. Hatalı üçlülerin yer aldığı örnek veri setine Çizelge 4.2’de yer verilmiştir.

Çizelge 4.2.
Çizelge 4.2. Deneysel veri seti

Hatalı/Gerçek	E1	R	E2
Gerçek	Fatin Rüştü Zorlu	ölümyeri	Bursa
Gerçek	Seydişehir	posta kodu	42360
Gerçek	Seydişehir	plaka kodu	42
Hatalı	Ahmet Altan	doğumyeri	İzmir
Hatalı	Fazıl Say	doğumyeri	İstanbul
Gerçek	Doğan Aksan	doğumyeri	İzmir
Hatalı	Ümit Özat	doğumyeri	İstanbul
Gerçek	Nikola Tesla	ölümsebebi	Kalp yetmezliği
Hatalı	Bahadır Baruter	doğumyeri	İstanbul
Hatalı	Turgut Uyar	doğumyeri	İstanbul
Gerçek	Ömer Nasuhi Bilmen	ölümyeri	Ankara
Gerçek	Latife Tekin	doğumyeri	Kayseri
Gerçek	Dâvûd-i Kayserî	doğumyeri	Kayseri
Gerçek	Kenan Çoygun	doğumyeri	Bursa
Gerçek	Necati Ateş	doğumyeri	İzmir
Gerçek	Emre Aşık	doğumyeri	Bursa
Gerçek	Çağdaş Atan	doğumyeri	İzmir
Gerçek	Mustafa Denizli	doğumyeri	İzmir
Gerçek	Ramazan Kurşunlu	doğumyeri	İzmir
Gerçek	Recep Adanır	doğumyeri	Ankara
Gerçek	Sertab Erener	doğumyeri	İstanbul
Hatalı	Sertab Erener	doğumyeri	Ankara
Gerçek	İhsan Eliaçık	doğumyeri	Kayseri
Hatalı	Ali Saydam	doğumyeri	İstanbul
Gerçek	Ali Saydam	doğumyeri	Türkiye
Gerçek	Ali Saydam	meslek	Gazeteci
Gerçek	Ali Saydam	meslek	yazar
Gerçek	Ali Saydam	meslek	iletişim profesyoneli
Gerçek	Ali Saydam	meslek	öğretim görevlisi
Hatalı	Ali Saydam	meslek	mühendis
Hatalı	Ali Saydam	doğumyeri	İstanbul

Çizelge 4.1'de yer alan kırmızı ile belirginleştirilmiş üçlüler, yanlış değerleri ifade etmekte olup, PTrustE'nin [129] ve DSKRL [130] yaklaşımı dikkate alınarak veri seti %20, %40, %60 oranında yanlış üçlü değerleri ile zenginleştirilmiştir.

4.4. Performans Değerlendirme Ölçütleri

KG'lerin oluşturulması sürecinde, varlıklar ve/veya ilişkiler açısından yanlış gerçekler yakalanabilir. Bu süreç, özellikle karma nitelikli veri kaynaklarından elde edilen bilgiler açısından hatalara açıktır.

Veri kalitesini korumaya hem yeni hem de yerleşik gerçeklere dayalı olarak Ortalama Karşılıklı Sıralama (MRR-Mean Reciprocal Rank), Doğruluk (Accuracy), Kesinlik (Precision), Geri Çağırma (Recall) ve F Ölçümü (F-Measure) gibi ölçütler kullanılarak değerlendirmeler yapıldığı görülmektedir.

Bununla birlikte bilgi çıkarımı, bilgi tabanları ve bilgi grafları, isimlendirilmiş varlık tespiti ve ilişki çıkarımının başarım performansını ölçmeye yönelik olarak, Precision, Recall ve F-Measure metriklerinin kullanıldığı görülmektedir. Söz konusu değerlerin hesaplanmasında kullanılan True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) kavramlarına aşağıda yer verilmiştir.

Doğru Pozitif (TP): Sınıfı Positive olan ve sistem tarafından da doğru şekilde Positive olarak sınıflandırılanları ifade etmektedir.

Doğru Negatif (TN): Sınıfı Negative olan ve sistem tarafından da doğru şekilde Negative olarak sınıflandırılanları ifade etmektedir.

Yanlış Pozitif (FP): Sınıfı Negative olan ancak sistem tarafından yanlış şekilde Positive olarak sınıflandırılanları ifade etmektedir.

Yanlış Negatif (FN): Sınıfı Positive olan ancak sistem tarafından yanlış şekilde Negative olarak sınıflandırılanları ifade etmektedir.

Precision (Kesinlik) ve Recall (Duyarlılık), sınıflandırma problemleri gibi makine öğrenimi ve bilgi geri çağırma uygulamalarında kullanılan iki önemli metriktir. Genellikle bir modelin ne kadar iyi çalıştığını ölçmek için kullanılmaktadır.

Precision, modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu ölçer. Yani, modelin yaptığı pozitif tahminlerin ne kadar "kesin" olduğunu belirtir. Sistemin bulamadığı veya çıkarım yapamadığı değerlerin sayısına bakılmaz.

Recall ise, sistemin çıkarması gereken verilerin ne kadarını çıkarabildiğinin oranını temsil etmektedir. Gerçekte pozitif olan tüm örneklerin ne kadarının model tarafından pozitif olarak sınıflandırıldığını ölçer. Başka bir deyişle, modelin pozitif sınıfı ne kadar iyi "hatırladığı" ile ilgili bir ölçümdür. Recall'da sistemin ne kadar yanlış çıkarımda bulunduğuna bakılmamaktadır.

Precision veya tek başına recall başarı ölçümü için yeterli değildir, F-Score hem precision ve hem de recall'u göz önüne alan performans ölçüm metriğidir.

Bu tanımlar ışığında precision, recall, F-Measure başarı ölçüm metriklerinin matematiksel olarak hesaplanmasına yönelik ifadelere, Eş. 4.1, Eşitlik 4.2. ve Eş. 4.3'te verilmiştir.

$$Precision = \frac{TP}{TP+FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.2}$$

$$F - Measure = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall} \tag{4.3}$$

4.5. Deneysel Sonuçlar

Bu araştırma kapsamında, heterojen Web veri kaynaklarından bilgilerin toplanması, doğrulanması ve sorgulanmasını sağlamak amacı ile doğal dil işleme, bilgi çıkarımı, bilgi grafları, Semantik Web, bağlantılı veri alanlarında literatürde öne çıkan çalışmalar incelenerek, orijinal ve özgün bir model geliştirilmiş, genel amaçlı TRPedia yazılım platformu ve bilgi grafi sunulmuştur.

Geliştirilen mimari kapsamında elde edilen genel karşılaştırmalı istatistiksel sonuçlara https://www.trpedia.org/search2/statistics.php adresinden erişim sağlanmaktadır.

Farber [69], Pillai [70] ve Wu [71] tarafından bilgi graflarının perforemans karşılaştırmasına yönelik yapılan araştırmalarda kullanılan karşılaştırma ölçütleri dikkate alınarak Wikipedia'ki Türkçe konu başlıklarından geliştirilen özgün model kapsamında yapılan yapılan bilgi çıkarımları (üçlüler), Dbpedia ve YAGO bilgi tabanları üzerinde yer alan çıkarımlar ile karşılaştırılmıştır.

Söz konusu karşılaştırmalarda, TRPedia üzerinden elde edilen üçlü çıkarımların, Wikipedia infobox verilerine kıyasla %6 oranında, DBpedia'ya kıyasla 4,28 kat daha fazla ve YAGO bilgi tabanına kıyasla 11,01 kat, bilgi çıkarımına yönelik daha iyi sonuçlar sunduğu görülmüştür. Çizelge 4.2'de bilgi çıkarımına yönelik karşılaştırmalı sonuçlara yer verilmiştir.

Çizelge 4.2. Bilgi çıkarımı karşılaştırmalı sonuçları

KG/ Parameters	**Wikipedia**	**DBpedia**	**Yago**	**Trpedia**
Toplam Konu Başlığı Sayısı	529.648	270.523	302.346	529.648
Toplam NER Sayısı	116.312	83.640	78.618	124.452
Toplam NER – Person Sayısı	100.711	72.869	66.103	107.760
Toplam NER – Location Sayısı	8.304	6.497	6.953	8.885
Toplam NER – Organization Sayısı	7.297	4.274	5.562	7.807
Toplam Triple Sayısı	5.415.502	1.353.875	526.780	5.794.587

Söz konusu başarımı şu şekilde açıklamak mümkündür, DBPedia ve YAGO bilgi tabanları sadece Wikipedia Infobox verileri üzerinden çıkarım yapmaktadır. TRPedia yazılım platformu kapsamında Wikipedia konu başlıklarına ait içerikler, infoboxlar ve metinsel içerik ve infoboxlarda yer alan varlıklar üzerinden eksik ilişkilerin de tamamlanması dahil olmak üzere çift yönlü olarak kontrol edilmiş, eksik ilişkilerin tamamlanması sağlanmıştır. Yine bu kapsamda doğal dil işleme ve NER üzerine uygulanan yöntemlerin daha iyi sonuç üretmektedir.

Büyük ölçekli Web verilerinin doğrulanması ve kalitesiyle ilgili araştırmalarda veri kalitesi ölçütlerinin belirlenmesi, kendine özgü bir dizi zorluğu beraberinde getirmektedir. KG'ları bağlamında kalitenin ne olduğunu tanımlamak, önemli ölçüde veri tüketicileri, veri kaynağı ve kullanım durumu için kalite boyutlarının tanımlanıp tanımlanmadığına bağlı olacaktır.

Literatürde bağlantılı veri, bilgi grafları ve bilgi çıkarımı alanında öne çıkan çalışmalar incelendiğinde doğruluk, tutarlılık, tamlık, zamanlılık, güvenilirlik ve kullanılabilirlik boyutlarının öne çıktığı görülmektedir [65,72,74,75,76]. Geliştirilen modelin bilgi çıkarımına yönelik performans sonuçlarına Çizelge 4.3'te yer verilmiştir.

