Doğruluk Problemi İçin Veri Kümesi Hazırlanması

Arif Kürşat Karabayır, Ozan Onur Tek, Özgür Fırat Çınar, ve Selma Tekir
arifkarabayir@gmail.com, ozanonurtek@gmail.com,
ozgurfiratcinar@gmail.com, selmatekir@iyte.edu.tr

İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü 35430 Urla/İzmir

Özet. İnternet günümüzde en önemli bilgi kaynaklarından biri haline gelmiştir. Ancak İnternet'teki bilgiler her zaman güvenilir olmayabilir. Bilgiye ulaşımın ve paylaşımın kolaylaşması, çelişkili bilgilerin açığa çıkmasına sebep olmuştur. Çelişkili bilgilerin artmasıyla, bunlar arasında doğru olanı bulmak da her geçen gün zorlaşmaktadır. Bu sorun ilk olarak 2008 yılında Doğruluk (Veracity) Problemi olarak tanımlanmıştır. Takip eden yıllarda bu problem üzerine çalışmalar sürdürülmüştür. Bu alanda geliştirilen algoritmalar girdi olarak yapısal veriyi kabul etmektedir. Bu algoritmaların İnternet üzerinde kullanılabilmesi için İnternet'teki yapısal olmayan verinin yapısal forma dönüştürülmesi gerekmektedir. İnternet'teki verinin çeşitliliği düşünüldüğünde bu işin konudan bağımsız, otomatik olarak gerçekleştirilmesi zordur.
Bu çalışmada, çelişkili bilgi içerme potansiyeli olan ilgi çekici bir alan belirlenerek bu alan için yapısal bir tanım kümesi oluşturulmuştur. Sonrasında İnternet'te, bu tanım kümesi ile ilgili bilgi sağlayan kaynaklar tespit edilmiştir. Kullanılacak kaynaklar seçilmiş, kaynaklardan veri çekilmiş, çekilen veri ön işlemeden geçirilmiş ve sonrasında yapısal bir veri kümesine dönüştürülmüştür. Özetle, doğruluk bulma algoritmalarının başarımını sınamak üzere kullanılabilecek bir veri kümesi oluşturulmuştur. Ayrıca bu veri kümesi hazırlama süreci kayıt altına alınarak İnternet'teki yapısal olmayan verinin doğruluğu doğrudan sınanabilecek yapısal veriye dönüştürme işinin daha iyi anlaşılması sağlanarak otomatize edilmesine yönelik bir katkıda bulunulmuştur.

Anahtar Kelimeler: Doğruluk Problemi · Veri Kümesi · Web

1 Giriş

İnternet günümüzde hayatımızın önemli bir parçası haline gelmiştir. İnsanlar gündelik hayatında internette önemli bir süre geçirmektedir. Bu durum birçok insanın interneti önemli bir bilgi kaynağı olarak görmesine neden olmaktadır. İnternetteki bilgilerin ve bilgi sağlayan kaynakların sayısı da gitgide artmaktadır. Fakat internetteki bilgilerin güvenilir olduğunun bir garantisi yoktur. Popüler kaynakların güvenilir bilgi sağlayacağı düşünülebilir fakat yapılan araştırmalar bu varsayımın her zaman doğru olmadığını göstermiştir [1].

İnternetteki çelişkili bilgilerden doğru olanı bulup, kaynakların güvenilirliğini saptamak için 2008 yılında Doğruluk Problemi [1] tanımlanmış ve bu problemin çözümü için bir model oluşturulmuştur. Takip eden yıllarda bu problem üzerine çalışmalar sürdürülmüştür [2,3,4,5]. Bu alanda geliştirilen algoritmalar girdi olarak yapısal veriyi kabul etmektedir. Ancak internetteki verilerin büyük kısmı yapısal olmayan formdadır. Bu yüzden, bu konudaki algoritmaların doğrudan kullanımı mümkün değildir. Algoritmaların sınanması için önceden hazırlanmış birkaç yapısal veri kümesi kullanılmaktadır.

Bu çalışmada, bir alan belirlenerek bu alan hakkında internette sunulan yapısal olmayan veri yapısal bir forma dönüştürülerek doğruluk bulma algoritmalarının sınanabileceği bir veri kümesi oluşturulmuş ve izlenen adımlar kayıt altına alınarak sürecin otomatize edilmesine yönelik bir katkıda bulunulmuştur.

