Linked Data Üzerinden Doğal Dil Sorgularını Cevaplayan Sistem

A. Talha Kabakuş [1], Erdoğan Doğdu [2]

[1] Abant İzzet Baysal Üniversitesi

[2] TOBB Ekonomi ve Teknoloji Üniversitesi

talha.kabakus@ibu.edu.tr
edogdu@etu.edu.tr

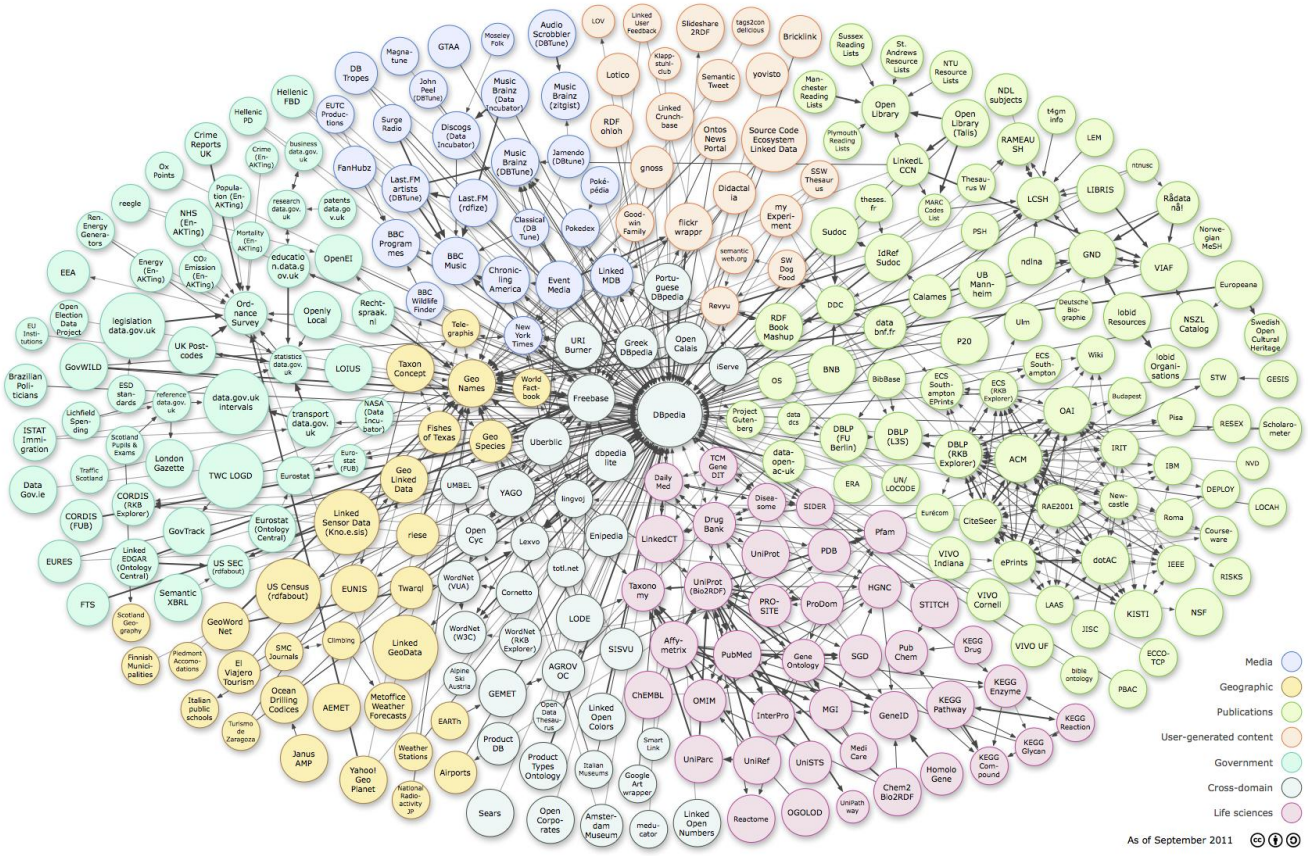
Özet. Günden güne kullanıcısı sayısı öngörülemez bir hızla artan web, özellikle Web 2.0 ile birlikte kullanıcılarıyla olan etkileşimini artırmıştır. Ondan öncesinde web *(Web 1.0)* kullanıcılarına bilgi veren, tek yönlü bir iletişim içinde bulunan bir araçken, günümüzde kullanıcı etkileşimi üst düzeyde olan günlük hayatın vazgeçilmez bir parçası haline gelmiştir. Günümüzde kullanıcılar etkileşimli web sayfaları, arama motorları sayesinde isteklerine en kısa zamanda ulaşmayı amaçlamaktadırlar. Semantik Web diğer adıyla Web 3.0 ise bu adımı da bir adım öteye götürerek web'in sadece insanlar tarafından değil, makinalar tarafından da anlaşılabilir hale gelmesini amaçlamaktadır. Ortak bir ontoloji sayesinde web'deki bilgilerin tüm dünyayı açılması, erişilebilir olması amaçlanmaktadır. DBpedia projesi, 2007 yılında Wikipedia üzerindeki yapısal verilerin semantik veritabanlarında depolanarak web üzerinden tüm dünyaya sunulabilmesi için başlatılmıştır. Linked Data ise başta DBpedia, FreeBase ve YAGO olmak üzere dünyadaki bütün bu bilgi tabanlarının *(knowledge base)* birleştirilmesini sağlamak üzere kurulmuştur. DBpedia sunduğu bir servis sayesinde kendisinde tanımlı olan verilerin sorgulanmasını sağlamaktadır. Bu sorgulama için kullanılan dil SPARQL (SPARQL Protocol and RDF Query Language) olduğundan konunun uzmanları haricindeki kişiler için bir anlam ifade etmemektedir. Bu çalışmada İngilizce doğal dil sorgularını algılayıp, SPARQL sorgularına çevirerek bu servis üzerinden sorgulamayı sağlayan bir semantik web projesinin detayları anlatılacaktır.

