VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Büyük Veri Kavramı ile İlgili Akademik Yayınların Metin Madenciliği Yöntemi ile Analizi

Hakan ÖZKÖSE[1*]
[1] *Bartın Üniversitesi , İ.İ.B. Fakültesi, Yönetim Bilişim Sistemleri Bölümü. 74100 Bartın/TÜRKİYE*

Makale Bilgisi

Başvuru:
25/12/2019
Kabul:
04/05/2020

Özet

Büyük veri son yıllarda büyük popülerlik kazanan çalışma alanlarından biridir. Walmart, Netflix, Shell, gibi bir çok firma büyük veri labarotuvarları kurarak ellerindeki verileri işlemekte ve elde ettikleri sonuçları planlamalarında kullanmaktadırlar. Beş kriteri bulunan bu çalışma alanında, en fazla bilinen kriter boyut kavramıdır. Daha sonra boyut kavramına, hız ve çeşitlilik kavramları eklenmiştir. En son olarak ise değer ve doğruluk kavramları dahil edilerek büyük veri kriterleri son halini almıştır. Bu çalışmada büyük veri üzerine yapılan akademik yayınlar incelenmiştir. Bu kapsamda, Science Citation Index Expanded, Social Science Citation Index ve Emerging Science Citation Index içerisinde yer alan dergiler göz önünde bulundurulmuştur. Bu dergiler içerisinde, başlık bilgisinde büyük veri kavramı geçen son 5 yıla ait 3868 adet ingilizce makalelerin tüm bilgileri Web of Science Core Collection veritabanından elde edilmiştir. İlk olarak en fazla yayın yapan ülkeler ile aralarındaki ilişkiler, daha sonra, ilgili alandaki yazarlar ile ilişkileri ve son olarak da makalelerin içerisindeki anahtar kelimeye göre öne çıkan terimler incelenmiş ve analiz edilmiştir. Bu işlemler sırasında iki programdan yararlanılmıştır. Verilerin birleştirilmesinde, indirgenmesinde ve temizlenmesinde BibExcel, verilerin ilişki haritalarının oluşturulmasında ve görselleştirilmesinde ise VosViewer programlarından yararlanılmıştır.

Anahtar Kelimeler: *Büyük Veri, Metin Madenciliği, Görselleştirme, Birliktelik Kuralları*

Analyzing Academic Papers Related to Big Data Concept by Text Mining

Abstract

The concept of big data is one of the areas that gained great popularity in recent years. Many companies, such as Walmart, Netflix, Shell, have established big data labs for processing their data and using the results they have obtained in their planning. Big data has five criteria and the volume is the most known among them Then, the concept of speed and diversity had been added over this concept. Finally, the concept of value and accuracy was added and the concept was finalized. In this study, academic publications on big data were examined. In this context, the journals which are inside of Science Citation Index Expanded, Social Science Citation Index and the Emerging Science Citation Index are taken consideration. In this journals, all the information of 3868 articles in English which were published in last 5 years and passing the big data concept in title information were taken from the Web of Science Core Collection database. Firstly, the relations between countries which publish most articles and then the relations between authors who publish most articles and lastly the most prominent terms according to the keywords in the articles were examined and analyzed. Two programs were

* İletişim e-posta: hakan_ozkose@hotmail.com

used during these operations. Merging, reducing and cleaning of data was done by the help of Bibexcel. Creating relationship maps and visualizations of maps were done by VosViewer.

Keywords: *Big Data, Text Mining, Visualization, Association Rules*

1 Giriş

Büyük veri kavramı günümüzde popüleritesini arttıran kavramlardan bir tanesidir ve son zamanlarda işletmelerin veriye daha fazla önem vermesine neden olmaktadır. Veri analiz yöntemlerinden bir tanesi olan büyük veri, son zamanlarda şirketlerin ilgisini çekmiş ve bu alan üzerine büyük yatırımlar yapmışlardır. Özellikle bu alanın öncüsü olan Google, yapmış olduğu büyük yatırımlar ile dikkat çekmektedir. Google'ı, Walmart, Netflix, Shell, Facebook, Linkedin, BBC, Airbnb, Amazon gibi dünya çapında bilinen büyük firmalar izlemiş ve rekabet ortamında ön plana çıkmayı başarmışlardır. Bu şirketler veriyi analiz etme şekilleri ile giderlerini büyük oranda düşürmüşler ve karlarını arttırarak alanlarında öncü olmayı başarmışlardır.

Büyük veri kavramının şekillenmesi zaman içerisinde gerçekleşmiştir. İlk başlarda sadece boyut yani volume kriteri ile bilinirken, zaman içerisinde değişime uğrayarak günümüzde 5 farklı kriterin birleşmesiyle oluşmuştur. Aslında bu beş kriter büyük verinin yol haritası olarak bilinmektedir ve 5V kısaltması ile anılmaktadır. 5V kavramı bizlere, volume (hacim), variety (çeşitlilik), velocity (hız), value (değer) ve veracity (doğruluk) kavramlarını hatırlatmaktadır. Bu sayede aslında büyük verinin ne olduğu gün ışığına tutulmuş olur. Büyük veri kavramında unutulmaması gerekenlerden bir tanesi de işleme hızının yüksek olması için yüksek teknoloji gerekliliğidir.

Veri analiz yöntemlerinden biri olan büyük verinin incelemesi için bu çalışmada metin madenciliği ile sınıflama ve kümeleme algoritmalarının yapısından yararlanılmıştır. Metin madenciliği kısaca metin verilerinin içerisinden anlamlı sonuçlar çıkartılmaya çalışılmasıdır. Veri madenciliğinde sıklıkla kullanılan yöntemlerden bir tanesi olarak da bilinir.

Bu çalışmada hem metin madenciliği hem de verinin görselleştirilmesinde Vosviewer programından yararlanılmıştır. İlk başta Web of Science Core Colection veritabanından elde edilen verilerin birleştirilmesi ise BibExcel programından yararlanılmıştır. Ayrıca, verinin uygun formata çevirilmesinde de BibExcel'den yararlanılmıştır.

Bu çalışma kapsamında büyük veri üzerine yapılan çalışmalar Web of Science Core Collection veri tabanından elde edilen 3868 adet makalenin verilerinden yararlanılarak yapılmıştır. Hem ülkelerin hem de yazarların alana yapmış oldukları katkı düzeyleri görselleştirme teknikleri kullanılarak gösterilmiş ve yorumlanmıştır.