Çizelge 4.3. Bilgi çıkarımı performans sonuçları

Modül	**Precision**	**Recall**	**F-Measure**
Knowledge Extraction - NER	0,81	0,79	0,80
Knowledge Extraction – RE	0,76	0,71	0,73
Overall -KE	0,78	0,75	0,76

Tez kapsamında yer verilen hususlara ilişkin deneysel sonuçlara https://trpedia.org/search2 sayfasında yer verilmiştir.

Söz konusu sayfa üzerinde

- Wikipedia Türkçe konu başlıkları üzerinden oluşturulan bilgi tabanı ve bilgi grafı,
- Bilgi grafı ve Infobox mimarisi
- Oluşturulan mimari ve mevcut veri seti üzerinden unutulma hakkının uygulanmasına yönelik demo,
- Açık kaynaklarda yer alan Türk İlça sektörüne ilişkin bilgiler üzerinden oluşturulan bilgi tabanı-graf demo (Ör: https://www.trpedia.org/demo-ilac/?id=e-222)
- Telekomunikasyon graf analitiği (https://trpedia.org/test5/)

Yer almaktadır.

Oluşturulan mimari aynı zamanda varlıklar arasındaki anomalileri kesişimlerin gösterimine olanak sağlamaktadır (https://www.trpedia.org/test7/).

İlerleyen sayfalarda trpedia.org/search2 adresinden gelen sayfa üzerinde Barış Manço aratıldıldığında ilgili sayfada tez kapsamında yer verilen anlatılan tüm içeriğe yer verilmiştir:

- Doğal dil işleme,
- Morfolojik analiz,
- Bilgi çıkarımı aşamaları
    - NER
    - NED
    - RE
- Karşılaştırmalı sonuçlar
- Bilgi çıkarımı sonuçları
- URI,
- İstatistiksel Sonuçlar,
- Bilgi çıkarımı aşamaları (NER, NED, RE)

Barış Manço'ya ait Wikipedia sayfası üzerinde yer alan bilgiler ve Infobox verileri (https://tr.wikipedia.org/wiki/Bar%C4%B1%C5%9F_Man%C3%A7o) üzerinden elde edilen veriler, Şekil 3.3'te yer verilmiş TRPedia blok gösterimi üzerinde yer alan; *verilerin toplanması, bilgi çıkarımı, bilgi füzyonu, güven değeri hesaplanması, bilgilerin graf gösterimi* süreçlerini içerecek şekilde ilerleyen sayfalarda yer verilmiştir.

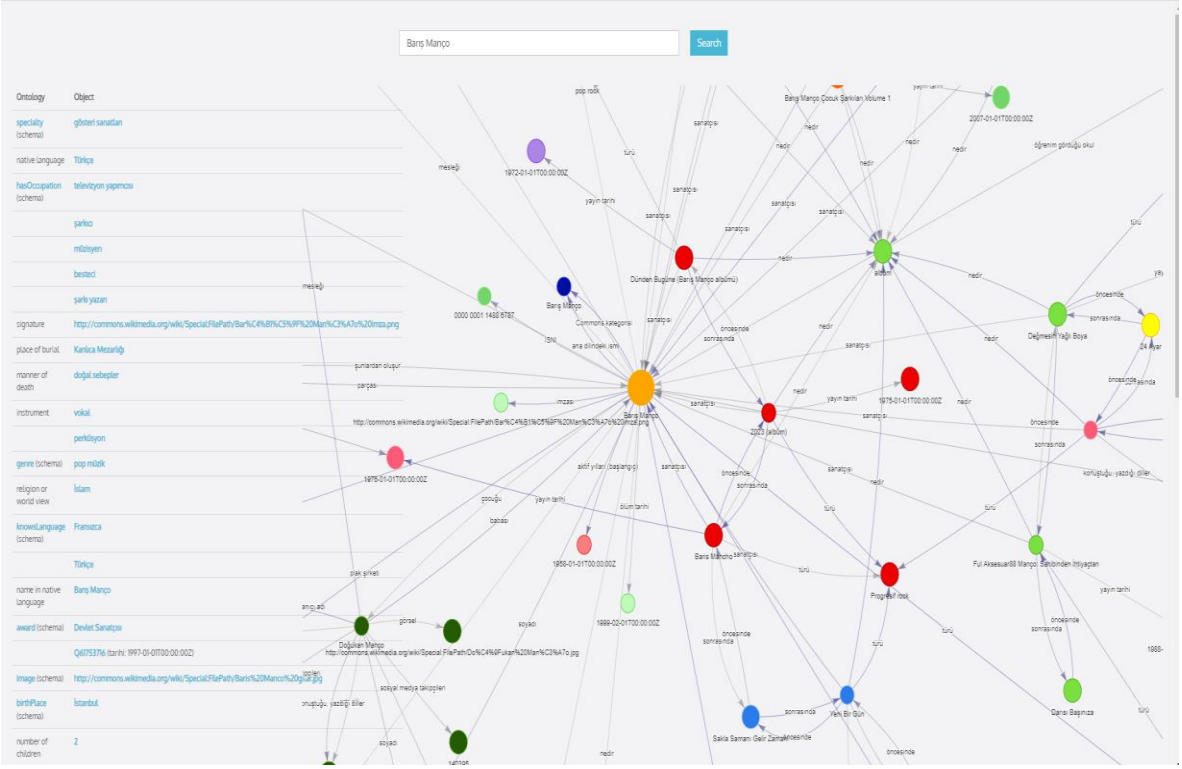
https://trpedia.org/search2/ adresi üzerinde arama kısmında “Barş Manço” yazıldığında

- Graph & Infobox
- Wikipedia
- Yago
- DBPedia
- Haberler
- Sosyal Medya
- Morfolojik Analiz
- NER- RE
- Karşılaştırmalı Sonuçlar
- Çıkarımlar
- URI

Sekmeleri yer almaktadır.

Söz konusu resimde Barış Manço’nun sahip olmuş olduğu ontolojiler üzerinden birinci ve ikinci dereceden ilişkili olduğu diğer varlıklara ait bağlantıları da gösterimi sağlanmıştır.

Resim 4.1'de Barış Manço'ya ait Wikipedia, Wikidata, DBPedia, Yago ve schema.org üzerinden yapılan bilgi çıkarımları ve ontoloji eşleştirmeleri sonucu üretilen üçlülerin infobox ve bilgi grafı gösterimine yer verilmiştir.

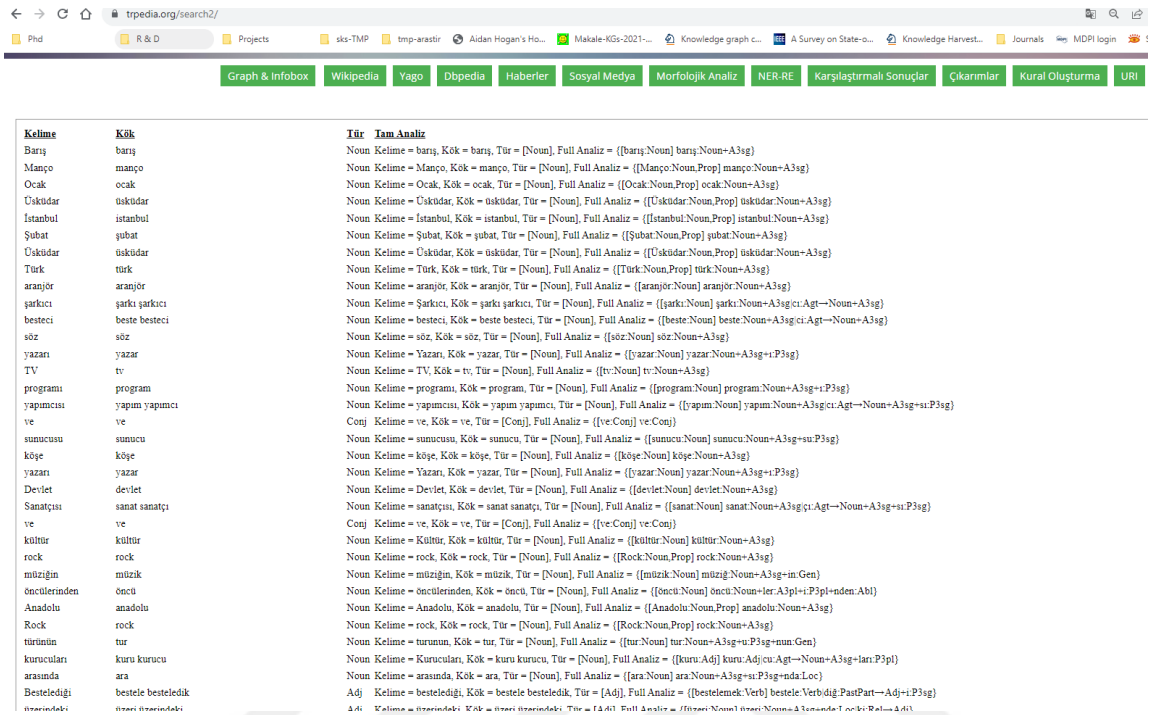


Resim 4.1 Wikipedia Barış Manço Graf Gösterimi

Resim 4.1'de Barış Manço'ya ait Wikipedia, Wikidata, DBPedia, Yago ve schema.org üzerinden yapılan bilgi çıkarımları ve ontoloji eşleştirmeleri sonucu üretilen üçlülerin infobox ve bilgi grafı gösterimine yer verilmiştir.

Barış Manço'ya ait Wikipedia, Wikidata, DBPedia, YAGO ve schema.org üzerinden yapılan bilgi çıkarımları ve ontoloji eşleştirmeleri sonucu üretilen üçlülerin infobox ve bilgi grafı gösterimine yer verilmiştir. Yine aynı zamanda Barış Manço'nun sahip olmuş olduğu ontolojiler üzerinden birinci ve ikinci dereceden ilişkili olduğu diğer varlıklara ait bağlantıları da gösterimi sağlanmıştır.