Anahtar Kelimeler. Soru Cevaplayan Sistem, Doğal Dil İşleme, Semantik Web, Linked Data

1 Giriş

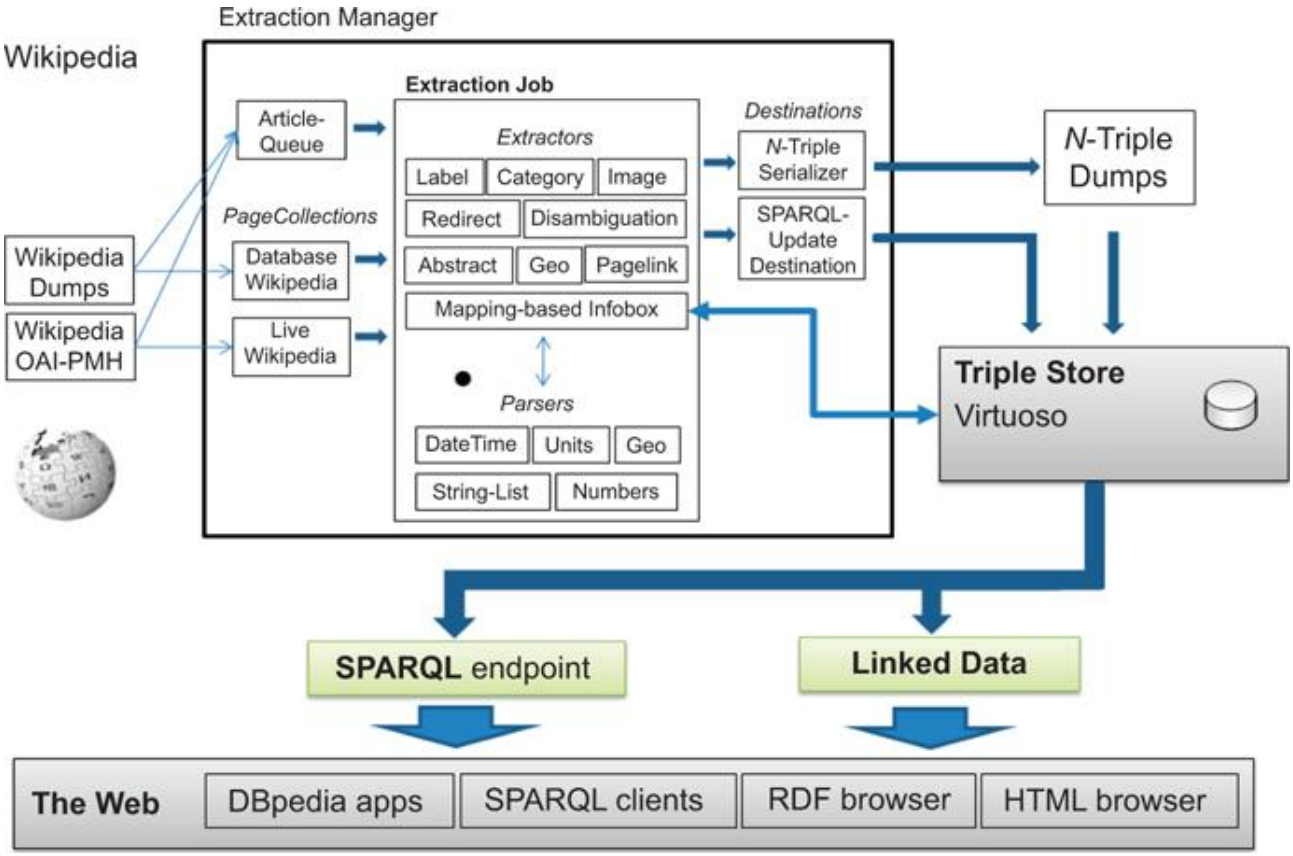
Semantik Web veya diğer adıyla Web 3.0 World Wide Web (WWW)'in kurucusu Tim Berners-Lee tarafından şu şekilde tanımlanmıştır: “Semantik Web, bağımsız bir web değil, şuan kullanılan web'in doğru tanımlanmış verilerin insanlar ve bilgisayarlar tarafından anlamlandırılmasına olanak sağlayan geliştirilmiş halidir.” [1] PCMag tarafından yapılan başka bir semantik web tanımı ise şöyledir: “Semantik Web, makinaların da insanlar gibi web sayfalarını okuyabildiği, arama motorlarının ve yazılım ajanlarının NET'i daha iyi irdeleyebildiği ve aradıklarımızı daha iyi bulabildiği bir ortamdır.” [2] Tanımlamaların da ifade ettiği üzere Semantik Web, Web 2.0'ın barındırdığı tüm özelliklerin üstüne web üzerinde anlamlı ve ilişkilendirilmiş veriler işlenmesi üzerine yoğunlaşmaktadır. Semantik Web, RDF (Resource Description Framework) üzerine inşa edilmiştir ve üzerinde taşınmak istenen veriler üçlüler (triple) halinde tanımlanmaktadır. Tanımlanan verileri sorgulamak için ise standart olarak W3C (World Wide Web Consortium) tarafından SPARQL adı verilen SQL benzeri bir dil kabul edilmiştir.