Bunun dışında yazarların yayınlarında belirlemiş oldukları anahtar kelimelere göre bir analiz yürütülmüş, alanda hangi konuların popüler olduğu ve aralarındaki ilişki düzeyleri ilişki haritaları ile gösterilmiştir. Yapılan analiz sonuçlarına bulgular aşamasında yer verilmiştir.

Çalışma şu şekilde oluşturulmuştur; ilk olarak büyük veri kavramına yer verilmiştir. Bu başlık altında, büyük verinin bileşenlerine ve bu çalışmadaki uygulama alanlarına değinilmiştir. Daha sonra ise, metin madenciliği kavramına değinilip uygulama alanlarından bahsedilmiştir ve metin madenciliğinde kullanılan programların bilgilerine yer verilmiştir. Sonraki başlıkta ise araştırma ile ilgili bilgilerden bahsedilmiştir. Araştırmanın amacı, modeli ve bulguları bu bölüm içerisinde yer almıştır. Çalışmanın sonunda ise sonuçlara ve önerilere değinilmiştir.

2 Literatür Çalışması

Büyük veri, ilk başlarda tek bir makine tarafından işlenemeyen veriler olarak bilinmekteyken, artık veri analitiği veya görselleştirme ile ilgili her şeyi kapsayan bir terim olarak kullanılmaktadır [1].

Verinin hacminin, çeşitliliğinin ve hızının sürekli olarak artması ile büyük veri kavramı ortaya çıkmıştır. "Kaydedilen, yeniden düzenlenebilen ve analiz edilebilen bilgi birimi" [2] şeklinde tanımlanan veri, özgün yapısındaki bu değişimler neticesinde gelişerek büyük veriye dönüşmüştür [1].

"Büyük Veri" terimi, analizi yapılan verinin büyüklüğünü göstermek amacıyla ilk olarak 1900'lü yılların ortalarına doğru ortaya çıkmıştır ve bir çok firma tarafından kullanılmaya başlanmıştır. Literatüre bakıldığı zaman büyük veri kavramının tam olarak anlaşılmadığı görülmektedir. Bu alanda çalışma yapanlar kendisi için önemli olarak gördüğü verinin karakteristik yapısını ön plana çıkararak tanımlamalar yapmışlardır. Bu sebeple tanımlama yapanlar farklı farklı ve birbiri ile çelişen tanımlamalar yapmışlardır. Bu tanımlamalardan bir kaçı üzerinde durulması, büyük veri kavramının açıklanması açısından faydalı olacaktır [1].

Bir tanımda; "büyük veri yüksek hacimde, yüksek çeşitlilikte ve hızla gelen verilerin toplanması, saklanması, temizlenmesi, görselleştirilmesi, analiz edilmesi ve anlamlandırılması eylemi" şeklinde tanımlanmıştır [3]. Bir başka tanımda ise; "bilgi, karar verme ve süreç otomasyonunu artıran, maliyet etkin, yenilikçi bilgi işleme biçimleri talep eden, yüksek hacimli, yüksek hızlı ve/veya çok çeşitli bilgi varlıkları" olarak açıklanmıştır [4]. Tüm bu tanımlarla beraber olarak "gerek insan gerekse makineler tarafından sayısal olarak kodlanmış her türden kurumsal veri ile internet ve sosyal medya paylaşımları aracılığıyla ortaya çıkan kişisel verilerin anlamlı ve işlenebilir biçime dönüştürülmesi durumu" olarak da tanımlanmıştır [5]. Yukarıda bahsedilen tanımlara göre büyük veri, yüksek hızlı yakalama, keşfetme ve/veya analiz sağlayarak çok değişkenli büyük veri setlerinden ekonomik bir şekilde değer ayıklamak için tasarlanmış yeni nesil teknolojiler ve mimarilerdir.

Büyük veri, geleneksel veri tabanı yöntemlerinin kullanılması yoluyla işlenmesi imkansız olan, farklı büyüklüklerdeki ayrı türden veriyi açıklayan yeni bir kavramdır ve farklı dijital içeriklerden oluşmuştur [6].

I. *Yapısal veri:* Yapısal veri modellenmesi, girdi olarak sokulması, muhafaza edilmesi, sorgulanabilmesi ve görselleştirilmesi mümkün ve basit olan her türlü veri çeşidini tanımlamaktadır. Genel olarak önceden tanımlaması yapılmış alanlarda belirli çeşit ve boyutlarda sunulmaktadır. İlişkisel veri tabanlarında ve tablolarda yönetilebilirler. Katı bir yapıya sahiptirler, süreçler yüksek performanslı yetenekler

ve paralel teknikler gerektirmezler. Bu sebeple diğer veri türlerine nazaran faydalı bilgilerin elde edilmesi daha kolaydır [7].

II. *Yarı yapısal veri:* Yarı yapısal veya kendi kendini açıklayan (self-describing) veri, yapısal bir veri türünü yansıtırlar fakat özünde sadece katı bir modeli barındırmazlar. Başka bir ifade ile yapısallığın tanımlandığı modellerle birlikte belirli ögeleri ve verideki farklı alanların aşama sırasına göre gösterilmesinin tanımlanması amacıyla kullanılmakta olan işaretler ve etiketler gibi meta modellerini de bulundurmaktadırlar. En çok bilinen örnekleri arasında XML (Extensible Markup Language) ve JSON (JavaScript Object Notation) programlama dilleri yer almaktadır [7].

III. *Yapısal olmayan veri:* Yapısal olmayan veri, belirli bir formatı bulunmadan sunulan ve depolanan kayıt türleridir. Genel itibari ile kitap, dergi, makale, e-posta gibi serbest boyutlardaki metinler ve ses video benzeri medya dosyalarından oluşurlar. Yapısal olmayan verilerde verinin katı bir şekilde sunulması zor olasından dolayı NoSQL (Not only SQL) gibi yeni mekanizmaları ortaya çıkarmıştır [7].

