İnsan ve Makine Bulutları Sinerjisi: Kitle Kaynaklı Çalışma ile Veri Temizleme Örnek Uygulaması

Deniz İren[1], Gökhan Kul[2] ve Semih Bilgen[3]

[1, 2]Bilgi İşlem Daire Başkanlığı, Orta Doğu Teknik Üniversitesi, Ankara
[3]Elektrik ve Elektronik Mühendisliği, Orta Doğu Teknik Üniversitesi, Ankara

[1]diren@metu.edu.tr, [2]gkul@metu.edu.tr,
[3]bilgen@metu.edu.tr

Öz. Kitle Kaynaklı Çalışma (KKÇ) ve Bulut Bilişim bilgi teknolojilerinde önem kazanan kavramlar arasında yer almaktadır. İnsan ve makine bulutlarının karma kullanımıyla bir sinerji oluşturulması ve bu şekilde iki dünyanın güçlü tarafları öne çıkarılırken zayıf noktalarının da üstesinden gelinmesi mümkün kılınabilir. Bu makale Orta Doğu Teknik Üniversitesi'nde kullanılmakta olan, ancak kullanım ömrünü tamamlamaya yakın bir akademik yayın kayıt takip yazılımının güncel sürümünün geliştirilmesi sürecinde karma bir çözüm kullanımını konu almaktadır. Bu yöntem yazılımın yeni sürümünün geliştirilmesine paralel olarak, kayıtlı tutulan büyük miktarda verinin temizlenerek yeni sisteme aktarılmasında kullanılmıştır. Veri aktarımı için geliştirilmiş olan kullan-at prototip sistem ile 53,822 akademik kayıt temizlenmiş ve aktarılmıştır. Bu çözümün ilk adımı harici web servislerinden Sayısal Nesne Kimliği (Digital Object Identifier) sorgulanması ve kayıtların etiketlenmesinden oluşmaktadır. İkinci adımda ise bu çözüm için özel tasarlanmış dizgi benzerlik algoritması ile kalan kayıtlar filtrelenmiştir. Son olarak benzer ancak birebir aynı olmayan kayıtlar KKÇ yöntemi kullanılarak analiz edilmiş ve veri kümesindeki tekrarlar elenmiştir. Bu karma yöntem sayesinde projede, yalnızca makinelerin kullanıldığı bir çözüm ile ulaşılamayacak bir doğruluk seviyesine erişilebilmiş ve iş yalnızca insanların kullanıldığı bir çözüm ile erişilemeyecek bir hızda sonuçlanmıştır. Çözümün KKÇ fazında ulaşılan hata payı 6.4% olarak gözlemlenmiş ve insan ve makine bulutlarının sinerjisine Türkiye'deki kamu üniversitelerinde bir ilk örnek olan bu çalışmanın başarısı belgelenmiştir.

Anahtar kelimeler: Kitle Kaynaklı Çalışma, bulut bilişim, dizgi benzeştirme

Giriş

Teknolojinin ulaştığı son noktada pek çok iş bilgisayarlar tarafından çok hızlı ve verimli biçimde yürütülebilmektedir. Buna rağmen bazı işlerde bilgisayarlar hala insanların performansına ulaşamamıştır. Bu sayede, insan ve makinenin birlikte çalışmasından sinerji doğması beklenebilmektedir. Üstelik bu birliktelik geniş ölçekte

uygulandığında, makine bulutları ve insan bulutlarının oluşturacağı bir küresel beyin [1] kurgusunda bu sinerji daha da vurgulanacaktır.