DBpedia, Wikipedia üzerinde sunulan yapısal verilerin web üzerinden sorgulanabilmesi için semantik veritabanlarında tutulması amacıyla Free University of Berlin, University of Leipzig ve OpenLink Software işbirliği ile 2007 yılında başlatılan bir projedir. DBpedia, Wikipedia üzerindeki bilgi kutularında bulunan yapısal verilerin semantik veritabanlarında depolanmasıyla oluşturulmuş, Linked Open Data bünyesinde bulunan en önemli merkezi bilgi tabanlarındandır. [3] DBpedia üzerinde 2.35 milyonu tanımlı ontoloji içerisinde sınıflandırılmış olmak üzere toplamda 3.77 milyon üçlü tanımlanmıştır. Bu üçlüler içerisinde 764.000 insan, 573.000 yerleşim yeri, 333.000 sanat ürünü ve 192.000 organizasyon tanımlıdır. [4] DBpedia sunduğu bir servis ile tüm bu verilerin sorgulanabilmesini sağlamaktadır. Bu veriler SPARQL adı verilen semantik web sorgu dili tarafından sorgulanmaktadır. Dolayısıyla alanın uzmanları dışındaki kişiler için bu servis bir anlam ifade etmemektedir. Bu çalışma bu servisin genel kullanıma açılması hedeflenerek yapılmıştır. Günümüzde var olan doğal dil işleme kütüphaneleri öncelikli olarak İngilizce'yi hedef almaktadır. Türkçe doğal dil işleme ile henüz stabil bir kütüphane bulunmadığından çalışmamızda başlangıç için İngilizce doğal dil sorgularını işlemek hedef alınmıştır. İlerleyen bölümlerde doğal dil sorgularının algılanıp, SPARQL sorgularına çevrilerek DBpedia üzerinde sorgulanabilecek hale getirilmesi için kullanılan metodoloji anlatılacaktır.



Şekil 1. DBpedia, Linked Open Data bulut diyagramının en merkezi bilgi tabanlarından birisidir, Eylül 2011. [5]

DBpedia Ayrıştırma Kütüphanesinin temel bileşenleri şu şekilde sıralanabilir: Wikipedia makalelerinin soyutlaştırılmış halini temsil eden PageCollection, çıkartılmış RDF üçlülerini depolayacak olan Destination, wiki biçimlerini üçlülere çeviren Extractor, çıkartılan verilere tip tanımlamasının yapılmasını sağlayan ve değişik birimler arasında dönüşüm sağlayan Parser. Kütüphanenin çekirdeğini ise Extraction Manager oluşturur ve Wikipedia makalelerinin Extractor tarafından üçlülere dönüşümünü ve sonrasında hedef birimlerde depolanmasını sağlar. [8]



Şekil 2. DBpedia bileşenlerinin önizlemesi [8]

2 Metodoloji

Geliştirilen sistem temel olarak 3 aşamadan oluşmaktadır:

1. Doğal dil sorularının doğal dil işleme yöntemleriyle anlamlandırılması
2. Anlamlandırılan doğal dil sorularına uygun olan SPARQL sorgularının üretilmesi
3. DBpedia servisi üzerinden SPARQL sorgularının işletilmesi ve sonuçların web arayüzünde gösterilmesi

İlerleyen kısımlarda herbir aşama detaylı olarak açıklanacaktır.

2.1 Doğal Dil Sorularının Doğal Dil İşleme Yöntemleriyle Anlamlandırılması

DBpedia servisinin web'in tüm kullanıcılarına hitap edebilmesi için kullanıcıların ortak dili olan doğal dil sorgularının SPARQL sorgu diline çevrilmesi gerekmektedir. Son kullanıcıların doğal dil ile sordukları soruların algılanabilmesi için geliştirilen sistemde Apache OpenNLP kütüphanesinden faydalanılmıştır. Kütüphane verilen cümlenin öğelerine ayırmakla beraber sunduğu eğitim verileriyle insan, zaman, para, lokasyon, yüzde ve organizasyon tespitinde de bulunabilmektedir. Geliştirdiğimiz bir Java sınıfı sayesinde muhtemel özneler birleştirilmekte ve mümkün olan en büyük nesne adayı bulunmaktadır. Örnek olarak verilecek olursa "Who is Michael Jordan?" gibi bir sorguya kütüphanenin "insan" eğitim verisiyle bulabileceği kelimeler ayrı ayrı "Michael" ve "Jordan" olarak bulunmaktadır. Bu aday kelimeler birleştirilerek en muhtemel arama kriteri bulunmaktadır. Benzer şekilde, "who" kelimesinin soru bildiren WH sorularına ait olduğu ve "is" kelimesinin ise yardımcı fiil olduğu belirlenmektedir. Ön tanımlı desenler elde edilen WH soru kelimelerinden yoğun bir şekilde faydalanmaktadır. Bu şekilde aranan öznenin türü belirlenmiş olmaktadır. Şekil 3.'de "who is Michael Jordan?" doğal dil sorusunun ögeleri etiketleriyle beraber ağaç yapısında görüntülenmiştir.

```
(S who
  (S (VP is
       (NP Michael Jordan))))
```

Şekil 3. "Who is Michael Jordan?" sorusunun ögelerine ayrılmış hali

2.2 Anlamlandırılan Doğal Dil Sorularına Uygun Olan SPARQL Sorgularının Üretilmesi

Doğal dil sorguları öğelerine ayrıldıktan sonra sistem tarafından belirlenmiş desenlerle (pattern) eşleştirmesi yapılmaktadır. Tablo 1.'de geliştirilen sistem üzerinde tanımlanan desenler ve beklenen cevap türleri listelenmiştir.