2001 yılında yayınlanan Gartner araştırma raporunda (Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety") şirketin analisti olarak çalışan Doug Laney, fırsatları ve değişimleri de göz önünde bulundurarak verinin hızı, çeşitliliği ve büyüklüğü üç boyutlu bir şekilde yani 3V olarak tanımlamıştır ve hâla bir çok endüstri firması bu tanımı kullanmaktadır. Daha sonra 2012 yılında Gartner tarafından yapılan tanım büyük veri, çok büyük hacim, çok büyük hız ve çok fazla çeşitlilik olarak güncellenmiştir. Başka şirketler tarafından bu tanıma daha sonra bir V daha eklenmiştir. Bu kriterin amacı ise verinin tutarlılığının kurum içi önemli kararlar alınmasına etkisinin olup olmadığının gösterilmesidir [8]. Bazı kaynaklarda 3V'ye ek olarak doğruluk ve değer de eklenerek 5V'den bahsedilmektedir [9]. 5V kavramına aşağıda sırası ile değinilmiştir.

I. *Çeşitlilik:* Herhangi bir veri kümesindeki yapısal heterojenliği belirtir ve bu heterojen yapıyı %95 oranla yapısal olmayan veriler oluşturur [10]. Örnek vermek gerekirse bir çağrı merkezindeki konuşma kayıtları; müşteri adı, konuşma zamanı, konuşma süresi, şikâyet konusu yapısal olmayan veri içerisinde yer almaktadır [11]. Çağrı merkezi örneğine baktığımız zaman konuşma verisi gruplandırılabilir fakat şikâyet konusu grubundaki veri yapısal olmayan veri içerdiği için ilişkisel veri tabanlarına yerleştirilememektedir [7].

II. *Hız:* Verinin daima hareket halinde olduğu varsayılmaktadır. Bu sebeple, veri akış analizi, veri bilimciler için oldukça önemli bir hal almıştır. Verinin üretilme hızı oldukça yüksektir ve her geçen gün katlanarak arttığı da bilinmektedir. Hız sadece büyük veri için değil diğer iş süreçleri içinde oldukça önem arz etmektedir. Örneğin; sadece Formula 1 yarış arabaları üzerinde yer alan 150 sensör sayesinde 20 gigabayt veri üretebilmek mümkündür. Başka bir örnek olarak ise CERN'de (Conseil Européen pour la Recherche Nucléaire - Avrupa Nükleer Araştırma Kuruluşu) gerçekleştirilen "Büyük Hadron Çarpıştırıcısı" deneyinde sensörler sayesinde 1 petabayt veri üretilmiştir [7]. Sadece bu iki örnek ile bile verinin üretim hızının ne kadar yüksek olduğu görülebilmektedir.

III. *Hacim:* Boyut, büyük veri problemlerinin en başında gelmektedir. Çünkü veri depolama ve veri erişimi için yenilikçi programlara gerek duyulmaktadır. Büyük veri, artık geleneksel veri analiz teknikleri ile işlenemeyecek ve klasik veri tabanlarına sığmayacak boyutlara ulaşmıştır. Hayatımızda vazgeçilmez hale gelen akıllı telefonlar ve internet temelli uzaktan kontrollü cihazlar gibi birçok cihaz, sersörleri ile uygulamalara ürettikleri veriyi aktarabilmektedirler. Bu sebeple üretilen, depolanan ve iletilen veri miktarında yüksek bir artış olmaktadır. Araştırma kuruluşlarından IDC (International Data Corporation) tarafından hazırlanan "Digital Universe Study" isimli çalışmada, 2020'de yıllık veri hacminin 35 zettabayt seviyesine ulaşacağı belirtilmiştir [12]. Bilgi teknolojileri alanında çok uluslu bir şirket olarak faaliyet gösteren CSC (Computer Sciences Corporation) tarafından yayınlanan bir rapora göre de 2020 yılında elde edilecek veri hacminin günümüze göre % 4.300 oranında artış göstereceği öngörülmektedir. Günümüzde orta ölçekli organizasyonlarda bile 1 terabayt hacminde veri çok kısa süre içerisinde üretilebilmekte ve bu veri birçok kaynak tarafından yüksek çeşitlilikte oluşturulabilmektedir. IBM'e (International Business Machines) göre, 2014 yılından bu tarafa dünya üzerindeki verinin yaklaşık %90'ı son iki senede üretilmiş olup; her gün 2,5 eksabayt büyüklüğünde veri üretimi gerçekleştirilmiştir. 2003 yılına kadar insanlık tarihi boyunca üretilen veri 5 eksabayt iken aynı miktardaki veri iki günde üretilmiştir [7].

IV. *Doğruluk:* Büyük verinin ne kadar güvenilir ve doğru olduğunu göstermektedir. Veri, alınan iş kararlarında kullanılabilecek derecede güvenilir olmalıdır. Büyük verinin çeşitliliğinin yüksek olması, veri kalitesinin güvenilir olma durumunu zorlaştırmaktadır [13]. Veri kalitesi verinin güvenilirliği ile değerlendirilebilmektedir. Çünkü yüksek kaliteli veri sadece güvenilir modellerle üretilebilmektedir. Başka bir deyişle veride bulunan aykırı ve eksik değerler gibi anomaliler tespit edilebilmektedir [9]. Bu anomaliler, bir takım veri kaynaklarına has güvensizlik durumunun mevcut olması ile ilişkilendirilmektedir. Bu sebeple kesinlik belirtmeyen ve belirsizlik içeren veriyle baş edebilme ihtiyacı, belirsizlik içeren verinin yönetimi, veri madenciliği için tasarlanan araçların ve analiz metotlarının kullanılması, incelenmesi gereken başka bir yönünü de yansıtmaktadır [10]. Bu noktada elde bulunan verinin doğruluğu ve geçerliliği önem taşımaktadır. Doğru ve geçerli olmayan büyük veriler hem yanlış yorumlamalara yol açabilecek, hem de analiz için temel teşkil etmeyecektir [6].

V. Değer: Büyük veriyi oluşturan bileşenlerin en önemlisi kuşkusuz ortaya bir değer konmasıdır. Verinin işlenmesi, depolanması ve analiz işleminden sonra karar alma esnasında o kuruluşa da ayrı bir değer katması gerekmektedir. Örneğin, bir havayolu şirketi müşterilerinin alışkanlıklarını biliyorsa yapılacak yatırımların çok daha faydalı olmasına sebep olacak ve ekonomik açıdan da büyük katkı sağlayacaktır [14, 15].