Bildirinin geri kalan bölümünde ilk olarak doğruluk problemi tanıtılmakta ve ilgili literatür verilmektedir. Metodoloji bölümünde doğruluk bulma algoritmalarını sınamak üzere özdeyişler alanında veri kümesi oluşturma süreci anlatılmaktadır. Bildirinin sonuç bölümünde ise bulgular ortaya konmakta ve gelecek çalışmalar üzerine düşünceler dile getirilmektedir.

2 Doğruluk Problemi ve İlgili Literatür

Doğruluk problemi ilk defa 2008 yılında Yin vd. [1] tarafından ortaya konmuştur. Doğruluk, birden fazla kaynağın, (örneğin web sitesi) bir veya birden fazla nesne (örneğin kitap) üzerinde iddia ettiği bilgilerin (web sitesinin kitap hakkında iddia ettiği yazar ismi) doğruluğunun çıkarımı üzerine kurgulanmış bir problemdir. Bu problem kapsamındaki temel terimler aşağıda verilmektedir:

- **Kaynak**, bilginin sağlandığı yer.
- **İddia**, bir kaynağın nesne hakkında sağladığı veri.
- **Gerçek**, en yüksek güvenilirlik puanına sahip iddia.
- **Kaynak Güvenilirliği**, bir kaynağın doğru bilgiyi verme olasılığı.
- **İddia Güvenilirliği**, bir iddianın gerçek olma olasılığı.

Bugüne kadar geliştirilen doğruluk bulma algoritmaları iki farklı veri tipi üzerine yoğunlaşmıştır. Bunlar sayısal ve kategorik verilerdir.

Sayısal Veriler genellikle bir ölçümü veya miktarı ifade eder. Bu tipteki veriler için kesin doğru veya kesin yanlış demek doğru değildir. Bunun yerine bu verilerin doğruluğu, kesinlik oranı ile ifade edilir. Örnek olarak, bir nesnenin bir özelliği hakkında 25 ve 90 olarak iki farklı iddia varsa ve gerçek değer 100 ise, 90 doğruya yakın sayılıp 25'e göre iddia güvenilirlik puanı yüksektir. **Tablo 1**'de sayısal veri tipinde bir veri kümesi örneği verilmektedir.

Kategorik Veriler ise daha çok nesnelerin karakteristiğini ifade ederken kullanılır. Bu tipteki veriler ya doğru ya da yanlış olarak sınıflandırılır. Kategorik

Tablo 1. Sayısal Veri Kümesi Örneği - Nüfus Veri Kümesi

Nesne	Kaynak	İddia
Abu Dhabi	Contributor#1513217: Mohammed	1850230.0
Amsterdam	Contributor#141597: Ilse@	741329.0
Amsterdam	Contributor #1300620: Krator	742884.0
Adelaide	Contributor #3922171: Pirate05	1124315.0
Athens	Contributor #1876487: El Greco	4200000.0
Athens	Contributor #3007532: Theiasofia	4242000.0
Athens	Contributor #1876487: El Greco	745514.0
Athens	Contributor #0 (87.203.23.2)	745514.0
Athens	Contributor #1876487: El Greco	745514.0
Navalcán	Contributor#363486: Emijrp	2238.0
Navalmoralejo	Contributor #363486: Emijrp	63.0
Los Navalmorales	Contributor#363486: Emijrp	2636.0
Los Navalucillos	Contributor#363486: Emijrp	2636.0

Tablo 2. Kategorik Veri Kümesi Örneği - Özdeyiş Veri Kümesi

İddia	Nesne	Kaynak
George Bernard Shaw	Beauty is a short-lived tyranny	famous-quotes
Socrates	Beauty is a short-lived tyranny	brainyquote
Albert Einstein	It is strange to be known so universally and yet to be so lonely.	famous-quotes
Napoleon Bonaparte	A soldier will fight long and hard for a bit of colored ribbon.	quotables

veri tipi ile ilgili örnek veri kümesi **Tablo 2**'de sunulmaktadır.

Doğruluk probleminin çözümü için tasarlanan algoritmalar bu kavramlar üzerinden modellerini oluşturmuştur. Her bir çalışma konuya farklı bir yönden yaklaşmakta olup, aralarında bazı temel farklılıklar bulunmaktadır. Bu çalışmalar ve öne çıkan farklılıkları takip eden bölümde açıklanmaktadır.