1900'lerde uygulanmaya başlayan ve işçilerin basit görevlerde uzmanlaşması yaklaşımı, son yıllarda yazılım ve sistemlerin küçük, sınırları tanımlı, niş sorunlara çözüm üreten servisler şeklinde geliştirilmesine önayak olmuştur. Son on yıl içinde gerçekleşen bir diğer önemli gelişme ise Kitle Kaynaklı Çalışma'nın (KKÇ) ortaya çıkmasıdır. KKÇ'nin bir iş modeli olarak yaygın olarak uygulanmaya başlamasıyla birlikte kalabalığın bilgeliği ve insan bilişselliği faydalanılabilir ve ölçeklenmesi mümkün kaynaklar arasında yer almaya başlamıştır. Günümüzde KKÇ, tişört tasarımından ansiklopedi makalesi yazılmasına, uydu görüntülerinden kaza enkaz belirlemeden, orman yangınlarının tespitine çok değişik tipte sorunların çözüme kavuşturulmasında kullanılmaktadır. Bazı KKÇ platform sağlayıcılarının sunduğu program arayüzleri (API) ile, yazılımın işleyişi sırasında çeşitli işler insan bulutuna yaptırılıp, sonuçları yazılım tarafından kullanılmakta ve böylece gerçek zamanlı, karma bir insan – makine bulut çözümü oluşturulabilmektedir.

Bu makale Orta Doğu Teknik Üniversitesi'nde (ODTÜ) yürütülen bir veri temizleme ve aktarım işi sırasında uygulanan karma çözümü konu almaktadır. ODTÜ'de yaklaşık 2,500 akademik, 3,000 idari personel görev yapmaktadır. Öğrenciler de hesaba katıldığında ODTÜ Bilgi İşlem Daire Başkanlığı (BİDB) tarafından sunulan BT servislerinden yararlanan kullanıcıların sayısı 30,000'in üzerine çıkmaktadır. Yerleşkedeki BT yapısı bünyesinde çok sayıda güncellenmesi gereken eski uygulamalar ve bu uygulamalar tarafından kullanılan büyük miktarda veri barındırmaktadır. Yakın geçmişte bu uygulamaların güncellenmesi ve süreç otomasyon yaklaşımı ile bütünleştirilmesi için bir program başlatılmıştır. Bu büyük değişim bazı eski verilerin yeni geliştirilen sistemlere aktarılması gereksinimini doğurmaktadır.

Bu uygulamalardan biri olan CV-Akademik, 1990'larda hayata geçirilmiş olup, ODTÜ'nün kurulduğu tarihten beri üniversite bünyesinde yapılmış yayınların takibi için kullanılmaktadır. Bu uygulama kullanıcıların serbest metin biçiminde yayın başlıkları ve diğer bilgilerini girmesini gerektirmektedir. Ayrıca birden fazla yazar tarafından yazılmış yayınların bilgisi, yazarlar tarafından sisteme ayrı ayrı girilebildiğinden veriler tekrarlanmaktadır. Serbest metin girişinde yapılan yazım hataları ve tekrarlı kayıtların bu sebeple tutarsız olması, yeni geliştirilen uygulamaya veri aktarımı ve veri tabanının normalizasyonu için aşılması gereken bir güçlük teşkil etmektedir. Aktarılması ve temizlenmesi gereken 53,822 kayıt satırı bulunmaktadır.

Bu makalede anlatılan karma çözüm yaklaşımı harici bir servis olan CrossRef DOI sorgu web servislerini, kurum bünyesinde iyileştirilmiş dizi benzerlik hesaplama algoritmalarını ve KKÇ kullanımını içermekte ve bahsedilen, gerçek hayatta karşılaşılmış, veri aktarım sorununun çözülmesinde uygulanmıştır. Bu araştırmanın çözmeyi hedeflediği iş sorunu veri kümesinde bulunan hatalı kayıtların tespit edilmesi ve düzeltilmesi, ayrıca gereksiz tekrarların elenerek normalize edilmesi ve harici yayın depolarındaki standart yayın kimlikleri ile etiketlenmesidir. Araştırma hedefi ise KKÇ'nin yazılım mühendisliği pratiklerinin bir parçası olarak etkili ve verimli bir biçimde kullanılabileceğine dair bir kavramsal tanıt sunmaktır.