Büyük veri, araştırmacıların kişisel davranışları ve topluluk eğilimlerinde, aradıkları sorulara cevap bulmaları açısından, araştırmacılara büyük kolaylıklar sağlamaktadır. Bununla beraber ekonomik ve ticari faaliyetler, kamu yönetiminde iletişim, ulusal güvenlik, bankacılık ve bilimsel araştırmalar gibi çok fazla sahada büyük veriden yararlanılmaktadır.

Büyük veri uygulamalarının önemli hedeflerinden bazıları; tüketici deneyiminin iyileştirilmesi, çok daha düzgün pazarlama stratejilerinin belirlenmesi, maliyetlerin düşürülmesi ve mevcut süreç etkinliğinin arttırılmasıdır. Buna ek olarak günümüzde veri ihlallerinden dolayı yaşanan olaylarda güvenliğin sağlanması da büyük veri kullanım amaçları arasında yer almıştır.

Bu çalışma kapsamında büyük veri terimlerinin analizi için metin madenciliğinden yararlanılmıştır.

Feldman ve Sanger' e göre metin madenciliği, kullanıcının analiz araçları kullanılarak bir doküman yığını ile etkileşime girdiği bir süreç olarak tanımlanabilir [16]. Sumathi [17] ise metin madenciliğini, dokümanlar için geleneksel arama tamamlandıktan sonra dokümanlar arasındaki karmaşık ilişkileri araştıran bir işlem olarak tanımlamıştır [18]. Metin madenciliği, metin kümeleri içerisinden yeni ve anlamlı bilgiler çıkarmayı amaçlayan ve sürekli kendini geliştirmekte olan bir alandır. Metin madenciliği, kelime ya da bilgilerin gerçek bağlantısının fonksiyonu olan metin ile ilgilenir [19].

Eskiden dokümanlar içerisinde bulunan bilgiye ulaşmak için elle indeksleme yapılmaktaydı. Günümüzde her şeyin dijitalleştiği ve elektronikleştiğinden yola çıkarsak ve sadece internet üzerinde iki milyara yakın web sayfası olduğu düşünülürse buna elle ulaşmanın imkânsız olduğu rahatlıkla söylenebilir. Bu problemi aşmak için otomatik bilgiye ulaşma yöntemleri geliştirilmiştir. Bu yöntemler metin madenciliği adı altında incelenmektedir [20].

Metin madenciliğinde kullanılan hem ücretli hem de ücretsiz yazılımlar aşağıda belirtilmiştir;

- Vosviewer
- Sas
- Text Mining
- SPSS Text Mining and Text Analysis For Surveys
- STATISTICA Text Miner
- GATE-Natural Language
- RapidMiner
- R Programming
- Perl
- ODM-Oracle Data Mining.

3 Materyal ve Metot

3.1 Araştırmanın Amacı ve Önemi

İşletmelerin gelecekte rekabette ayakta kalabilmeleri için teknolojik gelişmelere açık olmaları gerekmekte ve buna ek olarak ise veri biliminden anlaması gerekmektedir. Büyük veri kavramına, büyük firmalar tarafından son 5 yıl içerisinde büyük yatırımlar yapılmıştır. Yapılan bu yatırımlar sayesinde şirketler verilerini daha iyi anlamakta ve geleceklerine daha iyi yön verebilmektedir. Bu çalışma ile büyük veri alanında yapılmış olan çalışmalardaki veriler kullanılarak alanın genel hatları çizilmiştir. Bu sayede büyük verinin çalışma alanı ortaya konulmuş ve en fazla kullanılan terimler belirlenmiştir. Bu sayede çalışmaların nereye doğru yöneldiği anlaşılabilmektedir. Bunun dışında alanda hangi ülkelerde büyük veri çalışmalarının daha fazla yapıldığı gösterilmiştir. Ayrıca, büyük veri üzerine en fazla çalışma yapan bilim insanlarına da bu çalışmada yer verilmiş ve birbirleri ile ilişkileri gösterilmiştir. Bu sayede hem ülkeler hem de yazarlar arasındaki ilişkiler keşfedilebilir.

3.2 Araştırmanın Sınırlılıkları

Bu çalışmadaki verilere Web of Science'ın Core Collection veritabanından ulaşılmıştır. Sadece son 5 yıl içerisindeki (2014-2018) yılları arasındaki SCI, SCI-Exp ve SSCI indexlerinde yer alan ingilizce makalelerden yararlanılmıştır.

3.3 Araştırmanın Yöntemi

Bu çalışma içerisinde 4 farklı başlık incelenmiştir. Bunlar;

I. Ülkerin yapmış oldukları çalışma sayılarının ve ilişkilerinin belirlenmesi,

II. Yazarların yapmış oldukları çalışmalara göre birbileri İle ilişkilerinin belirlenmesi,

III. Atıf yapılan yazarların co-citation analizi,

IV. Yazarların vermiş olduğu anahtar kelimelere göre co-ocurence analizidir.

Yukarıdaki başlıkların incelenebilmesi için ilk olarak Web of Science Core Collection veritabanından elde edilen verilerin birleştirilmesi gerekmektedir. 8 farklı dosya içerisinde yer alan 3868 adet ingilizce malale birleştirme işleminden sonra sadece tek bir dosya içerisinde yer almıştır. Birleştirme işleminden sonra verilerin parçalanması işlemi yapılmıştır. Parçalama işlemi sonunda her bir soru maddesi için bir dosya oluşturulmuştur. Verilerin birleştirilmesi ve sonra ilgili sorular için parçalara bölünmesi için BibExcel programından yararlanılmıştır. Çalışmanın ikinci kısmında ise verinin görselleştirilmesi ve ilişki haritalarının çıkarılması vardır. Bu işlemlerde ise VosViewer programı kullanılmıştır. En son olarak ise elde edilen haritalar yorumlanmıştır.

4 Bulgular ve Tartışma

4.1 Ülkerin Yapmış Oldukları Akademik Çalışma Sayılarının ve İlişkilerinin Belirlenmesi

Bu analizde yazarların menşeileri yani ülkeleri dikkate alınmıştır. Ülke bazlı olarak yaplımış olan yayınların sayıları ve aldıkları atıf sayıları analizde kullanılmıştır.