Yin vd. [1] tarafından geliştirilen model temel "TRUTHFINDER" modelidir ve kaynak-iddia ilişkilendirmesi üzerinde çalışarak her kaynak ve her iddia için belirli bir güvenilirlik puanı ataması yapar ve bu puanları yinelemeli bir şekilde günceller. Bu kapsamda, kaynak güvenilirliği ($t(\omega)$) değeri, $F(\omega)$, ω kaynağının sağladığı iddiaların güvenilirlik $s(f)$ toplamının; $F(\omega)$, ω kaynağınının sağladığı iddiaların toplam sayısına oranıdır:

$$t(\omega) = \frac{\sum_{f \in F(\omega))} s(f)}{|F(\omega)|} \tag{1}$$

Daha sonra, bulunan kaynak güvenilirliği $t(\omega)$, iddia güvenilirliğini hesaplamak için kullanılır.

$$s(f) = \sum_{w \in W(f))} t(\omega) \tag{2}$$

Bu modelin en önemli noktası, bir nesne hakkındaki farklı iddiaların birbirinin güvenilirlik puanına etki etmesidir. Bulunan iddia güvenilirlik puanları bu etkiler göz önünde bulundurularak güncellenir.

Zhao ve Han [2] doğruluk problemini çözmek için "GTM" adında bayesçi olasılıksal bir model oluşturmuştur. Geliştirilen model, kategorik veri üzerine yoğunlaşan önceki çalışmaların aksine özellikle sayısal verilerde daha başarılı sonuçlar vermektedir.

Doğruluk problemi için hazırlanan veri kümelerinde, kaynaklardan olabildiğince çok veri sağlanması temel bir gereksinimdir. Ancak pratikte her kaynak yeteri kadar veri sağlamayabilir. Bu durum doğruluk bulma algoritmaları için çözülmesi gereken bir problemdir. Li vd. [3] çalışmasında, az sayıda iddiaya sahip kaynakların da bulunduğu bir veri kümesinde doğruluk bulma başarımını artıracak bir model öne sürülmüştür. Geliştiren model hem kategorik hem de sayısal veri kümelerinde çalışabilmektedir.

Xin vd. [4] kaynakların verileri birbirinden kopyalama durumu üzerine odaklanmıştır. İnsanlar sezgisel olarak bir iddiayı ne kadar çok kaynakta görürse o iddianın doğruluğundan o kadar emin olurlar. Ancak kaynakların verileri birbirlerinden kopyalamaları durumunda yanlış bilgiler hızla yayılabilir ve doğru bilgiyi saptamak bu durumda oldukça zorlaşabilir. Çalışmada, iddiaları birbirinden kopyalayan kaynakların tespit edilmesi ile gerçek iddiaların saptanmasını kolaylaştıran bir model geliştirilmiştir.

Qi vd. [5] doğruluk bulma algoritmalarında sayısal ve kategorik verileri farklı şekillerde ele almıştır. Literatürde her iki veri kümesi için de tasarlanan çalışmalar mevcut olmasına rağmen ilk defa bu çalışmada veri kümesinin heterojen olduğu bir başka ifade ile her iki veri tipini de kapsadığı durumda çalışabilecek bir model geliştirilmiştir. Modelin heterojen veri kümesinde çalışabilmesi, farklı veri tipleri için farklı uzaklık fonksiyonlarının tanımlanabilmesi ile sağlanmıştır.

3 Metodoloji

Doğruluk bulma algoritmalarının çalışması için yapısal bir veri kümesi gerekmektedir. Ancak internetteki verilerin çoğu yapısal olmayan formdadır. Bu kısımda, internetteki yapısal olmayan verilerin, doğruluk bulma algoritmalarının sınanabileceği, yapısal bir forma dönüştürülmesi aşamalandırılmıştır.

3.1 Tanım Kümesinin ve Kaynakların Belirlenmesi

İdealde internetteki hemen her konunun tanım kümesi olarak seçilebileceği düşünülebilir. Bununla beraber, doğruluk bulma algoritmalarının mevcut çalışma yapıları gereğince tanım kümeleri için belirli şartların sağlanması daha sağlıklı sonuçlar vermektedir. Bu şartlar şu şekilde sıralanabilir;

- Nesneler hakkındaki iddiaların değişiklik gösterebilmesi,
- Tanım kümesi ile ilgili birçok kaynaktan veri elde edilebilmesi,

– Her nesne hakkında tek bir kesin doğru veri bulunması.