Resim 4.2’de Wikipedia Türkçe konu başlığı üzerinden alınan Barış Manço’ya ait sayfa içeriğinin morfolojik analiz işlemi sonuçlarına yer verilmiştir.



Kelime	Kök	Tür	Tam Analiz
Barış	barış	Noun	Kelime = barış, Kök = barış, Tür = [Noun], Full Analiz = {[barış:Noun] barış:Noun+A3sg}
Manço	manço	Noun	Kelime = Manço, Kök = manço, Tür = [Noun], Full Analiz = {[Manço:Noun,Prop] manço:Noun+A3sg}
Ocak	ocak	Noun	Kelime = Ocak, Kök = ocak, Tür = [Noun], Full Analiz = {[Ocak:Noun,Prop] ocak:Noun+A3sg}
Üsküdar	üsküdar	Noun	Kelime = Üsküdar, Kök = üsküdar, Tür = [Noun], Full Analiz = {[Üsküdar:Noun,Prop] üsküdar:Noun+A3sg}
İstanbul	istanbul	Noun	Kelime = İstanbul, Kök = istanbul, Tür = [Noun], Full Analiz = {[İstanbul:Noun,Prop] istanbul:Noun+A3sg}
Şubat	şubat	Noun	Kelime = Şubat, Kök = şubat, Tür = [Noun], Full Analiz = {[Şubat:Noun,Prop] şubat:Noun+A3sg}
Üsküdar	üsküdar	Noun	Kelime = Üsküdar, Kök = üsküdar, Tür = [Noun], Full Analiz = {[Üsküdar:Noun,Prop] üsküdar:Noun+A3sg}
Türk	türk	Noun	Kelime = Türk, Kök = türk, Tür = [Noun], Full Analiz = {[Türk:Noun,Prop] türk:Noun+A3sg}
aranjör	aranjör	Noun	Kelime = aranjör, Kök = aranjör, Tür = [Noun], Full Analiz = {[aranjör:Noun] aranjör:Noun+A3sg}
şarkıcı	şarkı şarkıcı	Noun	Kelime = Şarkıcı, Kök = şarkı şarkıcı, Tür = [Noun], Full Analiz = {[şarkı:Noun] şarkı:Noun+A3sg|cı:Agt→Noun+A3sg}
besteci	beste besteci	Noun	Kelime = besteci, Kök = beste besteci, Tür = [Noun], Full Analiz = {[beste:Noun] beste:Noun+A3sg|ci:Agt→Noun+A3sg}
söz	söz	Noun	Kelime = söz, Kök = söz, Tür = [Noun], Full Analiz = {[söz:Noun] söz:Noun+A3sg}
yazarı	yazar	Noun	Kelime = Yazarı, Kök = yazar, Tür = [Noun], Full Analiz = {[yazar:Noun] yazar:Noun+A3sg+ı:P3sg}
TV	tv	Noun	Kelime = TV, Kök = tv, Tür = [Noun], Full Analiz = {[tv:Noun] tv:Noun+A3sg}
programı	program	Noun	Kelime = programı, Kök = program, Tür = [Noun], Full Analiz = {[program:Noun] program:Noun+A3sg+ı:P3sg}
yapımcısı	yapım yapımcı	Noun	Kelime = yapımcısı, Kök = yapım yapımcı, Tür = [Noun], Full Analiz = {[yapım:Noun] yapım:Noun+A3sg|cı:Agt→Noun+A3sg+sı:P3sg}
ve	ve	Conj	Kelime = ve, Kök = ve, Tür = [Conj], Full Analiz = {[ve:Conj] ve:Conj}
sunucusu	sunucu	Noun	Kelime = sunucusu, Kök = sunucu, Tür = [Noun], Full Analiz = {[sunucu:Noun] sunucu:Noun+A3sg+su:P3sg}
köşe	köşe	Noun	Kelime = köşe, Kök = köşe, Tür = [Noun], Full Analiz = {[köşe:Noun] köşe:Noun+A3sg}
yazarı	yazar	Noun	Kelime = Yazarı, Kök = yazar, Tür = [Noun], Full Analiz = {[yazar:Noun] yazar:Noun+A3sg+ı:P3sg}
Devlet	devlet	Noun	Kelime = Devlet, Kök = devlet, Tür = [Noun], Full Analiz = {[devlet:Noun] devlet:Noun+A3sg}
Sanatçısı	sanat sanatçı	Noun	Kelime = sanatçısı, Kök = sanat sanatçı, Tür = [Noun], Full Analiz = {[sanat:Noun] sanat:Noun+A3sg|çı:Agt→Noun+A3sg+sı:P3sg}
ve	ve	Conj	Kelime = ve, Kök = ve, Tür = [Conj], Full Analiz = {[ve:Conj] ve:Conj}
kültür	kültür	Noun	Kelime = Kültür, Kök = kültür, Tür = [Noun], Full Analiz = {[kültür:Noun] kültür:Noun+A3sg}
rock	rock	Noun	Kelime = rock, Kök = rock, Tür = [Noun], Full Analiz = {[Rock:Noun,Prop] rock:Noun+A3sg}
müziğin	müzik	Noun	Kelime = müziğin, Kök = müzik, Tür = [Noun], Full Analiz = {[müzik:Noun] müzi|ğ:Noun+A3sg+in:Gen}
öncülerinden	öncü	Noun	Kelime = öncülerinden, Kök = öncü, Tür = [Noun], Full Analiz = {[öncü:Noun] öncü:Noun+ler:A3pl+i:P3pl+nden:Abl}
Anadolu	anadolu	Noun	Kelime = Anadolu, Kök = anadolu, Tür = [Noun], Full Analiz = {[Anadolu:Noun,Prop] anadolu:Noun+A3sg}
Rock	rock	Noun	Kelime = rock, Kök = rock, Tür = [Noun], Full Analiz = {[Rock:Noun,Prop] rock:Noun+A3sg}
türünün	tur	Noun	Kelime = turunun, Kök = tur, Tür = [Noun], Full Analiz = {[tur:Noun] tur:Noun+A3sg+u:P3sg+nun:Gen}
kurucuları	kuru kurucu	Noun	Kelime = Kurucuları, Kök = kuru kurucu, Tür = [Noun], Full Analiz = {[kuru:Adj] kuru:Adj|cu:Agt→Noun+A3sg+ları:P3pl}
arasında	ara	Noun	Kelime = arasında, Kök = ara, Tür = [Noun], Full Analiz = {[ara:Noun] ara:Noun+A3sg+sı:P3sg+nda:Loc}
Bestelediği	bestele besteledik	Adj	Kelime = bestelediği, Kök = bestele besteledik, Tür = [Adj], Full Analiz = {[bestelemek:Verb] bestele:Verb|diğ:PastPart→Adj+i:P3sg}

Resim 4.2 Barış Manço Wikipedia Sayfası Morfolojik Analiz

Resim 4.2.’de Wikipedia Türkçe konu başlığı üzerinden alınan Barış Manço’ya ait sayfa içeriği paragraf, cümle ve kelimeler bazında ayrıştırılarak, Zemberek programına girdi olarak verilmiştir.

Resim 4.3'te Wikipedia Türkçe konu başlığı üzerinden alınan Barış Manço'ya ait sayfa içeriğinin bilgi çıkarımı (NER, NED, RE) tekniklerine ait bileşenlere tabi tutulması sonrası sonuçlara ve sayfa içeriğine yönelik olarak yer verilmiştir

Resim 4.3 Barış Manço Wikipedia sf üzerinden NER-NED-RE ve İstatistikler

Resim 4.3'te Wikipedia Türkçe konu başlığı üzerinden alınan Barış Manço'ya ait sayfa içeriğinde yer alan içerik bilgi çıkarımı teknikleri uygulanarak NER, NED, NEL ve RE aşamaları yerine getirilmiş, analiz işlemi sonuçlarının sayfa üzerinde gösterimi sağlanmıştır. Aynı zamanda söz konusu sayfaya yönelik istatistiksel sonuçlara sağ üstte yer verilmiştir.

Resim 4.4'te yapılan bilgi çıkarımı sonucu üretilen üçlülerin literatürde öne çıkan bilgi tabanları ve Wikipedia sayfası üzerinden karşılaştırmalı sonuçlarına yer verilmiştir.

Object	Property (schema)	Trpedia	Wikipedia	Dbpedia	Yago
Barış Manço	ad	Barış Manço	Barış Manço	var	var
Barış Manço	artalan	tek şarkıcı	tek şarkıcı	-	-
Barış Manço	çalgı	Perküsyon	Perküsyon	-	-
		piyano	piyano	-	-
		klavye	klavye	-	-
		gitar	gitar	-	-
Barış Manço	çocuklar	Doğukan Manço	Doğukan Manço	-	-
		Batıkan Manço	Batıkan Manço	-	-
Barış Manço	diğer adları	Barış Ağabey	Barış Ağabey	-	-
		Uzun Saçlı Dev Adam	Uzun Saçlı Dev Adam	-	-
		Barış Çelebi	Barış Çelebi	-	-
Barış Manço	doğum adı	Tosun Yusuf Mehmet Barış Manço	Tosun Yusuf Mehmet Barış Manço	var	var
Barış Manço	doğum tarihi	1943.1.2	1943.1.2	-	-
Barış Manço	birthPlace	Üsküdar	Üsküdar	var	-
		İstanbul		-	-
		İstanbul	İstanbul	-	-
		Türkiye	Türkiye	-	-
Barış Manço	spouse	Lale Manço	Lale Manço	var	-
Barış Manço	etkin yıllar	1958-1999	1958-1999	-	-
Barış Manço	ilişkili hareketler	Moğollar	Moğollar	var	-
		Kurtalan Ekspres	Kurtalan Ekspres	-	var
		Kaygısızlar	Kaygısızlar	-	-
Barış Manço	imza	Barış Manço imza.png	Barış Manço imza.png	var	-
Barış Manço	hasOccupation	Şarkıcı	Şarkıcı	-	-

Resim 4.4 Barış Manço Wikipedia Sf. Karşılaştırmalı Sonuçlar

Diğer taraftan Resim 4.4'te' gri ile seçili üçlülerin olduğu satırlar graf tabanlı bir varlığa ait unutulma hakkının uygulanmasına yönelik olarak unutulma hakkının uygulanmak amacı ile seçilen ontoloji ve değerlerin yer aldığı görülmektedir $(s, \boldsymbol{p}, \boldsymbol{o})$.