Bu analiz için seçim kriterleri bir ülkenin en az 10 çalışmaya katkı vermiş olması ve o ülkenin bu çalışmalar içerisinden en az 10 atıf almış olması gerektiğidir. Bu kriterleri sağlayan ülke sayısı bu alanda yayın yapmış toplam 105 ülkeden sadece 50'sidir. Bu kriterin konmasının en önemli nedeni alana yön veren ülkelerin belirlenmesi ve aralarındaki ilişkilerin ortaya konulmak istenmesidir. Bu bölümde ilk olarak, alana en fazla katkı veren 20 ülkenin yapmış oldukları çalışma sayıları Tablo 1'de verilmiştir.

Tablo 1. Ülkelerin akademik çalışma ve atıf sayıları

No	Ülkeler	Yayın Sayısı	Atıf	Atıf/Yayın
1	Usa	1238	9626	7,78
2	P.R. China	1133	8088	7,14
3	England	328	2951	9,00
4	Australia	233	2193	9,41
5	Canada	184	1774	9,64
6	France	120	1293	10,78
7	Spain	156	1204	7,72

8	Italy	162	984	6,07
9	South Korea	210	977	4,65
10	Germany	133	969	7,29
11	India	177	820	4,63
12	Malaysia	47	756	16,09
13	Sweden	55	625	11,36
14	Taiwan	92	573	6,23
15	Switzerland	65	559	8,60
16	Saudi Arabia	79	545	6,90
17	Netherlands	88	503	5,72
18	Japan	86	463	5,38
19	Singapore	47	462	9,83
20	Brazil	40	433	10,83

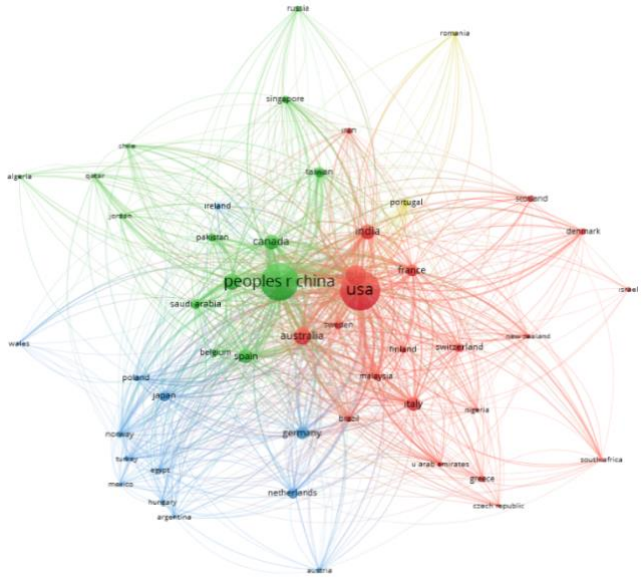
Bu tabloda ülkelerin isim, yayın sayıları, aldıkları atıflar ve yayın başına düşen atıf sayıları yer almaktadır.

Tablo 1 incelendiğinde, 1238 yayında aldığı 9626 atıf sayısı ile Amerika ilk sırada yer almaktadır. Amerikayı sırasıyla, 8088 atıf ile Çin, 2951 atıf ile İngiltere, 2193 atıf ile Avustralya ve 1774 atıf ile Kanada izlemektedir.

Tabloda yer almasa da Atıf/Yayın Sayısı oranının en yüksek olan ülke yayın başına 16.5 atıf ile Çek Cumhuriyeti'dir. Çek Cumhuriyeti menşeili 12 yayına 198 atıf yapılmıştır. Çek Cumhuriyetini, yayın başına 16,09 atıf sayısı ile Malezya izlemektedir. Amerika en çok atıf alan ülke olmasına rağmen yayın başına düşen atıf sayısı 7,78'dir. Bu bakımdan bakıldığı zaman Amerikanın yayın sayının çok olmasına rağmen aldığı atıf sayısının diğer ülkelere yakın olduğu görülmektedir.

Türkiye bu alanda atıf sayısı bakımından 31. sırada kendisine yer bulmuştur. Türkiye yapmış olduğu 25 çalışmayla 185 atıf almıştır. Yayın başına düşen atıf sayısı ise 7,40 olarak belirlenmiştir.

Aşağıdaki şekilde ülkelerin birliktelik ilişkileri görülmektedir. Şekil incelendiğinde Amerika (Kırmızı Küme) ve Çin'in (Yeşil Küme) merkezde yer aldıkları görülmektedir.



Şekil 1. Ülkelerin atıflarına göre ilişki haritası

Şekil 1. Şekil üzerindeki renklerin her biri bir kümeyi ifade etmektedir. Küme içerisindeki benzerliklerin yüksek kümeler arasındaki mesafelerin ise uzak olması istenmektedir. Aşağıdaki şekil incelendiğinde ise 4 farklı kümenin olduğu görülmektedir. Bu kümeler sarı, kırmızı, yeşil ve mavi renkler ile gösterilmiştir.

Şekil incelendiğinde, kırmızı küme içerinde yer alan Amerikanın, Fransa, Hindistan, Avustralya ile aynı kümede bulunduğu ve ilişkisinin diğerlerine göre daha yüksek olduğu görülmektedir. Yeşil küme içerisinde en fazla atıfı olan Çin ise Kanada, İspanya ve Belçika ile aynı küme içerisnde yer almaktadır. Türkiye ise mavi kümede yer almış ve Mısır, Norveç ve Meksikaya daha yakın bir konumda bulunmaktadır. Unutulmamalı ki bu birliktelik analizleri sonucunda yazarların ülkeleri göz önünde bulundurularak birlikte yapmış oldukları yayınlar göz önünde bulundurularak oluşturulmuştur.

4.2 Yazarların Yapmış Oldukları Çalışmalara Göre Birbileri İle İlişkilerinin Belirlenmesi

Bu analizde yazarların yapmış oldukları yayın sayısı ve aldıkları atıf sayılarına dikkat edilmiştir. Bu analiz içine seçim kriterleri; bir yazarın en az 10 çalışmaya katkı vermiş olması ve en az 10 atıf almış olması gerektiğidir. Bu kriterleri sağlayan yazar sayısı toplam 11529 yazardan sadece 22 tanesidir. Bu kriterin konmasının en önemli nedeni alana yön veren bilim insanlarının belirlenmesi ve aralarındaki ilişkilerin ortaya konulmak istenmesidir.