Bu şartlar göz önünde bulundurularak tanım kümesi olarak "Özdeyişler" belirlenmiştir.

Tanım kümesi belirlendikten sonraki aşama, verilerin çekileceği kaynakların seçilmesidir. Tanım kümesi kapsamında birçok kaynak bulunabilir ancak bu kaynakların hepsi veri kümesi oluşturulması için uygun olmayabilir. Özdeyiş tanım kümesi için bu çalışmada, internetteki popüler özdeyiş paylaşım siteleri tespit edilmiştir. Ayrıca sosyal medyada paylaşılan verilerin ne kadar güvenilir olduğunu görebilmek adına özdeyiş paylaşan belirli Twitter hesapları da kaynaklara dahil edilmiştir. Bu kaynakların belirlenmesinde;

– Kaynakların sözdizimi ve yazım kurallarını doğru kullanması,
– Verileri sağlarken tamamen 2. bir kaynağa bağımlı kalmaması,
– Yeterli sayıda veri sağlamış olması,
– Verilerin sağlandığı formatın tutarlı olması,

gibi kriterler dikkate alınmıştır.

3.2 Verilerin İşlenmesi

Kaynakların belirlenmesinin ardından kaynaklara ait programlama arayüzleri (API) tespit edilmiştir. Bu programlama arayüzleri kullanılarak veriye ulaşılmış ve veri düzenli bir formatta kayıt altına alınmıştır.

3.1'de de belirtildiği üzere, kaynaklar belirli kriterler çerçevesinde seçilmiş olmasına rağmen, kayıt altına alınan veri incelendiğinde formata uygun olmayan veya konu dışı içerikler, özetle kirlilik (noise) tespit edilmiştir. İlk basamakta, bahsedilen bu kirliliğin temizlenmesi için Jaro-Winkler Distance algoritmasından yararlanılmıştır. Jaro-Winkler distance temelde Jaro-Winkler Benzerliği üzerine oturtulmuş iki karakter dizisinin benzerliğini tespit etme algoritmasıdır. Jaro-Winkler Benzerliği ise temelde Jaro Benzerliği'ni kullanır. Bu benzerlik matematiksel olarak aşağıdaki formül ile hesaplanır;

$$sim_j = \begin{cases} 0 & if\ m{=}0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) & otherwise \end{cases} \tag{3}$$

s_i karakter dizisinin uzunluğunu,
m eşleşen karakter sayısını,
t karakterler için yer değişikliğini ifade eder.

İki dizinin birbiriyle eşleşir olarak sayılması için;

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1 \tag{4}$$

sağlaması gerekir. Jaro-Winkler Benzerliği ise bu noktadan yola çıkılarak;

$$sim_w = sim_j + (\ell p(1 - sim_j)), \tag{5}$$

eşitliği ile verilir. Bu eşitlikte;
sim_j *Jaro benzerliğini,*
l *ortak önek karakterlerin uzunluğunu,*
p *sabit ölçekleme faktörünü ifade eder.*

Son olarak Jaro-Winkler Distance değeri ise;

$$d_w = 1 - sim_w \tag{6}$$

şeklinde ifade edilir. Bu algoritmanın kullanımı sırasında, tüm veri ikililerinin benzerlik ilişkisine bakılmıştır. Benzerlik ilişkisi sayısal olarak çok düşük olan veriler, kümeden silinmiştir. Kirlilik minimize edildikten sonra, veri kümesi formatı bozulmadan tekrar kaydedilmiştir. İkinci basamakta ise bir kaynağın aynı iddiayı birden fazla kez barındırma sorunuyla karşılaşılmıştır. Bu sorun N-gram benzerlik algoritmasıyla, benzerlik ilişkisi sayısal olarak birbirine çok yakın ve aynı kaynaktan iddia edilmiş verilerin temizlenmesiyle çözümlenmiştir. Son basamakta ise özdeyişler, yine N-gram benzerlik özelliğinden yararlanılarak, özdeyiş-söyleyen ilişkisine göre gruplandırılmış ve TRUTHFINDER algoritması için uygun bir hale getirilmiştir. N-gram benzerlik algoritması, bir kaynak metinde, n uzunluğundaki tüm komşu kelime veya karakter gruplarının kombinasyonlarını baz alan yöntem olarak ifade edilir.