Makale şu şekilde düzenlenmiştir: Bölüm 1 araştırma ortamı, çözülmek istenen sorun ve önerilen çözüm yöntemini anlatarak konuya giriş yapmaktadır. Bölüm 2 literatüre geçmiş olan benzer uygulamalar hakkında bilgi sunmakta, Bölüm 3 ise önerilen karma çözüm hakkında detaylı bilgi vererek uygulanan yöntemi anlatmaktadır. Son olarak, Bölüm 4 araştırma sonuçlarını ve gelecekte yapılması hedeflenen çalışmaları belirtmektedir.

İlgili Çalışmalar

KKÇ ve bulut bilişim bileşenlerini içeren karma sistemler çağdaş BT uygulamalarında öncü bir etmen olarak değerlendirilmektedir [2]. Karma sistemleri ifade etmek için çok farklı terimler kullanılsa bile altta yatan fikir benzerdir: İnsan ve bilgisayar servislerinin sinerjisi ile katma değer yaratılmaktadır. İnsan ve bilgisayar servislerinin bütünleştirilmesi ile her iki tipte servisin sunabileceklerinin ötesinde, artırılmış bir servisin sunulması mümkün kılınabilmektedir [3]. Büyük ölçekli insan-bilgisayar iletişimi ile mümkün kılınan bu yeni varlık A. Bernstein tarafından küresel beyin olarak adlandırılmıştır [1]. Bulut ekosistemini servis katmanlarının bir yığıtı olarak tarif eden Lenk, KKÇ'yi bu katmanların en üstünde göstermeyi uygun görmüş ve servis-olarak-insan (human-as-a-service) olarak adlandırmıştır [4]. Lackermeir tarafından e-ticarete uyarlanarak tasvir edilen karma bulut mimarisi de servis-olarak-insan katmanı içermektedir [5]. Vukovic araştırmalarında bulut bilişim ile güçlendirilmiş bir KKÇ servisinden bahsetmiştir [6]. M. Bernstein internet aramaları için kullanıcı tecrübesinin iyileştirilmesi amacıyla otomatik sorgu madenciliği ve KKÇ içeren karma bir yapı kullanmıştır [7]. Bunların yanı sıra, literatürde karma yöntemlerin büyük veri sorunları [8], bilgi yönetimi [9], piyasa tahmin [10], kitlesel ortak çalışma [11], açık yenilikçilik [12] ve bilimsel sorunların çözümünde [13] kullanıldığına rastlanmaktadır.

Karma çözümler her zaman başarılı olmak zorunda değildir. Bernestein başarılı karma sistemler geliştirmek için, bu sistemlerin önemli özelliklerinin farkında olunması ve karma çözümleri geliştirmek ile geleneksel bilgisayar sistemlerini geliştirmek arasındaki farkların çok iyi anlaşılması gerektiğini vurgulamaktadır [1].

Karma Bir Çözüm Yaklaşımı

53,822 kaydın temizlenmesi işi, bir uzman tarafından el ile yapıldığında durumunda çok uzun zaman alacaktır. Bu sebeple KKÇ kullanımının zaman ve maliyet açısından verimlilik sağlayacağı öngörülmektedir.

CrossRef Veri Servislerinde DOI Sorguları

İlk aşamada CrossRef DOI sorgulama web servisleri kullanılmıştır. Bu harici web servislerini kullanan basit bir uygulama geliştirilmiştir. Uygulama, taşınan veri ve başlık boyutu verimliliği sağlamak için, her web servis çağrısında 20 kayıt gönderip,

yanıt almaktadır. Web servisinin parametrelerinden biri olan "bulanık arama" seçeneğinin seçilmesi sayesinde aramalar bire bir eş olan kayıtların yanı sıra benzer kayıtları da bulacak şekilde işletilmektedir. Web servis çağrısının yanıtları geldiğinde DOI bilgisi bulunan kayıtlar bu bilgi ile eşleştirilmektedir. İleriki aşamalarda DOI bilgisi var olan kayıtların benzerlik veya eşlik durumu bu alanların karşılaştırılması ile yapılabilecektir.