Bu bölümde ilk olarak alana en fazla katkı veren akademisyenlerin yapmış oldukları çalışma sayıları Tablo 5.2'de verilmiştir. Bu tabloda yazarların isimleri, yayın sayıları, aldıkları atıflar ve yayın başına düşen atıf sayıları yer almaktadır. Bu tablo atıf sayılarına göre büyükten küçüğe doğru düzenlenmiştir.

Tablo 2. Yazarların akademik çalışma ve atıf sayıları

No	Yazar	Yayın Sayısı	Atıf	Atıf/Yayın
1	Dubey, Rameshwar	12	331	27,58
2	Ranjan, Rajiv	16	328	20,50
3	Herrera, Francisco	19	321	16,89
4	Wang, Lizhe	14	307	21,92
5	Ahmad, Awais	19	270	14,21
6	Paul, Anand	20	268	13,40
7	Gunasekaran, Angappa	12	265	22,08
8	Chang, Victor	19	257	13,52
9	Yang, Laurence T.	20	246	12,30
10	Zhang, Yan	10	237	23,70
11	Wang, Kun	13	230	17,69
12	Rathore, M. Mazhar	14	214	15,28
13	Vasilakos, A. V.	13	201	15,46
14	Chen, Jinjun	10	184	18,40
15	Chen, Zhikui	10	184	18,40
16	Guo, Song	12	151	12,58
17	Zomaya, Albert Y.	15	147	9,80
18	Zhang, Yin	10	145	14,50
19	Li, Peng	11	137	12,45
20	Qiu, Meikang	10	122	12,20
21	Hossain, M. Shamim	12	112	9,33
22	Li, Keqin	12	95	7,91

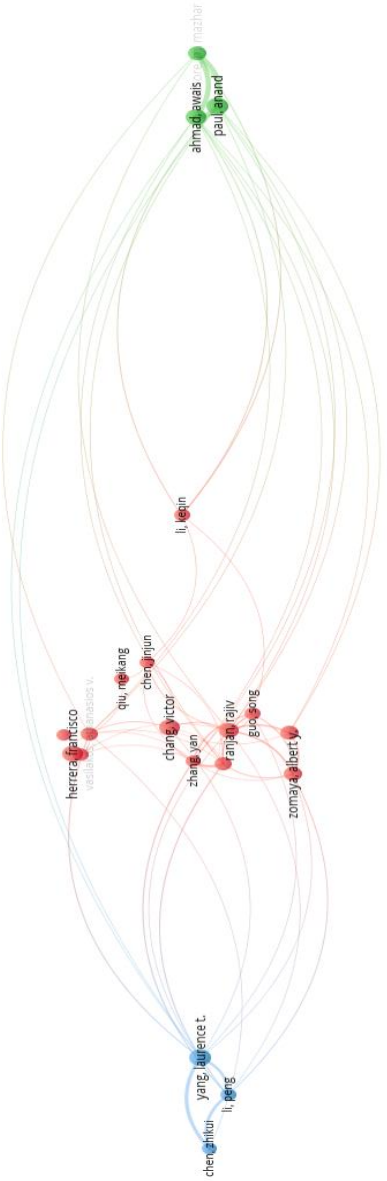
Tablo 2 incelendiğinde, 12 yayında aldığı 331 atıf sayısı ile Dubey, Rameshwar ilk sırada yer almaktadır. Dubey, Rameshwar sırasıyla, 328 atıf ile Ranjan, Rajiv, 321 atıf ile Herrera, Francisco, 307 atıf ile Wang, Lizhe ve 270 atıf ile Ahmad, Awais izlemektedir.

Atıf/Yayın Sayısı en yüksek olan yazarın Dubey, Rameshwar olduğu görülmektedir. Dubey, R. yayın başına 27,58 atıf alarak bu alanda ilk sırada yerini almıştır. Dubey, R.'yi, yayın başına 23.70 atıf sayısı ile Zhang, Yan izlemektedir.

Yayın sayısı bakımından ise, en çok makalesi bulunan yazarlar Paul, Anand (20 Yayın ve 268 Atıf) ve Yang, Laurence T.'dir (20 Yayın ve 246 Atıf). Türkiyeden herhangi bir akademisyen bu listeye girmeyi başaramamıştır.

Yazarların birbirleri ile ilişkilerinin daha net anlaşılabilmesi için görselleştirmeden yararlanılmıştır. Bu bağlamda aşağıdaki şekil elde edilmiştir. Böylece hangi yazarın hangi yazar ile ilişkisinin daha çok olduğu daha net görülebilmektedir.

Şekil 2 incelendiğinde üç tane küme olduğu görülmektedir. Bu kümeler mavi, yeşil ve kırmızı ile renklendirilmiştir. Mavi ve yeşil kümede üçer yazar bulunurken kırmızı kümede 16 yazar bulunmaktadır. Kırmızı küme diğer kümelere oranla daha güçlü bir küme konumundadır.



Şekil 2. Yazarlara göre ilişki haritası

Yeşil kümede, Ahmad, A., Rathore, M. ve Paul, Anand birbirleri ile güçlü ilişki içerisindeyken mavi kümede, Chen, Z., Yang L. ve Li, P. kendi aralarında güçlü bir bağ kurmuşlardır. Kırmızı kümede ise diğer yazarların ilişkileri diğer küme elemanları ile olan ilişkisine göre daha güçlüdür. Tüm birliktelikler ve birliktelik değerleri incelendiğinde Rameshwar DUBEY (Toplam bağlantı skoru = 130) alanın öncüsü olarak ortaya çıkmaktadır. İlişki ağı skoru incelendiğinde ise en güçlü ilişkisi olan yazarın M. Mazhar Rathore (Toplam bağlantı skoru = 133) olduğu görülmüştür

4.3 Atıf Yapılan Yazarlara Göre Co-Citation Analizi

Bu analiz kapsamında atıf yapılan yazarların birbirleri ile olan birliktelik ilişkileri ölçülmeye çalışılmıştır. Kesme değeri 20 atıf olarak belirlenmiştir. Kesme değerinin belirlenmesi ile birlikte çalışmada 79900 yazardan 583 tanesi görselleştime için kullanılmış ve aralarındaki ilişkiler haritalanmıştır.

Tablo 3 incelendiğinde, Web of Science Core Collection veritabanından elde ettiğimiz makalelerde en fazla atıf yapılan yazarın Dean, J. olduğu görülmektedir. 3868 makaleden 439 tanesinde Dean, J.'nin yapmış olduğu çalışmalara yer verilmiştir. Dean, J.'yi 292 atıf ile Manyika, J., 291 atıf ile Boyd, D., 287 atıf ile 275 atıf ile Davenport, Th. izlemektedir.