Resim 4.5'te Wikipedia Türkçe konu başlığı üzerinden alınan Barış Manço'ya ait sayfa içeriğinin bilgi çıkarım sonucu üretilen üçlülerin TRPedia üzerinden URI ile gösterimine yer verilmiştir.

Barış Manço

<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/birthPlace> <https://trpedia.org/search2/?q=Üsküdar> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/birthPlace> <https://trpedia.org/search2/?q=İstanbul> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/spouse> <https://trpedia.org/search2/?q=Lale Manço> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/hasOccupation> <https://trpedia.org/search2/?q=Şarkıcı> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/hasOccupation> <https://trpedia.org/search2/?q=müzisyen> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/hasOccupation> <https://trpedia.org/search2/?q=besteci> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/brand> <https://trpedia.org/search2/?q=Sayan Plak> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/photo> <https://trpedia.org/search2/?q=Barış Manço cropped.JPG> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=Anadolu rock> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=pop müzik> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=erkek> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=vokal> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=Fransızca> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=Türkçe> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=Galatasaray Lisesi> .
<https://trpedia.org/search2/?q=Barış Manço> <http://schema.org/genre> <https://trpedia.org/search2/?q=Kadıköy> .
<https://trpedia.org/search2/?q=Film Gibi> <http://schema.org/composer> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Baba Bizi Eversene> <http://schema.org/agent> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Murat Ertel> <http://schema.org/brand> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Mançoloji> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Barış Manço Live in Japan> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Müsaadenizle Çocuklar> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Mega Manço> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Darısı Başınıza> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=2023 (albüm)> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Baris Mancho> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Yeni Bir Gün> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Sözüm Meclisten Dışarı> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Estağfurullah... Ne Haddimize!> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=24 Ayar> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Değmesin Yağlı Boya> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .
<https://trpedia.org/search2/?q=Ful Aksesuar'88 Manço: Sahibinden İhtiyaçtan> <http://schema.org/creator> <https://trpedia.org/search2/?q=Barış Manço> .

Resim 4.5 Barış Manço Wikipedia sf. URI Gösterimi

Resim 4.5'te yer alan üçlülerin bağlantılı veri standartlarına uygun olarak URI formatında (s, p, o) formatında erişimin sağlanabildiği görülmektedir.

Diğer taraftan RDF ve JSON formatında örnek sonuçlara aşağıda yer verilmiştir.

RDF Format

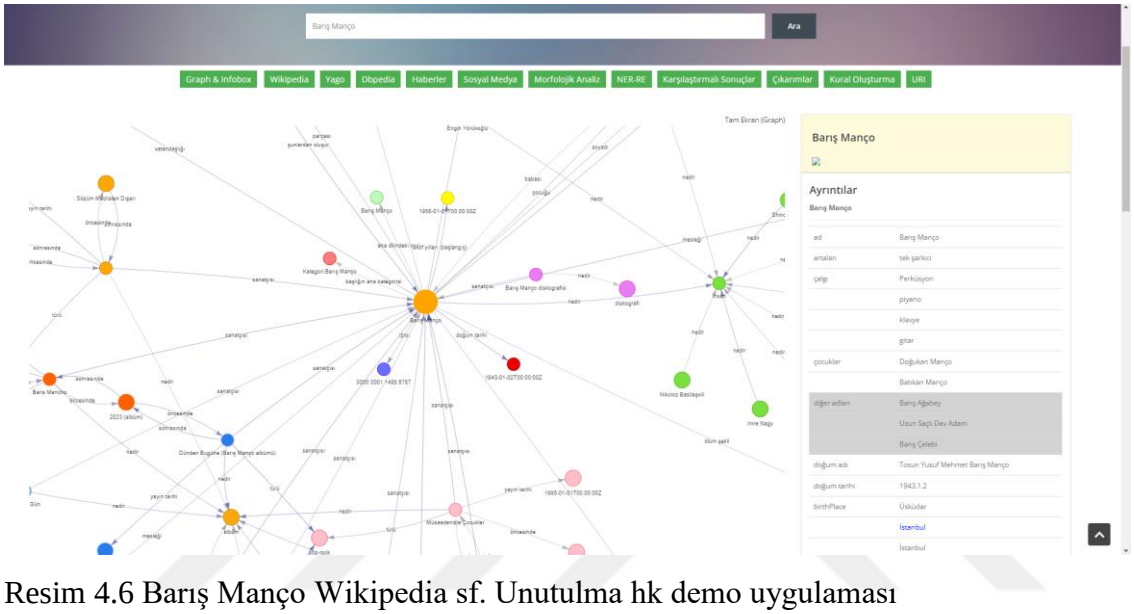
```
@prefix schema: <http://schema.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix trpedia: <http://trpedia.org/resource/> .

trpedia:Barış_Manço a schema:Person ;
    schema:name "Barış Manço"^^xsd:string ;
    schema:birthDate "1943-01-02"^^xsd:date ;
    schema:deathDate "1999-02-01"^^xsd:date ;
    schema:birthPlace trpedia:Istanbul ;
    schema:nationality "Turkish"^^xsd:string ;
    schema:occupation schema:Musician ;
    schema:description "Barış Manço, müzisyen."^^xsd:string .
```

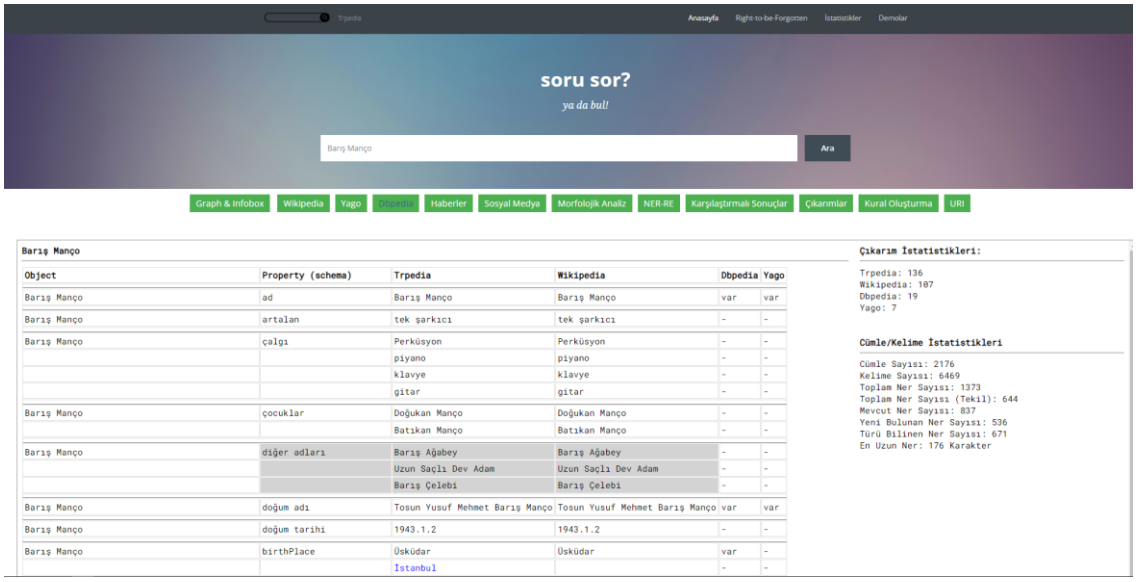
JSON Format

```
{
  "@context": "http://schema.org",
  "@type": "Person",
  "name": "Barış Manço",
  "birthDate": "1943-01-02",
  "deathDate": "1999-02-01",
  "birthPlace": {
    "@type": "Place",
    "name": "Istanbul",
    "url": "http://trpedia.org/resource/Istanbul"
  },
  "nationality": "Turkish",
  "occupation": {
    "@type": "Occupation",
    "name": "Musician"
  },
  "description": " Barış Manço, müzisyen, "
}
```

Resim 4.6 ve Resim 4.7'de varlığa ait üçlü değer/ontoloji ya da kendisinin de gösteriminin pasif edilmesini sağlayan unutulma hakkı gösterimine yönelik demo sayfasına yer verilmiştir. Graf tabanlı bir varlığa ait unutulma hakkının uygulanmasına yönelik olarak unutulma hakkının uygulanmak amacı ile seçilen ontoloji ve değerlerin yer aldığı görülmektedir (s, p, o).