Tablo 1. Geliştirilen sistem tarafından öntanımlı soru desenleri ve aranan cevap türleri

Soru Deseni	**Aranan Cevap Türü**
Who / whom	İnsan/Kişi
When	Tarih
Where	Yer, lokasyon
What	Herşey
Find	Yer, lokasyon
Locate	İçinde latitude & longitude bilgisi içeren lokasyon

Cümlenin yüklemi düzenli (regular) veya düzensiz (irregular) fiil olması ihtimaline yönelik olarak açık kaynak kodlu NoSQL veritabanı olan MongoDB üzerinde düzensiz fiiller ve halleri tutulmuştur. MongoDB, esnek, ölçeklenebilir ve oldukça hızlı çalışan döküman tabanlı olarak bir veritabanıdır. [6] Bulunan yüklemin veritabanında karşılığı bulunamaması durumunda düzenli olarak kabul edilmiştir. SPARQL sorguları üçlüler üzerinde arama yaptığından sorgunun nesnesi soru desenlerinin belirlendikten sonra ilgili eğitim verisi üzerinden aranmıştır. Özellikle kompleks sorgularda yanlış nesne eşleştirmesini önlemede bu süreç büyük önem arz etmektedir.

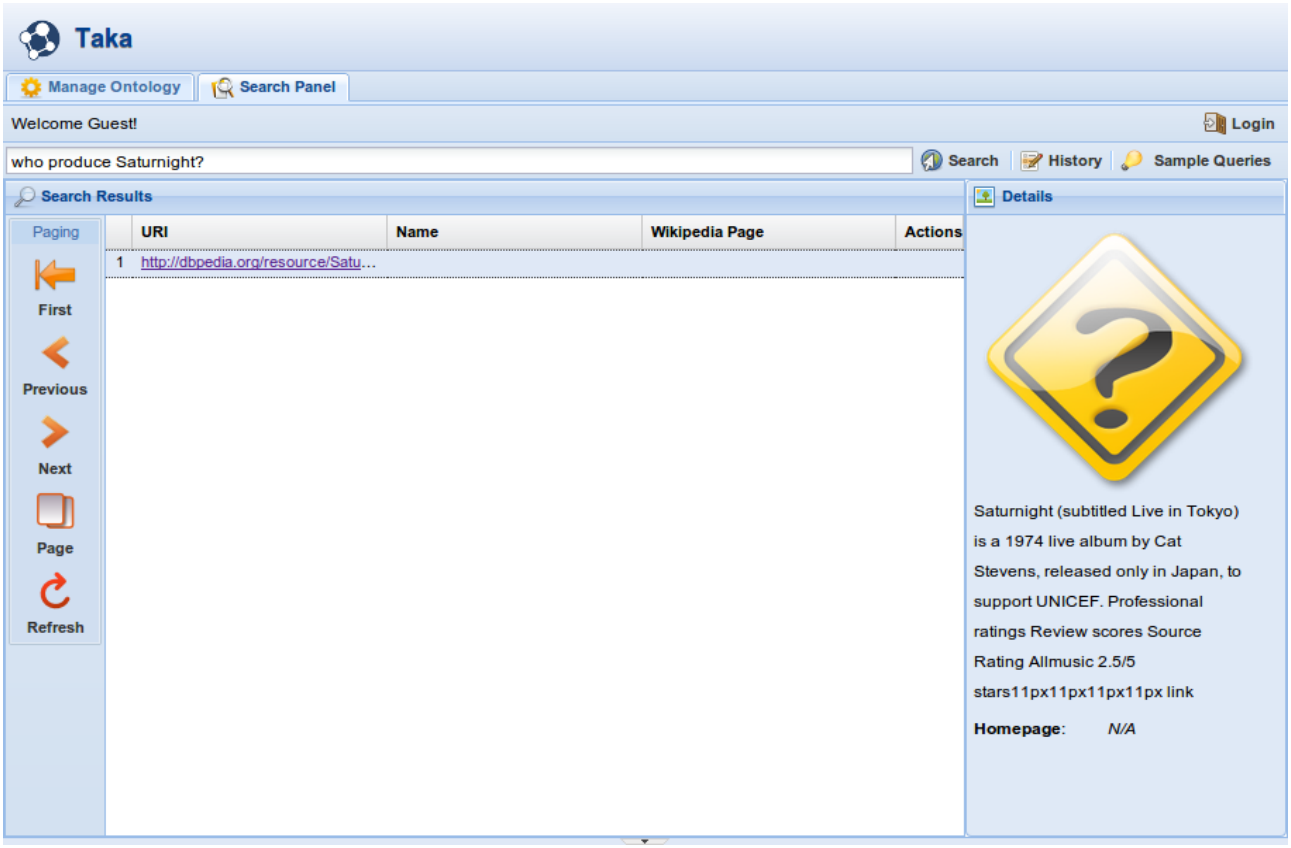
Sistem üzerinde test edilmiş olan “who produce Saturnight?” doğal dil sorusunu ele alarak sistemin çalışma şekli şöyle detaylandırılabilir: Sorgu öncelikle doğal dil işleme yöntemleriyle öğelerine ayrılmakta ve herbir kelimenin türü elde edilmektedir. Bu örnek için bu aşamada şu sonuçlar elde edilecektir: “Who : Soru eki” | “produce : yüklem” | “Saturnight : nesne”. Öğelerine ayrılan soru cümlesi ilk olarak ön tanımlı desenlerle eşleştirilmeye çalışılmaktadır. “Who” soru ekine yönelik bir ön tanımlı desen olduğundan uygun bir SPARQL sorgusu ön tanımlı desenler aracılığıyla elde edilecektir. Sonraki aşamada ise sadece doğal dil işleme yöntemleri kullanılarak bir sorgu üretilecektir. Bunun için üçlüyü oluşturacak olan özellik ve nesnenin belirlenmesi gerekmektedir. Doğal dil işleme yöntemleriyle yukarıdaki örnek için nesne olarak “Saturnight”, yüklem olarak ise “produce” kelimesi bulunmuştu. Bu yükleme karşılık gelecek olan muhtemel DBpedia özelliği DBpedia ontoloji modeli üzerinden aranmış ve Tablo 2.'deki sonuçları listelenen özellikler arasından en benzer olan “producer” özelliği seçilmiştir. Bu yaklaşımla |Özne: ?x| |Özellik: producer| |Nesne: Saturnight| olan bir üçlü üretilmiştir. Elde edilen bu iki SPARQL sorgusu birleştirilerek Şekil 4.'deki nihai sorgu elde edilmiştir. Şekil 5.'de ise sorgu sonucunu gösteren ekran görüntüsü sunulmuştur.