Tablo 3. En fazla atıf alan yazarlar

No	Yazar	Aldığı Atıf	Toplam Bağlantı Kuvveti
1	Dean, J.	439	4715
2	Manyika, J.	292	4055
3	Boyd, D.	291	3727
4	Chen, M.	287	3459
5	Davenport, Th.	275	4620
6	Zaharia, M.	247	2494
7	Mayer-Schonberger, V.	244	2945
8	Chen, Hc.	234	3785
9	Mcafee, A.	233	3515
10	Laney, D.	224	3194
11	Kitchin, R.	222	2691
12	Lazer, D.	170	1835
13	Zhang, Y.	161	1518
14	Wang, Y.	159	1756
15	Wu, Xd.	155	2209
16	Wamba, Sf.	150	3125
17	Chen, Clp.	149	2275
18	Gandomi, A.	146	2483
19	White, T.	127	1513
20	Breiman, L.	119	1046

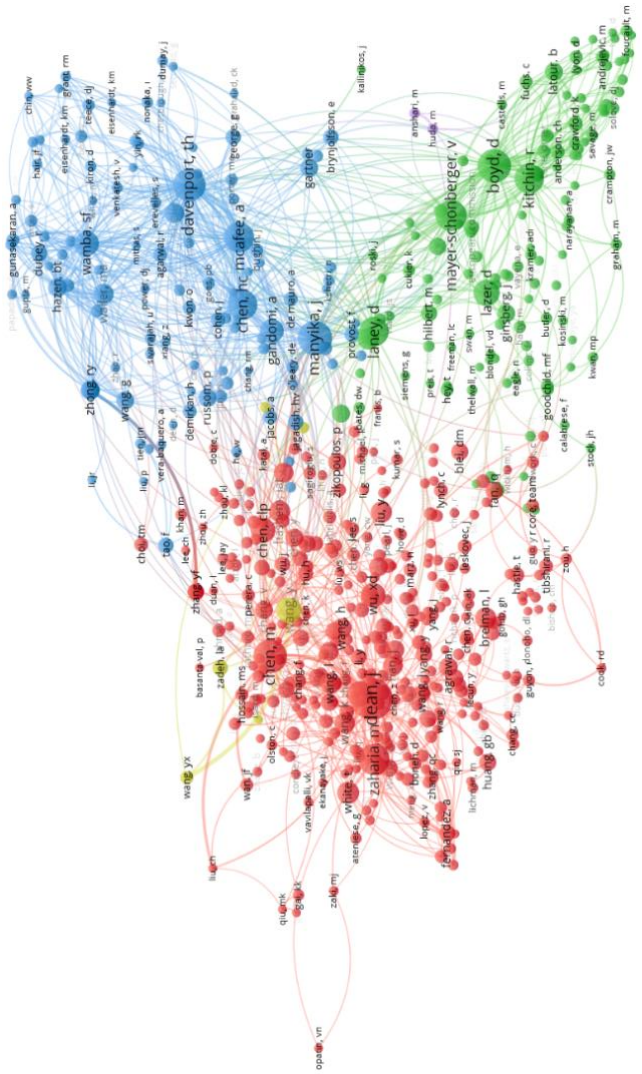
Web of Science Core Collection veritabanından elde ettiğimiz makalelerde en fazla atıf yapılan yazarın Dean, J. olduğu görülmektedir. 3868 makaleden 439 tanesinde Dean, J.'nin yapmış olduğu çalışmalara yer verilmiştir. Dean, J.'yi 292 atıf ile Manyika, J., 291 atıf ile Boyd, D., 287 atıf ile 275 atıf ile Davenport, Th. izlemektedir.

Toplam bağlantı kuvveti en yüksek olan yazar yine 4715 bağlantı kuvveti ile Dean, J.'dir. Dean, J'yi 4620 bağlantı kuvveti ile en fazla atıf alanlar listesinde 5. sırada olan Davenport, Th. izlemektedir.

Şekil 3'te elde ettiğimiz makalelerde atıf yapılmış yazarların co-citation analizinin görselleştirilmiş hali bulunmakadır. Birliktelik atıfı anlamına gelen bu analiz ile birliktelik ilişkileri belirlenerek hangi yazarların kimlerle çalıştığını bulmak olasıdır. Aşağıdaki şekilde yazarların birbirleri ili birliktelik ilişkileri görülmektedir.

Şekil 3 incelendiğinde, 5 farklı kümenin şekilde yer aldığı görülmektedir. Bu kümeler kırmızı, mavi, yeşil, sarı ve mor renk ile betimlenmiştir. Kırmızı küme 340 yazarı içinde bulundurarak en fazla elemana sahip olan kümedir. Bu küme

içerisinde Dean, J.'de yer almakta ve kümenin merkezinde bulunmaktadır. Ayrıca, Chen, M., ve Zaharia, M.'de kırmızı küme içerisinde yer almıştır. 123 elemana sahip yeşil küme ise en fazla elemana sahip 2. küme konumundadır. Burada dikkat çeken yazalar ise Laney, D., Mayer-Schonberger, V., Boyd, D., Kitchen, R. ve Lazer, D.'dir. Mavi küme ise 110 elemandan oluşmaktadır. Davenport, Th., Mcafee, A., Gandomi, A., Chen. Hc. ve Manyika, J. gibi alanın önemli isimleri burada yer almakta ve bu kümeyi güçlü kılmaktadırlar. Sarı küme 6 elemandan oluşurken mor küme ise sadece iki elemandan oluşmaktadır. Bu iki kümenin güç ilişkisi diğer 3 kümeye göre daha düşüktür.



Şekil 3. Atıflara göre yazarların birliktelik analizi

4.4 Anahtar Kelimelere Göre Co-Occurence Analizi

Bu analiz için yazarların vermiş olduğu anahtar kelimelerden yararlanılmıştır. Bu analiz sonucunda tekrar sayısı en az 1 olan 9340 adet farklı kelime elde edilmiştir. Çalışma için farklı eşik değerleri denenmiş ve en uygun eşik değerinin programın da önermiş olduğu 5 olduğu görülmüştür. 5 eşik değerini geçen toplam 335 adet kelime bulunmaktadır. Bu kelimelerin birbirleri ile ilişkileri birliktelik kuralları yani association rules değerleri ilgili program sayesinde belirlenmiştir. Küçük kümeler anlamlandırılamadığı için resolution değeri 0.80 olarak ayarlanmış böylece kümeler daha anlamlı hale getirilmeye çalışılmıştır. Bu sayede dağıtılan küçük kümelerin içerisindeki elemanlar kendilerine ilişki bakımından yakın olan kümelerin içerisine dahil edilmiştir.