Tüm kayıtlar için DOI çözümleme süreci 40 saatte tamamlanmıştır. Web servis ile DOI sorgulama sürecinin sonunda 5,681 kayıt geçerli bir DOI ile eşleştirilmiş, 39,415 kayıt için DOI kaydı bulunamamış, 391 kayıt ise içerdikleri özel karakterler sebebiyle işlem görmemiştir. Geri kalan 8,335 kayıt ise DOI kaydı olması beklenmeyen (ör: ulusal yayınlar) yayınlar içermeleri sebebiyle işlem gören kayıt kümesine dâhil edilmemiştir.

Dizgi Benzerlik Hesaplamaları

İkinci aşamada kayıt benzerliği çeşitli algoritmalar ile değerlendirilmiştir. Bu durum özelinde kayıt eşitliği DOI bilgisi olmayan kayıtlar için başlık, yazar ve yayıncı alanlarının aynı olması olarak tanımlanmıştır. DOI'nin küresel geçerlilikte eşsiz olarak belirlenmiş bir anahtar veri olması sebebiyle karşılaştırmalarda öncelikli alan olarak kullanılmaktadır. Öyle ki, DOI bilgisi olan kayıtlar için yalnızca DOI alanlarının aynı olması yeterlidir. Bununla birlikte, dizgi karşılaştırma yoluyla benzer kayıtları tespit ederken doğru olmayan sonuçlara varmak da mümkün olmaktadır. Kayıt kümesinde aynı yayın için olmasına rağmen hatalı yazılmış kelimeler veya kısaltmalar yüzünden farklı olarak kabul edilen kayıtlar bulunabilmektedir. Bu yüzden kayıtların aynılığının değil, benzerliğinin tespit edilmesi hedeflenmiştir. Her kayıt diğer tüm kayıtlarla karşılaştırılarak her bir benzerlik durumuna bir benzerlik skoru ile benzerlik anahtar verisi atanmaktadır. Benzerlik skorunu hesaplamak için, Levenshtein Distance (LD) yöntemi ve bir Jaccard Index varyantı (JI') birlikte kullanılmıştır.

Hesaplanan benzerlik skorları hem LD hem de JI' için 1'e eşit olan kayıtlar, aynı kabul edilerek kayıt kümesinden çıkarılmıştır. Benzerlik skorları bir eşik değerle karşılaştırılarak hem LD hem de JI' değerleri belirli bir eşik değerden daha yüksek olanlar bir araya getirilerek benzerlik grupları oluşturulmuştur. Kullanılan eşik değer, büyük ölçüde kullanılan dile göre belirlenmektedir. Belirtilen algoritmalar çeşitli eşik değerleriyle 50 kayıtlık örnek gruplarla denenmiş ve sonuç olarak LD ve JI' için en uygun eşik değerlerinin sırasıyla 0.7 ve 0.5 olduğu gözlemlenmiştir. Test sonuçlarına göre, aynı yayına ait olan kayıtların tamamı aynı benzerlik grubunda yer almaktayken, gerçekte farklı yayınlara ait olmasına rağmen hatalı şekilde benzer kabul edilen kayıtların oranı yalnızca % 18 olarak hesaplanmıştır. Farklı olmasına rağmen benzer kabul edilme hatası, bir diğer hata olan, gerçekte benzer olan kayıtların farklı kabul edilme hatasına tercih edilmiştir. Bu tercihin nedeni, birinci tipteki hataların sonraki aşamalarda giderilebilir olmasına karşın, ikinci tip hata durumunda böyle bir olanağın olmamasıdır.

LD skoru 0.7 den küçük olan kayıtlar farklı kayıt olarak kabul edildiğinden benzerlik skoru 0.7 ye eşit veya daha büyük olanların dışında kalan kayıtlar, benzeri olmayan kayıtlar olarak nitelendirilip, kayıt kümesinden çıkarılmıştır.