Tablo 2. Produce yüklemiyle benzerlik gösteren DBpedia özellikleri ve Levenshtein benzerlik değerleri

DBpedia Özellik Adı	Levenshtein Benzerlik Değeri
producer	1
coProducer	4
coExecutiveProducer	13
wineProduced	6
executiveProducer	11

```
PREFIX owl: <http://www.w3.org/2002/07/owl#> PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/> PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/> PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/> PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbont: <http://dbpedia.org/ontology/> PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT * WHERE
{?x rdfs:label "Saturnight"@en .
  ?x dbont:abstract ?abstract .
     OPTIONAL
       {?x foaf:name ?name .
          ?x dbont:thumbnail ?thumbnail .
          ?x foaf:isPrimaryTopicOf ?wiki .
          ?x foaf:homepage ?homepage .
          ?x owl:sameAs ?sameAs .
           FILTER regex(str(?sameAs), '^http://rdf.freebase.com', 'i') }
FILTER (lang(?abstract) = 'en')  }
UNION
{?y dbont:producer ?x .
 ?y rdfs:label ?target .
 ?x rdfs:label ?name .
 ?x dbont:abstract ?abstract .
    OPTIONAL
    {?x foaf:isPrimaryTopicOf ?wiki .
     ?x dbont:thumbnail ?thumbnail . }
FILTER(regex(?target, '^Saturnight', 'i') && lang(?name) = 'en' && lang(?target) = 'en' && lang(?abstract) = 'en')  }
} LIMIT 1 OFFSET 0
```

Şekil 4. "Who produce Saturnight?" doğal dil sorusuna karşılık üretilen SPARQL sorgusu



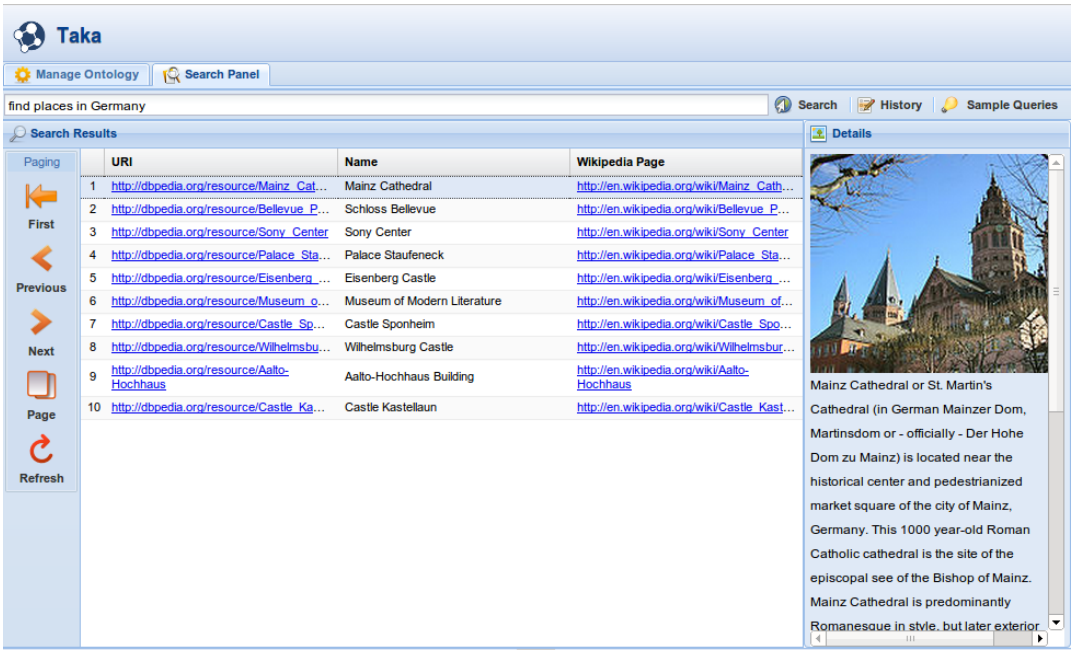
Şekil 5. "Who produce Saturnight?" sorgusunun cevaplarını gösteren ekran görüntüsü

İki örnekte görüldüğü üzere hem ön tanımlı desenler hem de sadece doğal dil işleme yöntemleriyle elde edilen sorgular birbirini tamamlamaktadır. Böylelikle ön tanımlı desenler aracılığıyla karşılığı bulunmayan doğal dil sorguları Levenshtein benzerlik algoritması ve doğal dil işleme yöntemleriyle cevaplandırılmaktadır. Benzer şekilde doğal dil işleme yöntemleriyle elde edilen yükleme karşılık DBpedia özelliği metin tabanlı bir benzerlik algoritması aracılığıyla elde edildiğinden içerik olarak benzeyen ancak anlamsal (semantik) olarak birbiri ile ilişkisi olmayan sonuçlar doğurabilmektedir. Geliştirilen sistemde bu tip durumlardaki açık, ön tanımlı desenler aracılığıyla kapatılmaya çalışılmıştır.