Resim 4.6 Barış Manço Wikipedia sf. Unutulma hk demo uygulaması



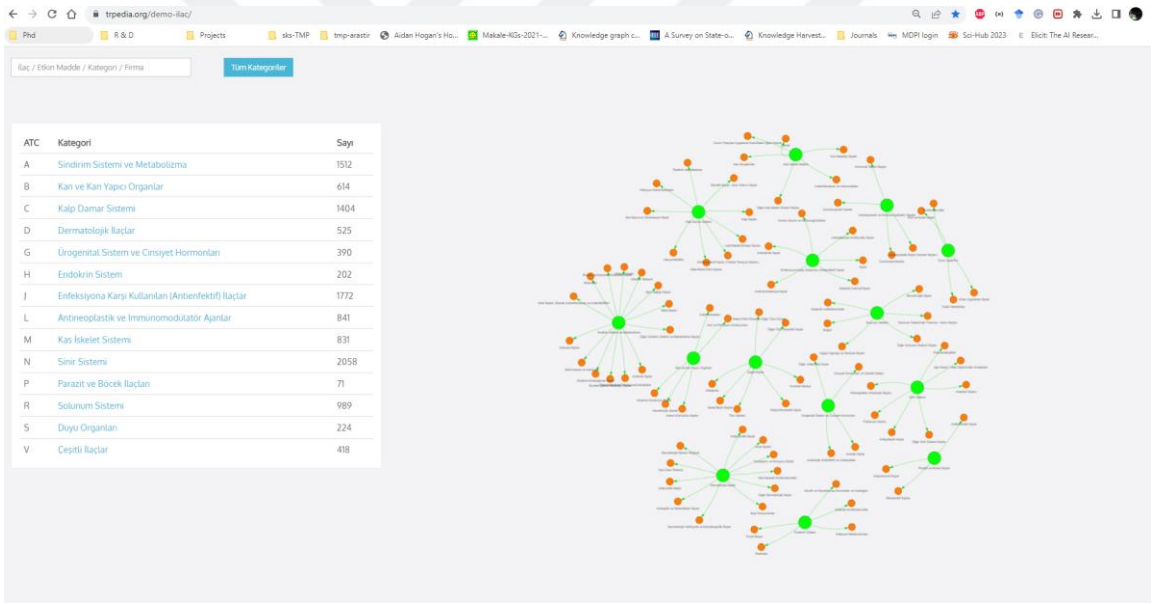
Resim 4.7 Barış Manço Wikipedia sf. – Unutulma hk demo uygulaması- ontolojiler

Resim 4.8'de https://www.ilactr.com adresi üzerinden elde edilen veriler üzerinden oluşturulan graf analitiğinde "Türk ilaç sektöründe ilaç gruplarının" gösterimine yer verilmiştir.

Söz konusu gösterimi sağlamak amacı ile

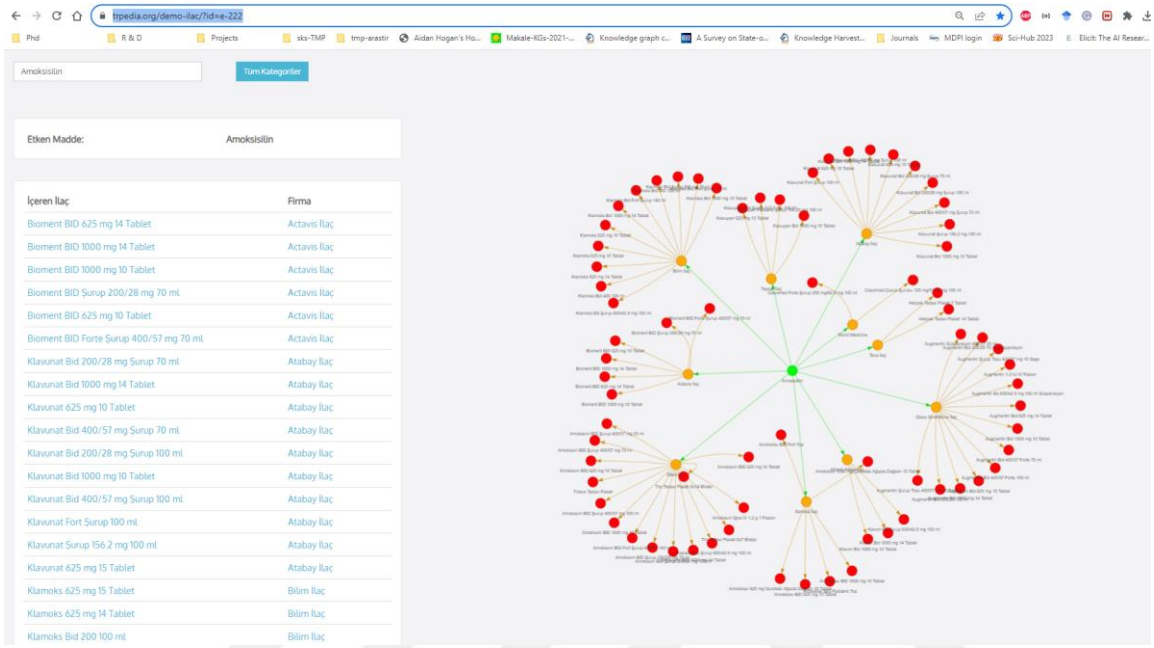
- ATC Kodları
- Genel Hastalık Kategorileri
- İlaç Firmaları
- Etkin Maddeler
- Ürünler

Arasında ontolojiler tanımlanmış graf üzerinde yatay seviyede ve dikey seviyede düğümler arasında erişim ve bilgiye ulaşım mümkün kılınmıştır.



Resim 4.8 Türk ilaç sektöründe ilaç grupları

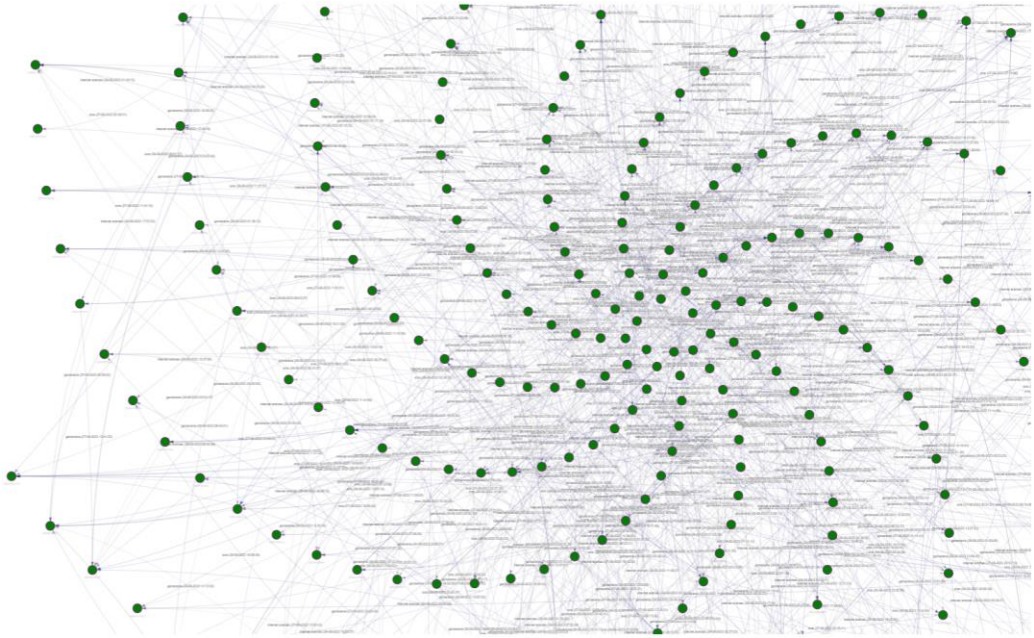
Resim 4.9’da https://www.ilactr.com adresi üzerinden elde edilen veriler üzerinden oluşturulan graf analitiğinde “Türk ilaç sektöründe Amoksisilin etkin maddesini kullanan ilaç ve firmaların” gösterimine yer verilmiştir.



Resim 4.9 Türk ilaç sektöründe amoksisilin etkin maddesini kullanan firmalar

Yine bu gösterimde de varlıklar arasında tanımlanmış ontolojiler üzerinden yatay seviyede ve dikey seviyede düğümler arasında erişim ve bilgiye ulaşım mümkün kılınmıştır.

Resim 4.10’da bilgi grafları ile telekomünikasyon analitiğinin gösterimine, yer verilmiştir.



Resim 4.10 Bilgi grafları ile telekomunikasyon analitiği

Resim 4.10’da telekomünikasyon analitiğinin bilgi graflarının geliştirilen özgün model kapsamında uygulanabilirliğini gösterebilmek amacı ile sentetik veriler üzerinden arama ve mesajlaşma kayıtları ve ontolojiler oluşturulmuş, bilgi grafı üzerinde gösterimine yer verilmiştir.

5. SONUÇ VE ÖNERİLER

Bu tez kapsamında "İnternette Heterojen Veri Kaynaklarından Verinin Toplanması, Doğrulanması ve Sorgulanmasını" Türkçe dilinde gerçekleştirmek amacıyla özgün bir bilgi grafı modeli ve buna yönelik uygulama platformu geliştirilmiştir. Geliştirilen özgün model Türkçe diline yönelik olarak, genel amaçlı, uçtan uca, web ölçekli, doğal dil işleme, bilgi çıkarımı ve güven değeri hesaplamalarını içermektedir.

Bu çalışma, Türkçe için tasarlanmış ilk bilgi grafiği modeli olması, özgün bir model önermesi ve heterojen veri kaynaklarından Türkçe diline yönelik genel amaçlı bir bilgi tabanı ve bilgi grafı sunması nedeniyle literatürdeki bir boşluğu doldurmaktadır.

Doğal dil işleme ve bilgi çıkarımı teknikleri kullanılarak genel amaçlı bir bilgi tabanı oluşturulması, bilginin üçlülere indirgenmesi ve bilgi grafı mimarisinde gösteriminin gerçekleştirilmesi amacıyla Türkçe konu başlıklarına sahip yaklaşık 530.000 Wikipedia sayfasından oluşan veri kümesi kullanılmıştır. Geliştirilen model birçok yenilikçi özelliği bünyesinde barındırmaktadır.

Çalışmamızın literatüre katkıları şu şekilde gerçekleşmiştir:

i. Türkçe dilindeki heterojen veri kaynaklarından bilgi çıkarılması, makine tarafından okunabilir bir formatta saklanması ve grafik gösterimi sağlanmıştır.
ii. Türkçe için genel amaçlı, Web ölçekli, doğal dil işleme ve bilgi çıkarma süreçlerini içeren ilk bilgi grafiğinin literatüre kazandırılmıştır.
iii. Geliştirilen modelin özgün mimarisine makine öğrenmesi algoritmaları kullanılarak topluluk keşfi ve düğüm sınıflandırması gibi yenilikçi özelliklerin eklenmesi ve farklı problemlere uyarlanabilmesi olanaklı kılınmıştır.
iv. TRPedia üçlü çıkarımlarının Wikipedia'daki Türkçe konu başlıklarından yapılan çıkarımlarla karşılaştırılması sonucunda Wikipedia Infobox verilerine göre (bilgi kutularına göre) %6, DBpedia'ya göre 4,28 kat, YAGO bilgi tabanına göre ise 11,01 kat daha iyi performans elde edilmiştir.