Benzerlik skoru hesaplama aşaması tamamlandığında, 4.558 kayıt aynı olarak değerlendirilirken 38.830 kaydın benzersiz olduğu belirlenmiştir. Bu kayıtlar normalize edilmiş ve kayıt kümesinden çıkarılmıştır. Geriye kalan 10.434 kayıt daha tekrar değerlendirilmek üzere diğer aşamalara aktarılmıştır.

Levhenstein Distance (Levhenstein Aralığı).

LD değeri ile karşılaştırma yapılmadan önce dizgiler büyük harflere dönüştürülür. Özel karakterler ASCII karşılıklarına dönüştürülür veya dizgiden çıkartılır. Yazar (Author) alanı basit dizgi işlemleri kullanılarak standart hale getirilir.

LD, dizgileri karşılaştırmak ve bir dizgiyi diğerine çevirmek için yapılması gereken işlem sayısını temsil eden aralık değerini hesaplamak için kullanılır. LD Algoritması literatürde tanımlandığı şekliyle kullanılır [14], [15].

Jaccard Index Variant (Jaccard İndisi Varyantı).

LD hesaplama işlemi tamamlandıktan sonra kayıtları, içerdiği kelimelere göre karşılaştırmak için JI' kullanılır. Algoritmayı kullanmadan önce, "THE", "FROM", "FOR" kelimeleri ile 1 ve 2 harfli kelimeler kayıtlardan çıkartılır.

Jaccard İndisi'nin [16] bir varyantı olan JI', aynı kelimeleri farklı sıralama ile kullanmış olan dizgilerle ilgili hatalı sonuçlara varmayı engellemek amacı ile LD ile birlikte kullanılarak onu tamamlar.

Jaccard Göstergesi ile JI' arasındaki fark (1) ve (2) de gösterilmektedir.

(1) Jaccard İndisi = $A \cap B / A \cup B$
(2) JI' = $A \cap B / A, \quad |A| \geq |B|$

JI' algoritmasında farklılaşmaya gidilmesinin sebebi, algoritmanın yeni biçimiyle daha yüksek kesinlik ile kayıt farklılıklarını tespit edebilmesidir. Bu değişiklik sayesinde sık rastlanan hatalardan biri olan başlık, yayıncı ve yıl bilgilerinin tümünün bir arada başlık alanına girilmesi durumu elenebilmektedir. Jaccard İndisi, A U B kümesinde bulunan kelime sayısının fazla olmasından dolayı JI'ya kıyasla daha düşük bir benzerlik skoru hesaplar. Bu durum için JI', Jaccard İndisi'nden daha doğru sonuç vermektedir.

Kitle Kaynaklı Çalışma

Hem özelleştirilmiş web servisleri hem de algoritmalar tarafından sınıflandırılamayan kayıtlar, bir sonraki aşama olan KKÇ aşamasına aktarılmıştır. Bu aşamada insan algısının, benzer metin alanlarındaki farklılıkları teşhis edebilme yeteneğinden faydalanılması hedeflenmiştir.

Söz konusu 10,434 kayıt, benzerlik anahtar verilerine göre 4,359 gruba ayrılmıştır. Çeşitli büyüklüklerdeki bu gruplara ait olan kayıt sayıları Tablo 1'de gösterilmektedir.

Tablo 1. Benzerlik gruplarındaki yayın sayıları

Group Size	Record Count	Group Size	Record Count
15	1	8	9
14	0	7	14
13	1	6	32
12	1	5	53
11	3	4	248
10	6	3	637
9	9	2	3,345

KKÇ'nin tasarımının, işçilerin görev performansı üzerinde önemli etkisi olduğundan benzerlik grupları çiftler halinde yeniden düzenlenmiştir. Böylelikle, "İki kayıt aynı mı yoksa farklı mı?" biçiminde basit ve ikili yanıtlı sorular sorabilme olanağı bulunmuştur.