2.3 DBpedia Servisi Üzerinden SPARQL Sorgularının İşletilmesi ve Sonuçların Web Arayüzünde Gösterilmesi

Doğal dil sorularının algılanıp, ilgili SPARQL sorguları elde edildikten sonra DBpedia servisi programatik olarak kullanılabilmesi için açık kaynak kodlu semantik web kütüphanesi olan Apache Jena kütüphanesinden faydalanılmıştır. Apache Jena barındırdığı ARQ (A SPARQL Processor for Jena) modülü sayesinde belirlenen semantik modeli üzerinde SPARQL sorgularının yürütülebilmesi ve sonuçlarının elde edilmesini sağlamaktadır. ARQ modulü, SPARQL ve SPARQL/Update(SPARQL 1.1) standartlarını sorgu dili olarak kullanabilmekle beraber birleştirme (aggregation), Lucene metin arama motoru ile serbest metin arama (free text search) gibi eklentileri de desteklemektedir. [7] Elde edilen sonuçlar JSON (JavaScript Object Notation) formatına dönüştürülerek web arayüzünde listelenmiştir. Geliştirilen web arayüzü DBpedia servisi için geliştirilmiş web arayüzünde bulunmayan bir özellik olarak sayfalama (paging) desteği de sunmaktadır. Bu da sonuçların daha hızlı olarak listelenmesini sağlayarak yanıt hızı daha yüksek bir web (responsive web) ortamı sunmaktadır.

Şekil 6.'da sistem üzerinden yapılan "find places in Germany" sorgusuna yönelik cevapların listelendiği ekran görüntüsü sunulmuştur. Öncelikle sorgu, Tablo 1.'de listelenmiş olan ön tanımlı desenlerle eşleşip-eşleşmediği denetlenmektedir. Sonrasında ise doğal dil işleme yöntemleriyle aranan sorgu öğelere ayrıştırılmıştır. Bu örnek için sorgunun yüklemi "find", nesnesi "Germany" olarak bulunmuştur. DBpedia ontoloji modeli üzerinden bulunan yükleme en yakın DBpedia'da tanımlı özellik olarak "dFe" bulunmuştur. Ön tanımlı desenden ise sorguya ait özellik olarak "locationCountry" elde edilmiştir. Hem ön tanımlı desenle hem de doğrudan doğal dil işleme ve Levenshtein benzerlik algoritmaları yaklaşımlarıyla elde edilen sorgular birleştirilerek Şekil 7.'de görüntülene nihai sorgu elde edilmiştir. Burada ontoloji üzerinden ön tanımlı desenler kullanılmadan oluşturulacak olan SPARQL sorgusu, dFe özelliğinin find yükleminin anlamsal olarak karşılığı olmadığından kullanıcıyı aradığı sonuca götüremeyecektir. Burada da ön tanımlı desenlerin önemi birkez daha ortaya çıkmaktadır.



Şekil 6. Sorgu sonuçlarını detayları ve resim önizlemesiyle beraber görüntülemeyi sağlayan web arayüzü