    Söz konusu başarım, kıyaslama yapılan çalışmaların sadece Wikipedia Infobox'ları üzerinde yer alan yarı yapılandırılmış veriler ile çıkarımlar yaparken, geliştirilen özgün model kapsamında doğal dil işleme ve bilgi çıkarımı yöntemleri Türkçe diline yönelik olarak etkili biçimde uygulanmış olması neticesinde gerçekleşmiştir.

v. Modelin benzersiz mimarisi, farklı sorunlara ve görevlere uyarlanmasına olanak sağlamaktadır. Bu esneklik, modelin çeşitli alanlarda kullanılabilirliğini ve uygulanabilirliğini artırmaktadır.
vi. Geliştirilen özgün model kapsamında arama motorlarında unutulma hakkının uygulanmasına yönelik graf tabanlı çözüm mimarisi önerisi sunulmuştur. Literatürde unutulma hakkının teknik uygulanabilirliğini sağlamaya yönelik bu alanda önerilmiş ilk çalışmadır. Arama motorlarında unutulma hakkının uygulanmasına yönelik etkili mekanizmaların geliştirilmesine ve nihayetinde bireylerin gizlilik haklarının korunmasına katkıda bulunması beklenmektedir.
vii. Çalışma kapsamında Türkçe dili için literatürdeki en geniş ve kapsamlı veri seti oluşturulmuştur.

Literatürde 2009 yılından itibaren yer alan çalışmalarda çok çeşitli türler ve kategoriler, alanlar arası bilginin graf tabanlı gösterimlerini amaçlayan genel kapsamlı ve Web ölçekli bilgi tabanlarının (DBpedia, Freebase, YAGO vb.) kullanımına yönelik ilgilinin arttığı, istatistiksel ve veri odaklı yöntemlere doğru bir eğilim görülmektedir.

Diğer taraftan geçtiğimiz son birkaç sene içerisinde yapay zekâ tabanlı gelişmiş soru cevaplama ve içerik üretme sistemlerine yoğun ilgi olduğu görülmektedir. Bu sistemler öncelikle metin verilerindeki istatistiksel kalıplar üzerinde eğitilmiş bir konuşma aracısı olsa da karmaşık sistemler için bir ön uç arayüzü olarak hizmet edebileceği değerlendirilmektedir. Teorik olarak, varlıklar hakkındaki ilişkileri ve gerçekleri yapılandırılmış bir formatta yakalayan bilgi grafları gibi yapılandırılmış veri kaynaklarıyla etkileşime girebilir.

Bununla birlikte söz konusu soru cevaplama ve içerik üretme sistemleri bilgi grafları, yapılandırılmış veritabanlarından yapılandırılmamış metin ve multimedyaya kadar internet üzerinde bulunan çeşitli veri havuzları olan heterojen Web kaynaklarından gelen verilerle oluşturulabilir veya geliştirilebilir. Bu çeşitli verilerin birleşik bir bilgi temsiline dahil edilmesi, diyalog sisteminin ilgili, doğru ve bağlama duyarlı yanıtlar veya sonuç üretme becerisini önemli ölçüde destekleyecektir.

Bilgi graflarının bağlantılı veriler ile kullanımı, diğer veri kümeleriyle bağlantı kurulmasını mümkün kılacak ve daha geniş, gezilebilir bir anlamsal alan sunacaktır. Bununla birlikte bağlantılı veri ilkeleri, veri noktalarını Web üzerinden adreslenebilir ve birbirine

bağlanabilir hale getirmek için standart Web protokollerinin ve benzersiz kaynak tanımlayıcılarının (URI'ler) kullanılmasını gerektirmektedir.

Söz konusu bilgi grafı tabanlı veri modeli, World Wide Web'in daha akıllı ve sezgisel bir katmanı olan Semantik Web'in daha geniş vizyonunu aynı zamanda beslemekte, aynı zamanda makineler tarafından da bir dereceye kadar anlaşılır kılınması amaçlamaktadır.

Semantik Web standartları, veriler hakkında otomatik akıl yürütmeyi mümkün kılarak mevcut verilerden yeni gerçekler çıkarma yeteneği sağlamaktadır. Bu nedenle, ChatGPT gibi sohbet robotları bu teknolojilerden yararlanabildikleri takdirde potansiyel olarak çok daha akıllı hale gelebileceği, bağlamsal olarak birbirine bağlı gerçekler ve ilişkiler ağında derinlemesine yanıtlar sunacağı öngörülmektedir.

Dolayısıyla, bu teknolojiler ve metodolojiler bağımsız olarak var olabilirken, entegrasyonları herhangi bir teknolojinin tek başına başarabileceğinden çok daha güçlü, yetenekli ve kullanıcı dostu sistemleri ortaya çıkaracaktır. Bu teknolojiler sadece büyük veri analitiğini değil, aynı zamanda akıllı- bağlam, anlayış ve anlamlı etkileşim yeteneğine sahip olduğu bir geleceği de mümkün kılacaktır.

Diğer taraftan metin benzerliği algoritmaları, birbirleriyle ne kadar yakından ilişkili olduklarını bulmak için farklı belgeleri veya metin parçalarını analiz edebilir ve karşılaştırabilir. Arama motorları, öneri sistemleri ve içerik kategorizasyonu dahil olmak üzere çok sayıda uygulama için metin benzerliği algoritmaları hayati önem taşımaktadır. Belge özetleme ve belge üretmeye yönelik teknolojiler, yalnızca hızlı içgörülere ihtiyaç duyan son kullanıcılar için değil, aynı zamanda büyük miktarda veriyi ayrıştırması gereken sistemler için de faydalı olacaktır. Belge benzerliği ve özetleme teknolojilerinin sohbet robotları, bilgi grafları, heterojen Web veri kaynakları, bağlantılı veri ve Semantik Web ortamına entegre ederek daha akıllı, verimli ve kullanıcı dostu sistemlere yönelik araştırmaların önemli katkılar sağlayacağı öngörülmektedir. Bu teknolojilerin birlikteliği birbirine bağlı, akıllı bir veri ekosisteminin gerçekleştirilmesine katkıda bulunacaktır.

İnternet veri kaynakları üzerinden çalışmak beraberinde zorlukları getirmektedir. Web siteleri arasındaki standartlaştırılmış veri formatlarının eksikliği, ilgili verilerin doğru bir şekilde çıkarılması için özelleşmiş yöntemlerin kullanılmasını gerektirmektedir. Ayrıca, web sayfaları dinamik olarak değişmesi, zaman içinde veri tutarlılığını sürdürmekte zorluklar yaratması nedeni ile güçlü doğrulama yaklaşımlarına ihtiyaç duyulmaktadır. Gerek

arama motoru sorguları gerekse video paylaşım platformları üzerinden yapılan sorgulama işlemlerinde anlamsal benzerliği olmayan hatta olumsuz içerikler ile karşılaşılmaktadır. Söz konusu sorunların aşılabilmesi amacı ile gerek anlamsal eşleştirmeler gerek benzerlik hesaplamaları gerekse kaynakların güvenilirliklerinin hesaplanması gerekliliği görülmektedir.

Bununla birlikte Türkçe diline ait doğal dil işleme, bilgi çıkarımı, anlamsal ağ ve bağlantılı verilere yönelik Web ölçekli çalışma yok denecek kadar azdır.

Çalışmamız ile bu alanda araştırma yapmak isteyen araştırmacılara graf analitiği de dahil olmak üzere genel bir bakış açısı sunulması aynı zamanda amaçlanmıştır.

KAYNAKLAR

1. Wang, K. (2015, May). The knowledge Web meets big scholars. *In Proceedings of the 24th International Conference on World Wide Web,*577-578.

2. Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM Sigkdd Explorations Newsletter*, 14(2), 20-28.

3. Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2017). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1), 17-37.

4. Shi, C., & Philip, S. Y. (2017). Heterogenous Information Network Analysis and Applications. *Cham: Springer International Publishing*. Pp:5-24

5. Weikum, G., Hoffart, J., & Suchanek, F. (2019). Knowledge harvesting: achievements and challenges. *Computing and Software Science: State of the Art and Perspectives*, 217-235.

6. Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2022). Heterogeneous information networks: the past, the present, and the future. *Proceedings of the VLDB Endowment*, *15*(12), 3807-3811.

7. Internet: A. Singhal. "Introducing the Knowledge Graph: Things, not Strings", https://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html, Son Erişim : 04/08/2023

8. Sheth, A., Padhee, S., & Gyrard, A. (2019). Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Computing*, 23(4), 67-75.

9. Hür, A., Janjua, N., & Ahmed, M. (2021, December). A survey on state-of-the-art techniques for knowledge graphs construction and challenges ahead. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 99-103).

10. Weikum, G., & Theobald, M. (2010). From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. In *PODS '10: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (s. 65-76). ACM.

11. Axel Polleres, Aidan Hogan, Andreas Harth, and Stefan Decker. Can we ever catch up with the Web? Semantic Web, 1(1-2):45–52, 2010. DOI:10.3233/SW-2010-0016.

12. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34-43.

13. Tiwari, S., Gaurav, D., Srivastava, A., Rai, C., & Abhishek, K. (2021). A Preliminary Study of Knowledge Graphs and Their Construction. Emerging Technologies in Data Mining and Information Security (s. 11-20). Springer. https://doi.org/10.1007/978-981-15-9774-9_2

14. Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4), 211-226.

15. Speer, R. R. R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4444–4451).

16. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., ... & Yates, A. (2004, May). Web-scale information extraction in KnowItall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web* (pp. 100-110).

17. Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., & Soderland, S. (2007, April). TextRunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (pp. 25-26).

18. Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from Wikipedia and Wordnet. *Journal of Web Semantics*, 6(3), 203-217.

19. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1247-1250).

20. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Journal Of Web Semantics*, 7(3), 154-165.

21. Zimmermann, A., Gravier, C., Subercaze, J., & Cruzille, Q. (2013, May). Nell2RDF: Read the Web, and Turn it into RDF. In *KNOW@ LOD* (pp. 2-8).

22. Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

23. Ferrucci, David A. (2012). Introduction to This Is Watson. IBM Journal of Research and Development, 2012,56.3.4: 1: 1-1: 15.

24. Internet: IBM Watsons Healthcare, https://tw.rpi.edu/project/HEALS, Son Erişim: 10/08/2023

25. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).

26. Internet: N. Torzec, "Yahoo's Knowledge Graph". https://research.yahoo.com/publications/9194/yahoo-knowledge-graph, Son Erişim: 10/08/2023

27. Weikum, G., Hoffart, J., & Suchanek, F. (2016). Ten Years of Knowledge Harvesting: Lessons and Challenges. IEEE Data Eng. Bull., 39(3), 41-50.

28. Tiwari, S., Al-Aswadi, F. N., & Gaurav, D. (2021). Recent Trends in Knowledge Graphs: Theory and Practice. *Soft Computing*, 25, 8337-8355.

29. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, Acquisition, and Applications. *IEEE Transactions on neural networks and learning systems,* 33(2), 494-514.

30. Färber, M., & Rettinger, A. (2018). Which Knowledge Graph is Best For Me? arXiv preprint arXiv:1809.11099.

31. Berti-Equille, L., & Ba, M. L. (2016). Veracity of Big Data: Challenges of Cross-Modal Truth Discovery. *Journal of Data and Information Quality (JDIQ)*, *7*(3), 1-3.

32. Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information Extraction Meets The Semantic Web: a survey. *Semantic Web*, *11*(2), 255-335.

33. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., & Decker, S. (2011). Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. Journal of Web Semantics, 9(4), 365-401.

34. Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). Foundations of semantic web technologies. CRC press.

35. T. Heath, C. Bizer. (2011) Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool.

36. Janev, V., Graux, D., Jabeen, H., & Sallinger, E. (2020). Knowledge Graphs and Big Data Processing (p. 5-30). Springer Nature.

37. Balog, K. (2018). Entity-oriented search. Springer Nature. Pages 25-52

38. Suchanek, F., & Weikum, G. (2013, April). Knowledge Harvesting from Text and Web sources. In 2013 IEEE 29th International Conference on Data Engineering (ICDE) (pp. 1250-1253). IEEE.

39. Mahdisoltani, F., Biega, J., Suchanek, F.M. (2015). YAGO3: A Knowledge Base From Multilingual Wikipedias. In: Seventh Biennial Conference on Innovative Data Systems Research, CIDR ’15.

40. Suchanek, F. M., & Weikum, G. (2014). Knowledge Bases in the Age of Big Data Analytics. PVLDB, 7(13), 1713-1714.

41. Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and Its Technical Challenges. *Communications of the ACM*, *57*(7), 86-94.

42. Rajaraman, V. (2016). Big data analytics. *Resonance*, *21*, 695-716.

43. Zakir, J., Seymour, T., & Berg, K. (2015). Big Data Analytics. *Issues in Information Systems*, *16*(2).

44. Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big Data Analytics and Big Data Science: A Survey. *Journal of Management Analytics*, *3*(1), 1-42.

45. Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big Data Analytics: A Survey. *Journal of Big Data*, *2*(1), 1-32.

46. Kaur, A. (2016). Big data: A review of challenges, tools and techniques. *International journal of scientific research in science, engineering and technology*, 2(2), 1090-1093.

47. Babu, S. A., Vijay, N., & Gadde, R. (2017). An Overview of Big Data Challenges, Tools and Techniques. International Journal of Research and Applications, 596-601.

48. Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. Web Scraping Techniques and Applications: A Literature.

49. Ren, X., Wang, H., & Dai, D. (2020, December). A Summary of Research on Web Data Acquisition Methods Based on Distributed Crawler. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)* (pp. 1682-1688). IEEE.

50. Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web Data Extraction, Applications and Techniques: A survey. *Knowledge-based systems*, 70, 301-323.

51. Li, H. (2021, May). Research on Big Data Analysis Data Acquisition and Data Analysis. In 2021 International Conference on Artificial Intelligence, *Big Data and Algorithms (CAIBDA)* (pp. 162-165). IEEE.

52. Hitzler, P. (2021). A review of the semantic web field. *Communications of the ACM*, 64(2), 76-83.

53. Kejriwal, M., Sequeda, J. F., & Lopez, V. (2019). Knowledge graphs: Construction, Management and Querying. *Semantic Web*, *10*(6), 961-962.

54. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., & Decker, S. (2012). Scalable and Distributed Methods for Entity Matching, Consolidation and Disambiguation over Linked Data Corpora. Journal of Web Semantics, 10, 76-110.

55. Hogan, A. (2016). Linked Data & the Semantic Web Standards. In *Linked Data Management* (pp. 31-76). Chapman and Hall/CRC.

56. Polleres, A., Hogan, A., Delbru, R., & Umbrich, J. (2013). RDFS and OWL reasoning for linked data. In Reasoning Web International Summer School (pp. 91-149). Berlin, Heidelberg: Springer Berlin Heidelberg.

57. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An Empirical Survey of Linked Data Conformance. *Journal of Web Semantics*, *14*, 14-44.

58. Ali, W., Saleem, M., Yao, B., Hogan, A., & Ngomo, A. C. N. (2022). A survey of RDF stores & SPARQL engines for querying knowledge graphs. *The VLDB Journal*, 1-26.

59. Hogan, A. (2020). The Semantic Web: Two decades on. *Semantic Web*, *11*(1), 169-185.

60. Hogan, A., & Hogan, A. (2020). *Web of data*. Springer International Publishing.

61. Salas, J., & Hogan, A. (2021). Semantics and Canonicalisation of SPARQL 1.1. *Semantic Web,* (Preprint), 1-65.

62. Umbrich, J., Hogan, A., Polleres, A., & Decker, S. (2015). Link traversal querying for a diverse web of data. Semantic Web, 6(6), 585-624.

63. Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoč, D. (2017). Foundations of modern query languages for graph databases. *ACM Computing Surveys (CSUR)*, *50*(5), 1-40.

64. Berti-Équille, L. (2007). Quality awareness for managing and mining data (Doctoral dissertation, Universté de Rennes).

65. Batini C., Scannapıeco M. (2018). *Data and Information Quality: Dimensions, Principles and Techniques*. Sprınger.

66. Internet: ISO, Data Quality Model, https://www.iso.org/standard/35736.html, Son Erişim: 11/08/2023

67. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93.

68. Färber, M., Ell, B., Menne, C., & Rettinger, A. (2015). A comparative survey of Dbpedia, Freebase, Opencyc, Wikidata, and Yago. *Semantic Web Journal*, 1(1), 1-5.

69. Pillai, S. G., Soon, L. K., & Haw, S. C. (2019). Comparing DBpedia, Wikidata, and YAGO for web information retrieval. In *Intelligent and Interactive Computing: Proceedings of IIC 2018* (pp. 525-535). Springer Singapore.

70. Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3), 489-508.

71. Pan, J. Z., Vetere, G., Gomez-Perez, J. M., & Wu, H. (Eds.). (2017). *Exploiting linked data and knowledge graphs in large organizations* (p. 281). Heidelberg: Springer.

72. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., & Gómez-Pérez, A. (2018). A comprehensive quality model for Linked Data. *Semantic Web,* 9(1), 3-24.

73. Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked Data Quality Of Dbpedia, Freebase, Opencyc, Wikidata, and Yago. *Semantic Web,* 9(1), 77-129.

74. Huaman, E., Kärle, E., & Fensel, D. (2020). Knowledge graph validation. *arXiv preprint arXiv:2005.01389*.

75. Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., ... & Chen, H. (2021). Knowledge graph quality control: A survey. Fundamental Research, 1(5), 607-626.

76. Huaman, E. (2022). Steps to Knowledge Graphs Quality Assessment. arXiv preprint arXiv:2208.07779.

77. Ringler, D., & Paulheim, H. (2017). One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & Co. *In KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI*, Dortmund, Germany, September 25–29, 2017, Proceedings 40 (pp. 366-372). Springer International Publishing.

78. P.N. Mendes, H. Mühleisen, C. Bizer, Sieve: Linked Data Quality Assessment and Fusion, in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 2012, pp. 116–123

79. Berners-Lee, T. (2006). Linked Data (Personal notes). World Wide Web Consortium (W3C). Retrieved from https://www. w3. org/DesignIssues/LinkedData.html

80. Davenport, T.H.: Analytics 3.0 (2013). https://hbr.org/2013/12/analytics-30

81. Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. Semantics (Posters, Demos, SuCCESS), 48, s. 1-4.

82. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724-2743.

83. Hogan, A. (2020). Knowledge graphs: Research directions. Reasoning Web. Declarative Artificial Intelligence: 16th International Summer School 2020, Oslo, Norway, June 24–26, 2020, Tutorial Lectures 16, 223-253.

84. Wang, C., Song, Y., El-Kishky, A., Roth, D., Zhang, M., & Han, J. (2015, August). Incorporating world knowledge to document clustering via heterogeneous information networks. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1215-1224).

85. Suchanek, F., & Weikum, G. (2013, June). Knowledge harvesting in the big-data era. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 933-938).

86. Nakashole, N., Theobald, M., & Weikum, G. (2011). Scalable knowledge harvesting with high precision and high recall. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 227-236).

87. Galárraga, L., Heitz, G., Murphy, K., & Suchanek, F. M. (2014, November). Canonicalizing open knowledge bases. In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (pp. 1679-1688).