Kayıtları çiftler haline getirmek KKÇ görev sayısında artışa neden olmuştur. Bir grupta bulunan benzer kayıtlardan oluşturulacak benzerlik çitlerinin sayısı aşağıdaki formülle hesaplanabilir.

$$\#\text{ Benzerlik Çiftleri} = \text{Grup Boyu} \cdot (\text{Grup Boyu} - 1) / 2$$

Böylece toplam görev sayısı 9.308 olarak hesaplanmıştır. Bu görevler Amazon Mechanical Turk (AMT) platformunda yayınlanmıştır. Her görevde işçilerden dört benzerlik çifti içeren kayıt kümesini değerlendirmeleri istenmiştir. Başarıyla tamamlanan her görev için 0.02$ ödenmiştir. Görevlerdeki dörtlü çiftlerden birisi altın standart kümesinden seçilerek elde edilmiş iken kalan çiftler olağan kayıt çiftleri kümesinden seçilmiştir. Görevin başarılı olup olmadığı altın standart kümesinden seçilen çift için verilen cevabın doğruluğuna bakılarak değerlendirilmiştir.

İşveren-Çalışan ilişkilerinin zayıf olması, anonimlik özelliği ve beceri düzeylerindeki çeşitlilik, bekleneceği üzere KKÇ ile üretilen nihai ürün kalitesinde düşüklüğe neden olmaktadır. Bu yüzden, KKÇ uygulayıcıları bazı kalite güvence yöntemleri [17] kullanmak durumundadırlar.

Bu çalışmada altın standart, tekrarlama (redundancy) ve otomatik kontrol [17], [18] kalite güvence yöntemleri ile bir arada kullanılmıştır.

Birinci düzey kalite kontrol yöntemi olarak altın standart mikro-görevleri kullanılmıştır. KKÇ öncesi, 100 çiftten oluşan bir altın standart çiftler kümesi geliştirilmiştir. Bu kümedeki çiftlerin yarısı kolay bir şekilde aynılığı saptanabilecek şekilde olumlu örneklerden oluşturulmuştur. Kümedeki çiftlerin diğer yarısı ise, olumsuz örnek olarak, belirgin bir biçimde farklı kayıtlardan oluşturulmuştur. Çalışanlara atanacak görevlerin her birinde bulunan 4 kayıt çiftinin birisi altın standart kümesinden seçilmiştir. Sorulan 4 soru içindeki altın standart çiftine ait soruya verilen yanıtın doğruluğuna göre tüm yanıtlar kabul edilmiş veya reddedilmiştir.

Her bir görev 3 defa, 3 farklı çalışana atanmıştır. Daha sonra çoğunluk kararı tekniği kullanılarak, söz konusu yayın kayıtlarının doğruluğu ile ilgili nihai karar verilmiştir.

Son olarak, aynı benzerlik gruplarında bulunan kayıtların geçişkenliklerinin tutarlı olup olmadığı otomatik olarak kontrol edilmiştir. Örneğin, birbirine benzediği belirlenen üç yayın kaydından (A, B, C) üç oluşturulan üç benzerlik çifti ((A,B),(A,C),(A,D)) hakkında verilen karara göre A, B ile aynı ise, ve A, C ile aynı ise, B ile C benzerlik çiftinin de aynı olarak değerlendirilmiş olması beklenmektedir.

Bu otomatik geçişkenlik kontrolünün sonucunda az sayıda tespit edilen tutarsızlık, çoğunluk kararı tekniği ile ortadan kaldırılmıştır.

KKÇ fazı 17 günde tamamlanmıştır ve 186$'a mal olmuştur. 1.385 işçi, 9.308 mikro-görev icra etmiş ve çalışma esnasında toplam 27.924 karar toplanmıştır. Bunlardan 1.920'si altın standart görev başarısızlığı yüzünden kabul edilmemiştir. Bir günde tamamlanan ortalama görev sayısı 1.643 olarak tespit edilmiştir. Bir mikro-görevin tamamlanma süresi ortalama olarak 52 saniye olarak hesaplanmıştır.