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX :<http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbont: <http://dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT * WHERE {
  { ?x rdf:type dbont:Place . ?x dbpedia2:locationCountry ?tk . ?tk dbpedia2:commonName ?cName . ?x dbpedia2:name ?name . ?x dbont:abstract ?abstract .
    OPTIONAL { ?x dbont:thumbnail ?thumbnail . ?x foaf:isPrimaryTopicOf ?wiki . } FILTER(regex(?cName, '^Germany', 'i') && (lang(?abstract) = 'en'))
}
UNION
  { ?x dbont:fc ?y . ?y rdfs:label ?target . ?x rdfs:label ?name . ?x dbont:abstract ?abstract .
    OPTIONAL { ?x foaf:isPrimaryTopicOf ?wiki . ?x dbont:thumbnail ?thumbnail . }
FILTER(regex(?target, '^Germany', 'i') && lang(?name) = 'en' && lang(?target) = 'en' && lang(?abstract) = 'en')  }
} LIMIT 10 OFFSET 0
```

Şekil 7. Şekil 5.'de görüntülenen doğal dil sorgusuna yönelik sistem tarafından üretilen SPARQL sorgusu

Sonuç bulunamayan aramalarda kullanıcıya daha önceki aramalardan benzer olanları sunulmaktadır. Benzerlik tespitinde aranan nesne türü baz alınmaktadır. Böylelikle yazım hatası gibi son kullanıcıdan kaynaklı hatalar elimine edilmiş olmakta ve kullanıcı sonuca ulaşması için yönlendirilmektedir. Şekil 8.'de "who is Brack?" araması sonucunda yazım hatasından kaynaklı sonuç bulunmama durumunda kullanıcıya yöneltilen önceki benzer aramaların listelendiği ekran görüntüsü sunulmuştur.

Şekil 8. Yazım hatasından kaynaklanan sonuç bulunamama durumuna karşılık sistemin geçmiş aramalar baz alınarak arama önerilerini gösteren ekran görüntüsü

3 Benzer Çalışmalar

Ell ve arkadaşları [16] SPARTIQULATION isimli SPARQL sorgularını son kullanıcı tarafından da okunabilmesi, anlaşılabilmesi doğal dil ifadelerine dönüştüren ve bunu sözsel bir ifadeyle son kullanıcıya sunan bir sistem geliştirmiştir. SPARTIQULATION, doğal dil üretim sistemlerinden (Natural Language Generation - NLG) esinlenerek ve dil bilimi ağırlıklı olarak kullanılarak geliştirilmiş. Soru cevaplarken geliştirdiğimiz sistemle bu sistem benzer bir yaklaşım öne sürmektedir. Aranan özne öncelikle en genel tip olan herhangi bir şey ("thing") olduğu kabul edilerek, sonrasında soruda var olan kısıtlamalara göre öznenin türü şekillendirilmektedir. SPARTIQULATION, sadece RDF ve RDFS tabanlı bir yaklaşım sunmaktadır, bu da FOAF, OWL gibi gelişmiş ve oldukça sık kullanılan kütüphanelerden faydalanılamamasına sebep olmaktadır. Ayrıca mesaj tipleri sistemde gömülü olarak çalışması esneklik açısından bir dezavantaj oluşturmaktadır. Bizim sistemimizde eğitim verileri, kullanılan Apache OpenNLP kütüphanesi sayesinde sürekli güncellenmekte ve gelişmektedir. Sistemin oluşturulan SPARQL sorgularının sözsel ifadesinin elde edilip, son kullanıcıya sunulması bizim geliştirdiğimiz sistemden başlıca farkıdır ve bu aşama kullanıcıya hangi algıyla cevap verildiğinin aydınlatırken soru cevaplama hızını düşüreceği düşünülmektedir.

Lopez ve arkadaşları [9] PowerAqua adında soru cevaplayan sistem sunmuştur. Bu sistem sorguları işlerken birden çok veri kaynağı (datasource) kullanmaktadır. Bu yaklaşım heterojen veri kaynakları ve veri modelleri sebebiyle uyumsuz sonuçlar doğurma gibi bir tehlike barındırmaktadır. Geliştirilen sistem veri kaynağı olarak verinin homojenliğinden faydalanmak ve uyumsuz ve düşük kaliteli sonuçlardan korunmak içinde temelde sadece DBpedia'yı kullanmaktadır.

Damljanovic ve arkadaşları [10] FREyA isimli bir sistemi sunmuştur. Bu sistemin en büyük dezavantajı sistemin soruları cevaplamaya başlayabilmesi için eğitimden geçmesi gerekliliğidir. Sistemin soruları cevaplama mekanizması kullanıcıların geri bildirimlerine (feedback) dayanmaktadır. Bu da beklenmeyen ya da tutarsız sonuçlara sebebiyet verebilecektir. Geliştirilen sistem eğitilmeye ihtiyaç duymamakta, herhangi bir geri bildirime ihtiyaç duymaksızın soruları cevaplamaya başlamaktadır. Geliştirdiğimiz sistemin diğer bir avantajı ise DBpedia servisi dışında herhangi bir harici bağımlılığının (dependency) olmamasıdır. Dolayısıyla bizim sistemimiz daha kararlı ve güvenilir kabul edilebilir.

Ferrucci ve arkadaşları [11] Watson adında tanınmış ve komplek bir sistem sunmuştur. Watson'un diğer sistemlerden farklı olan yaklaşımı, soruları birçok alt parçalara ayırması ve sonunda bu parçaları birleştirip, sonuçları değerlendirmesidir. IBM'in bildirdiğine göre Watson, doğal dilleri analiz etmek, kaynaklarını teşhis etmek, varsayımları (hypotheses) bulmak, üretmek ve bu varsayımları birleştirip değerlendirmek için 100'den fazla farklı tekniğe sahiptir. [12]

Unger ve arkadaşları [13] şablon sorgular temeline dayanan bir sistem sunmuştur. Bu yaklaşım ön tanımlı şablonlar sebebiyle genel sorgulara yönelik bazı kısıtlamalar getiriyor olsa da bu yaklaşım özellikle kompleks sorgularda yanıt oranı daha yüksek bir soru cevaplama sistemi sunmaktadır. Çünkü her zaman için sorunun yüklemine uygun özelliği bulmak mümkün olmamaktadır. Bu yaklaşımda sorulara yanıt, şablonu tanımsız sorularda daha genel de olsa her zaman vardır ve yükleme uygun özellik bulunamasa bile tüm sorular cevaplandırılmaktadır.

Pythia [14] birçok soru cevaplama sisteminden farklı bir yaklaşım öne sürmüştür. Dilbilimsel analizler temeline dayanan sistem kendi sözlüğünü oluşturmakta ve kompleks soruları dilbilimsel yaklaşımlarla ele almaktadır.

Yahya ve arkadaşları [15] soruları kelime gruplarına ayırıp, sonrasında kelime gruplarını semantik kaynaklara çeviren bir metot öner sürmüştür. Geliştirdikleri sistem kompleks sorguları bu şekilde ele almaya çalışmaktadır.

4 Tartışma