88. Ryen, V., Soylu, A., & Roman, D. (2022). Building Semantic Knowledge Graphs From (Semi-) Structured Data: A Review. *Future Internet*, 14(5), 129.

89. Kejriwal, M. (2022). Knowledge graphs: A Practical Review of The Research Landscape. *Information*, 13(4), 161

90. Weikum, G., Dong, X. L., Razniewski, S., & Suchanek, F. (2021). Machine knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Foundations and Trends® in Databases,* 10(2-4), 108-490.

91. Hogan A, Blomqvist E, Cochez M, et al. (2021) Knowledge Graphs. *ACM Computer Surveys (CSUR)* 54(4):1–37

92. Fensel, D., Simsek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., ... & Wahler, A. (2020). Knowledge Graphs. Cham: Springer International Publishing.

93. Fan, Y., Wang, C., Zhou, G., & He, X. (2017). Dkgbuilder: An Architecture for Building a Domain Knowledge Graph From Scratch. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017*, Suzhou, China, March 27-30, 2017, Proceedings, Part II 22 (pp. 663-667). Springer International Publishing.

94. Tamašauskaitė, G., & Groth, P. (2023). Defining a knowledge graph development process through a systematic review. *ACM Transactions on Software Engineering and Methodology*, 32(1), 1-40.

95. Simsek, U., Umbrich, J., & Fensel, D. (2020, January). Towards a Knowledge Graph Lifecycle: A pipeline for the population of a commercial Knowledge Graph. In *Qurator*.

96. Chaudhri, V. K., Baru, C., Chittar, N., Dong, X. L., Genesereth, M., Hendler, J., . . . Wang, K. (2022). Knowledge graphs: Introduction, history, and perspectives. AI Magazine, 43(1), 17-29.

97. Weikum, G. (2021). Knowledge Graphs 2021: A Data Odyssey. Proceedings of the VLDB Endowment, 14(12), 3233-3238. doi:10.14778/3476311.347639

98. Zhao, Z., Han, S. K., & So, I. M. (2018). Architecture of Knowledge Graph Construction Techniques. *International Journal of Pure and Applied Mathematics, 118*(19).

99. Grishman, R. (2015). Information Extraction. *IEEE Intelligent Systems*, 30(5), 8-15.

100. Singh, S. (2018). Natural Language Processing for Information Extraction., (s. 1-24). doi:https://doi.org/10.48550/arXiv.1807.02383

101. Helali, M., Vashisth, S., Carrier, P., Hose, K., & Mansour, E. (2023). Linked Data Science Powered by Knowledge Graphs. Cornell University.

102. Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A Survey on Open Information Extraction. Computation and Language, 1-13. doi:https://doi.org/10.48550/arXiv.1806.05599

103. Muhammad, I., Kearney, A., Gamble, C., Coenen, F., & Williamson, P. (2020). Open Information Extraction for Knowledge Graph Construction. In *Database and Expert Systems Applications: DEXA 2020 International Workshops BIOKDD, IWCFS and MLKgraphs* (s. 103-113). Bratislava, Slovakia: Springer.

104. Dong, X. L., & Srivastava, D. (2015, May). Knowledge curation and knowledge fusion: challenges, models and applications. In *Proceedings of the 2015 acm sigmod international conference on management of data* (pp. 2063-2066).

105. X. L. Dong et al., "From data fusion to knowledge fusion," In *Proc. VLDB Endow.*, vol. 7, no. 10, pp. 881–892, 2014, doi:10.14778/2732951.273

106. Equille L. B. Truth Discovery. (2018), Springer; Springer International Publishing, pp.1-8, 2018

107. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., ... & Han, J. (2016). A survey on truth discovery. ACM Sigkdd Explorations Newsletter, 17(2), 1-16.

108. Pasternack, J., & Roth, D. (2010, August). Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 877-885).

109. Galland, A., Abiteboul, S., Marian, A., & Senellart, P. (2010, February). Corroborating information from disagreeing views. In Proceedings of the third ACM international conference on Web search and data mining (pp. 131-140).

110. Yin, X., Han, J., & Yu, P. S. (2007, August). Truth Discovery With Multiple Conflicting Information Providers On The Web. *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1048-1052).

111. Waguih, D. A., & Berti-Equille, L. (2014). Truth discovery algorithms: An experimental evaluation. arXiv preprint arXiv:1409.6428.

112. Berti-Equille, L., & Borge-Holthoefer, J. (2022). Veracity of Data. Springer Nature.

113. Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment,* 2(1), 550-561.

114. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014, June). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 1187-1198).

115. Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. (2015). On the discovery of evolving truth. In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), pages 675-684.

116. X. Yin and W. Tan. Semi-supervised truth discovery. In Proc. of the International Conference on World Wide Web (WWW'11), pages 217{226, 2011

117. X. Yin and W. Tan. Semi-supervised truth discovery. In Proc. of the International Conference on World Wide Web (WWW'11), pages 217{226, 2011

118. Ba, M. L., Berti-Equille, L., Shah, K., & Hammady, H. M. (2016, April). VERA: A platform for veracity estimation over web data. In Proceedings of the 25th international conference companion on world wide web (pp. 159-162).

119. Esteves, D., Rula, A., Reddy, A. J., & Lehmann, J. (2018). Toward veracity assessment in RDF knowledge bases: an exploratory analysis. *Journal of Data and Information Quality (JDIQ)*, 9(3), 1-26.

120. Lamine Ba, M., Berti-Equille, L., & Hammady, H. M. (2016). Discovering the Truth on the Web Data: One Facet of Data Forensics. In *Qatar Foundation Annual Research Conference Proceedings* (Vol. 2016, No. 1, p. ICTPP3179). Qatar: HBKU Press.

121. Kejriwal, M. (2019). Domain-specific knowledge graph construction. Cham: Springer International Publishing.

122. Goyal, A., Gupta, V., & Kumar, M. (2018). Recent Named Entity Recognition and Classification Techniques: A Systematic Review. Computer Science Review, 29, 21-43. doi:https://doi.org/10.1016/j.cosrev.2018.06.001

123. Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., & Stefanidis, K. (2020). An Overview of End-to-End Entity Resolution for Big Data. ACM Computing Surveys, 53(6), 1-42.

124. Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. ACM Computing Surveys, 1-43. doi:https://doi.org/10.1145/3604931

125. Augenstein, I., Maynard, D., & Ciravegna, F. (2016). Distantly supervised web relation extraction for knowledge base population. *Semantic Web*, 7(4), 335-349.

126. Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. arXiv preprint arXiv:1712.05191.

127. Cerezo-Costas, H., & Martín-Vicente, M. (2018). Relation Extraction for Knowledge Base Completion: A Supervised Approach. In Semantic Web Challenges: 5th SemWebEval Challenge at ESWC 2018, Heraklion, Greece, June 3–7, 2018, Revised Selected Papers 5 (pp. 52-66). Springer International Publishing.

128. Jiang, H., Bao, Q., Cheng, Q., Yang, D., Wang, L., & Xiao, Y. (2020). Complex relation extraction: Challenges and opportunities. arXiv preprint arXiv:2012.04821.

129. Ma, J., Zhou, C., Wang, Y., Guo, Y., Hu, G., Qiao, Y., & Wang, Y. (2022). PTrustE: A high-accuracy knowledge graph noise detection method based on path trustworthiness and triple embedding. Knowledge-Based Systems.

130. Shao, T., Li, X., Zhao, X., Xu, H., & Xiao, W. (2021). DSKRL: A dissimilarity-support-aware knowledge representation learning framework on noisy knowledge graph. Neurocomputing.

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı	: Serdar Kürşat SARIKOZ
Uyruğu	:
Doğum yeri	:
Medeni hali	:
E-mail	:

Eğitim

Derece	**Eğitim Birimi**	**Mezuniyet Tarihi**
Doktora	Gazi Üniversitesi/Bilişim Enstitüsü/Bilgisayar Bilimleri	: Devam ediyor
Yüksek Lisans	Gazi Üniversitesi/ Bilişim Enstitüsü/Bilgisayar Bilimleri	: 2011
Lisans	Sakarya Üniversitesi/ Bilgisayar Mühendisliği	: 2002
Lise	Ankara Yenimahalle Teknik Lisesi	: 1996

İş Deneyimi

2006 – Halen	Bilgi Teknolojileri ve İletişim Kurumu	: Bilişim Uzmanı
2003 – 2006	Türk Telekomünikasyon A.Ş.	: Uzman Yardımcısı
2003 – 2003	Adalet Bakanlığı Bilgi İşlem Daire Başkanlığı	: Çözümleyici
2001	Türkiye İş Bankası B.İ.M.	: Stajyer
2000	Türkiye İş Bankası B.İ.M.	: Stajyer
1995	TRT Genel Müdürlüğü	: Stajyer
1994	T.C.D.D. Genel Müdürlüğü	: Stajyer

Yabancı Dil

İngilizce

Yayınlar

Sarıkoz S.K., Akcayol M.A., Examining Knowledge Extraction Processes from Heterogene ous Data Sources, Brilliant Engineering, 1(2023), 4798. https://doi.org/10.36937/ben.2023. 4798

Sarıkoz, S. & Küçükali, M. (2020). Hücresel Şebekelerde Makineler Arası İletişim ve Güv enlik Önerileri, *TÜBAV Bilim Dergisi*, 13 (4) , 11-26 .

Cubukcu A., Sarıkoz S.K, Kişisel Veriler ve Mahremiyetin Korunmasının Elektronik Habe rleşme Perspektifinden İncelenmesi: Türkiye İçin Öneriler, 7. International Management In formation Systems Conference "*Health Informatics and Analytics", 2020,*

Taskin, C., & Sarikoz, S. K. (2008, June). An overview of image compression approaches. In *2008 The Third International Conference on Digital Telecommunications (icdt 2008)* (p p. 174-179). IEEE.

GAZİLİ OLMAK AYRICALIKTIR