6.224 kayıt çifti aynı, 3.084 kayıt çifti ise farklı olarak değerlendirilmiştir. Bu kararlar, her bir benzerlik grubundaki kayıtlar aynı olsun olmasın, otomatik olarak, nihai yargıyı oluşturmakta doğrudan kullanılmıştır.

Kitle kaynaklı çalışmanın doğruluk derecesi, rastgele örnekleme yoluyla uzman değerlendirmeleri kullanılarak saptanmıştır. Karşılaştırma işine temel oluşturması amacı ile uzman değerlendirmelerinden oluşan bir küme geliştirmek için, 1.500 rastgele seçilmiş mikro-görev uzmanlar tarafından el ile icra edilmiştir. Uzman değerlendirmeleri sonucunda uymazlık gösteren 96 kayıt tespit edilmiştir. Bu da seçilen örneklemin 6.4%'üne denk gelmektedir.

Sonuç ve Gelecekteki Çalışmalar

Bu çalışmada veri temizleme sorununu çözmeye yönelik bir yazılım prototipi geliştirilmiştir. Bu karma çözümde, harici kaynaklardan DOI sorgulanması, dizgi benzerlik hesaplama algoritmaları ve KKÇ kullanılmıştır.

Böyle bir işin, firma çalışanına atanması yerine, KKÇ ile yapılması önemli ölçüde zaman ve maliyet tasarrufu sağlamaktadır.

Ayrıca, aynı mikro-görevlerin tekrar tekrar yapılmasının çok sıkıcı ve psikolojik olarak zorlayıcı olduğu gözlemlenmiştir. Böylece mikro-görevler üzerinde çok sayıda işçinin çalışması ve psikolojik yükü paylaşıyor olmaları açısından KKÇ kullanımını avantajlı kılmaktadır.

KKÇ, tasarım açısından bakıldığında mükemmel olmaktan uzaktır. KKÇ ile vasat kalitede ürünler kolaylıkla elde edilebilir. Bu çalışmanın KKÇ aşamasında gözlenen hata oranı 6.4%'dür. Bu özellikteki bir iş için kabul edilebilen bir hata oranı olarak değerlendirilmiştir. Doğruluk oranını yükseltmek için daha iyi kalite güvencesi sağlayabilen tasarımlar gerekmektedir. KKÇ aşaması tamamlandığında elimizde kalan hatalı kayıtlar (53.822'de 596), yeni geliştirilen sisteme aktarılacaktır. Yeni sistem yazarların kendi yayın bilgilerini düzeltebilmelerine olanak sağlayacaktır.

KKÇ'nin etkinliğini yönetmek kadar kalite maliyetini yönetmek de önemlidir. Bu yüzden, kalite maliyetlerini kestirmek için bazı maliyet modellerinin kullanılması ve maliyet açısından en iyileştirilmiş kalite güvence yönteminin seçilmesi önerilmektedir.

Sonuç olarak, bazı problemlerin çözümü için karma yaklaşımların uygun olabileceği kararına varılmıştır. Bilgisayarların veri işleme gücü ile insanların algı ve kavrayış becerilerinin bir arada kullanılmasının her iki yöntemin güçlü yönlerinin, zayıf yönlerini dengelemesi yoluyla, daha iyi sonuçlar alınmasına katkıda bulunacağı öngörülmektedir. KKÇ'nin sadece yazılım geliştirme veya veri analizinde değil ayrıca üniversite araştırmalarının önemli bir bölümünde problem çözümü için kullanılabilecek değerli bir yöntem olarak görülmesi tarafımızca önerilmektedir.

Bu çalışmanın birincil katkısı veri temizleme ve aktarım sorununun çözülmesi için uygulanan çözümün çıktılarıdır. KKÇ kullanmanın avantajlı olacağı durumlarla karşılaşan araştırmacılar veya uygulamacılara yol göstereceği düşünülmektedir. Kaliteyle ilgili olarak sunulan gözlemler, uygulamacılara gerçekçi beklentiler oluşturmalarında fayda sağlayacaktır. Ayrıca, yazılım geliştirme süreçlerindeki problemlere çözüm olarak karma yaklaşımların faydalı olabileceğine dair bir örnek gösterilmiştir. Bu örneklerin sayısının artmasıyla uygulamacıların karma yöntemlerin kullanılması yönünde karar vermeleri beklenmektedir.