Tüm doğal dil sorularına cevap verebilecek bir sistem şuan için geliştirilmemiştir ve %100 başarıya sahip bir sistem birçok dil bilimcisi ve bilgisayar bilimcisi tarafından öngörülmemektedir. Çünkü doğal dilde sorulan bir soru aynı manada birçok şekilde sorulabilmekte ve bu soruların hepsinin tek bir sorguya denk gelmektedir. Bu çok çeşitliliğin anlaşılması şuan için üzerinde çalışılan konular arasındadır. Örnek verecek olursak, “Türkiye'nin nüfusu kaçtır?” sorusunun semantik web SPARQL

sorgu karşılığı şu üçlüden oluşmaktadır: |Özne : Turkey| |Özellik: dbprop:populationEstimated| |Nesne: ?x|. Aynı soru "Türkiye'de ne kadar insan yaşamaktadır?" şeklinde de sorulması mümkündür. Bunun gibi birçok örnek verilebilir ve tüm bu çeşitliliği algılayıp, uygun forma dönüştürmek şuan için çözüm bulunamamış bir konudur. Geliştirilen sistemde bu durum sorunun nesnesi belirlendikten sonra ilgili direk yanıt bulunamasa bile nesneye yönelik bilgiler görüntülenerek aşılmaya çalışılmıştır. Bu yaklaşımın kullanıcının sorusunu yanıtsız bırakmak yerine kullanıcıyı aradığı cevaba bir adım daha yakınlaştıracağı düşünülmektedir.

Bu zamana kadar birçok doğal dil sorularını cevaplayan sistemler üzerinde çalışmalar yapılmış ve her çalışmanın izlediği kendine özgü bir yaklaşım bulunduğu gözlemlenmiştir. Ön tanımlı desen kullanmayan sistemlerde tüm sorulara yönelik uygun DBpedia özellik (property) eşleştirmesi yapılması mümkün olmadığından özellikle kompleks sorgularda sistemlerin soruları yanıtsız bıraktığı gözlemlenmiştir. Bu aşamada ise tamamen doğal dil işleme yöntemleri ve DBpedia ontolojisinden faydalanılarak elde edilen sorgulardan faydalanılmaktadır. Bu iki yaklaşım birbirini eksikliklerini tamamlamakta önem arz etmektedir. Ayrıca yüklemle ilişkili DBpedia özelliği bulunamasa bile geliştirilen sistem nesneye ait sonuçları görüntülemektedir. "Sonuca yakın cevap, cevapsız bırakmaktan daha iyidir" yaklaşımı temel alınarak kullanıcıyı sonuca iletecek yanıtlar listelenmiştir. Ayrıca DBpedia servisinde bulunmayan ancak geliştirilen sistemde sunduğumuz sayfalama seçeneği sayesinde arama sonuçları kullanıcıya daha hızlı bir şekilde gösterilmekte ve sunucu daha performanslı olarak çalışabilmektedir.

5 Sonuçlar

Geliştirdiğimiz semantik web projesi ile doğal dil sorularını algılayan ve uygun SPARQL sorgularına çevirerek DBpedia servisi üzerinden cevaplandıran bir sistem geliştirdik. Doğal dil işleme yöntemlerimden faydalanarak algılanan sorular, tanımlı desenler üzerinden sınıflandırılmıştır. Soruların yüklemleri ve nesneleri ile DBpedia üzerinden tanımlı olan ontolojiye çevrilmesi tam olarak eşleşemeyen durumlarda aranan cevaba götürecek yakın cevaplar listelenmiştir. Geliştirilen sistem tüm doğal dil sorgularını cevaplayabilir durumda olmasa bile, elde ettiğimiz ilk test sonuçları çok umut verici ve ileriye yönelik geliştirme için teşvik edici olmuştur. Geliştirilen sistem http://app.ibu.edu.tr:8080/taka adresinden kullanıma açıktır.

Geliştirme olarak, daha fazla soru deseni tanımlanarak sistemin daha fazla soru çeşidine cevap vermesi iyileştirme hedefleri arasında gözükmektedir. Ayrıca birden çok veritabanından faydalanmak, daha kompleks doğal dil sorularını cevaplayabilmek için dış bağlantıları doğru tanımlanmış sorgular üretebilmek için veritabanları üzerindeki yapısal tanımlamaların ilişkilendirilmesi ve başta Türkçe olmak üzere diğer doğal dil sorgularını işleyebilme geliştirilen sistem için gelecekteki iyileştirme hedefleri arasında yer almaktadır.

Kaynaklar

1. Berners-Lee T., Hendler J., and Lassila O., Scientific American, 2001.
2. Metz, C.: Tim, Lucy, and The Semantic Web, http://www.pcmag.com/article2/0,2817,2102852,00.asp, (2007).
3. 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: Science, Services and Agents on the World Wide Web, 7(3), 154-165 2009.
4. DBpedia Wiki, http://dbpedia.org/About.
5. Linking Open Data cloud diagram, http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.png, September 2011.
6. Banker, K.: MongoDB in Action. Manning Publications (2011).
7. Lindörfer F., "Semantic web frameworks," pp. 1–5, 2010.
8. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. 7, 154–165 (2009).
9. Lopez V., Fernandez M., Stieler N., & Motta E.: PowerAqua : supporting users in querying and exploring the Semantic Web content. Semantic Web Journal, (2011).
10. Damljanovic D., Agatonovic M., & Cunningham H. FREyA: An interactive way of querying Linked Data using natural language. Proceedings of the 1st Workshop on Question Answering over Linked Data (QALD-1), (ESWC 2011).
11. Ferrucci D., Brown E., Chu-Carroll J., Fan J., Gondek D., Kalyanpur A. A., Welty C.: Building Watson: An overview of the DeepQA project. AI Magazine, 59–79, (2010).
12. Watson – A System Designed for Answers, IBM Whitepaper, (February 2011).
13. Unger C., Bühmann L., Lehmann J., Ngomo A.-C. N., Gerber D., & Cimiano P.: Template-based question answering over RDF data. Proceedings of the 21st international conference on World Wide Web - WWW 2012 - Ontology Representation and Querying: RDF and SPARQL, 639–648, (2012).
14. Unger C. and Cimiano P. Pythia: Compositional meaning construction for ontology-based question answering on the Semantic Web. In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011).
15. Yahya M., Berberich K., Elbassuoni S., Ramanath M., Tresp V., & Weikum G.: Natural Language Questions for the Web of Data. Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP 2012).
16. Ell B., Vrandecic D., and Simperl E.: Spartiqulation: Verbalizing sparql queries. Proceedings of ILD Workshop, (2012).