4 Sonuç

4.1 Bulgular

Yapılan bu çalışmanın temel amacı, doğruluk problemi ve ilgili algoritmaları destekleyici bir alt yapı oluşturmak ve söz konusu alana katkı sağlamaktır.

Bahsedilen temel kazanımların sağlanması için öncelikli olarak alan üzerindeki benzer çalışmalar incelenmiş ve varolan yaklaşımlar farklı veri kümelerinde sınanmıştır. Kitap-yazar, şehir-nüfus gibi eşleşmeler içeren kategorik ve sayısal veri kümeleri üzerinde çalışılmıştır.

Çalışma kapsamında oluşturulan özdeyişler içeren yeni bir veri kümesi ile truth-finder algoritması çalıştırılmıştır. Truth-finder algoritması nesne-gerçek çiftlerine ihtiyaç duyduğu için kişilerin nesne, özdeyişlerin ise gerçek olarak kullanılması planlanmıştır fakat, bu durum truth-finder algoritmasının genel sezgisel yaklaşımlarından en temeli olan "Her nesne yalnızca bir tane gerçek doğruya sahiptir." önermesine ters düştüğü için bu fikir terkedilmiş ve tam tersi bir analoji ile devam edilmiştir.

Truth-finder algoritması çeşitli kaynaklardan alınan, özdeyişlerden oluşan veri kümesinde çalıştırılmıştır. Elde edilen, "kaynakların güvenilirlik skorları" **Tablo 3**'de verilmektedir.

Tablo 3. TRUTHFINDER algoritmasının özdeyiş veri kümesi üzerindeki sonuçları

Kaynak	Kaynak Güvenilirliği
AboutCheGuevara	0.680000
FranzKafkaQtss	0.680430
SFreudSayings	0.692487
mahatmaa_gandhi	0.695539
GreatestQuotes	0.691252
QuoteDaily	0.711092
NietzscheQuotes	0.718113
einsteinquotes	0.719569
PureNietzsche	0.725397
NietzscheQ	0.738335
gandhhii	0.767866
quotables	0.827372
famous-quotes	0.832783
brainyquote	0.884118
Einsteiin_Quote	0.980225
GandhiiQuotes	0.981335
einssttein	0.983000

4.2 Gelecek Çalışmalar

Özellikle son yıllarda -internetteki bilgi kirliliğinin de artmasıyla- doğruluk problemi önem kazanmıştır ve bu önemini koruyacağı öngörülmektedir. Bu konuda yapılan hem akademik hem de kurumsal çalışmaların sayısı her geçen gün artmaktadır.

Haberin doğruluğunu editörler aracılığı ile kontrol eden, bunu kamuoyu ile paylaşan, çalışmalarına Türkiye'de başlamış, Uluslararası Doğruluk Kontrolü Ağı (International Fact-checking Network) tarafından yayınlanan İlkeler Kılavuzu'nu (Code of Principles) tanıyarak imzalayan teyit.org Facebook'un üçüncü taraf doğrulama programının Türkiye uygulayıcısı olarak duyurulmuştur [6].

Çalışmanın devamında, yukarıda verilen örnek gibi; haber kaynakları tarafından iletilen bilgilerin doğrulanmasında ya da ansiklopedik veri kaynaklarının doğruluğunun denetlenmesinde kullanılabilecek otomatize araçlar geliştirilebilir.

Sistem, gerçek zamanlı veri üzerinde ve/veya farklı veri tipleri üzerinde çalışabilir hale getirilerek (sıralanmış veri, sayısal-kategorilenmiş veri, grafiksel veri vs.) otomatize çalışan bir çok platform için bir iskelet oluşturabilir.

Kaynaklar

1. Xiaoxin Yin, Jiawei Han, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web.
2. Bo Zhao, Jiawei Han. 2012. A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources
3. Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data

4. Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava. 2009. Integrating Conflicting Data: The Role of Source Dependence
5. Qi Li, Yaliang Li, Jing Gao, Bo Zhao2, Wei Fan and Jiawei Han. 2014. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation
6. https://teyit.org/facebookun-dogrulama-programi-turkiyede-teyit-org-is-birligiyle-hayata-geciyor/