Bu çalışmanın ikincil katkısı ise özel olarak uyarlanmış JI' algoritması ve bu algoritmanın LD ve belirtilen eşik değerlerle bir arada kullanılmasının örneklenmesidir. Öyle ki, belirtilen yöntem benzer veri temizleme problemlerinde doğrudan kullanılabilecektir.

Bu alandaki araştırmalarımız birbiriyle ilişkili iki odak ekseninde sürdürülecektir. Bunlar, KKÇ ile büyük veri analizi süreçlerinin bütünleştirilmesi için etkin ve elverişli yöntemler geliştirmek ve KKÇ kalite güvence maliyetlerinin kestirilmesi ve kalite güvence yöntemlerinin seçilmesi konusunda ilkelerin oluşturulmasıdır.

Teşekkür. Bu araştırma projesi ODTÜ Bilimsel Araştırma Projeleri (BAP) kapsamında desteklenmiştir. Proje ODTÜ Bilgi İşlem Daire Başkanlığı'nda gerçekleştirilmiştir.

Kaynakça

1. A. Bernstein, M. Klein, and T. W. Malone, "Programming the Global Brain," *Commun. ACM*, vol. 55, no. 5, pp. 41–43, May 2012.
2. S. Amer-Yahia, A. Doan, J. Kleinberg, N. Koudas, and M. Franklin, "Crowds, Clouds, and Algorithms: Exploring the Human Side of 'Big Data' Applications," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 2010, pp. 1259–1260.
3. J. G. Davis, "From Crowdsourcing to Crowdservicing," *Internet Comput. IEEE*, vol. 15, no. 3, pp. 92–94, May 2011.
4. A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, "What's Inside the Cloud? An Architectural Map of the Cloud Landscape," in *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, 2009, pp. 23–31.

5. G. Lackermair, “Hybrid cloud architectures for the online commerce,” *Procedia Comput. Sci.*, vol. 3, no. 0, pp. 550–555, 2011.
6. M. Vukovic and J. Laredo, “PeopleCloud Service for Enterprise Crowdsourcing,” in *IEEE International Conference on Services Computing*, 2010, pp. 538–545.
7. M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz, “Direct Answers for Search Queries in the Long Tail,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 237–246.
8. L. von Ahn and L. Dabbish, “Labeling images with a computer game,” *Proc. 2004 Conf. Hum. factors Comput. Syst. - CHI ’04*, pp. 319–326, 2004.
9. E. Fast, D. Steffee, L. Wang, J. Brandt, and M. S. Bernstein, “Emergent, Crowd-scale Programming Practice in the IDE,” 2014.
10. G. Tziralis and I. Tatsiopoulos, “Prediction Markets: An Extended Literature Review,” *J. Predict. Mark.*, vol. 1, no. 1, pp. 75–91, 2007.
11. “Wikipedia.” [Online]. Available: www.wikipedia.org.
12. “Innocentive.” [Online]. Available: www.innocentive.com.
13. “Fold-it.” [Online]. Available: fold.it.
14. V. Levenshtein, “Binary codes capable of correcting spurious insertions and deletions of ones,” *Probl. Inf. Transm.*, vol. 1, pp. 8–17, 1965.
15. G. Navarro, “A Guided Tour to Approximate String Matching,” *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.
16. M. Levandowsky and D. Winter, “Distance between Sets,” *Nature*, vol. 234, no. 5323, pp. 34–35, Nov. 1971.
17. D. Iren and S. Bilgen, “Cost models of crowdsourcing quality assurance mechanisms,” 2013.
18. A. Quinn and B. Bederson, “Human computation: a survey and taxonomy of a growing field,” in *... Conference on Human Factors in Computing ...*, 2011.