İlk olarak en fazla tekrarlanan 25 kelimenin listesi Tablo 4'te verilmiştir. Tablo 4 incelendiğinde, yazarların Büyük Veri çalışmalarında en fazla kullandıkları anahtar kelimenin "Big Data" olduğu görülmektedir. Büyük veri analizinde ilk sırada bu değerin çıkması oldukça normal ve anlaşılabilir bir durumdur. Önemli olan bu kelimeyle birlikte hangi kelimelerin kullanıldığıdır. Bu bağlamda, Cloud Computing, Machine Learning, Mapreduce, Data Mining, Hadoop, Internet of Things, Social Media, Privacy ve özellikle Data Analytics gibi kelimelerin ön plana çıktığı ve büyük veri kavramında önemli yer edindikleri görülmektedir. Unutulmamalı ki Büyük Veri kavramı veri analizi olmadan pek bir önem arz etmeyecektir. Bu yüzden tabloda içerisinde Analytic geçen farklı kelime yapıları bulunmaktadır.

Bu gibi çalışmalarda bazen kelime tekrarları ile karşılaşılabilmektedir. Bunun nedeni anahtar kelimelerinin tek bir formatta kullanılmasının zor olmasıdır. Örneğin, yazarlar bazen kısaltma kullanırken bazen de anahtar kelimenin uzun formatını (IOT veya Internet of Things gibi) kullanabilmektedir.

"Big Data" anahtar kelimesinden sonra en fazla bağlantı gücüne sahip olan kelimenin 365 bağlantı gücü ile Cloud Computing olduğu görülmektedir. Cloud Computing'i sırasıyla, Machine Learning (310), Mapreduce (299), Hadoop (275), Big Data Analytics (258) ve Data Mining (256) izlemektedir.

Büyük veri kavramı denildiği zaman ilk akla gelenler Mapreduce algoritması ile Hadoop mimarisidir. Bu yüzden Big Data anahtar kelimesi ile birlikte ön planda olmaları şaşırtıcı değildir.

Tablo 4. Anahtar kelimelerin tekrar sayısı ve toplam bağlantı gücü

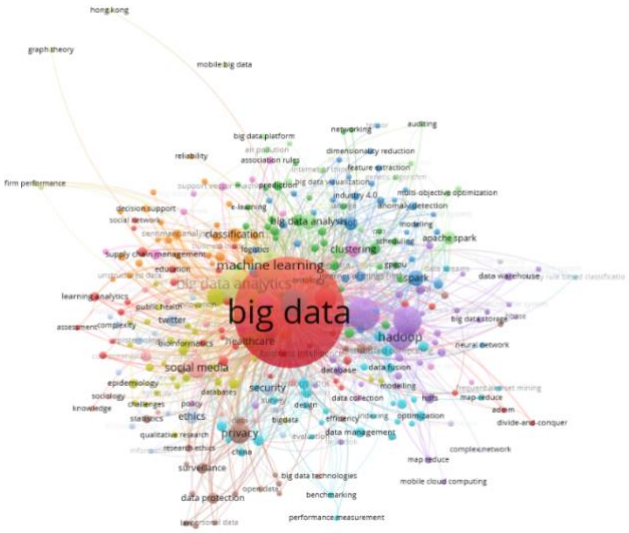
No	Anahtar Kelime	Tekrar Sayısı	Toplam Bağlantı Gücü
1	Big Data	1906	2596
2	Big Data Analytics	171	258
3	Cloud Computing	166	365
4	Machine Learning	137	310
5	Mapreduce	131	299
6	Data Mining	112	256
7	Hadoop	102	275
8	Internet of Things	73	197
9	Privacy	70	184
10	Social Media	65	134
11	Analytics	57	141
12	Data Analytics	55	141
13	Data Science	42	93
14	Spark	39	91
15	Big Data Analysis	38	31
16	Deep Learning	33	66
17	Clustering	32	74
18	Ethics	31	76
19	Security	30	76
20	IOT	30	69
21	Classification	29	60
22	Data Analysis	29	53
23	Healthcare	29	71
24	Apache Spark	26	57
25	Twitter	25	52

Bu çalışma kapsamında bir co-occurence analizi yapılmış ve görselleştirmesine ise Şekil 4'te yer verilmiştir.

Elde edilen görsel incelendiğinde, 15 farklı kümenin oluştuğu görülmektedir. Bu kümelerin analizi için uzmanlık bilgisi gerekmekte ve kümeler içerisinde yer alan tüm elemanlar detaylı bir şekilde incelenerek ve tamamı göz önünde bulundurularak kümelerin adları belirlenmelidir.

Kümelerin içerikleri incelendiğinde şu kanılara varılmıştır;

- Big Data kavramının yoğun olduğu kırmızı küme genellikle "Veri" kavramı üzerine odaklanmıştır ve bu kümede 34 eleman bulunmaktadır.
- Yeşil küme ise daha çok "Optimizasyon" üzerine yoğunlaşmış ve içerisinde 34 melemanı barındırmaktadır.
- Mavi küme ise "Endüstri 4.0" kavramlarını içerisinde barındırmakta ve 34 elemana ev sahipliği yapmaktadır.
- Sarı küme ise "Sağlık" üzerine yoğunlaşmış ve 34 elemanı bulunmaktadır.
- Mor küme ise "Hadoop Mimarisi ve MapReduce algoritması" üzerine yoğunlaşmıştır. Bu küme 33 elemandan oluşmaktadır.
- Açık mavi küme ise "Parallel Computing" üzerine odaklanmış ve 33 elemanı bulunmaktadır.
- Turuncu kümede "Business Analytics ve Inteligence" ile ilgili kelimeler yer almaktadır. Bu kümede 27 eleman yer almaktadır.
- Kahverengi küme "Statistics" üzerine yoğunlaşmıştır. Bu küme 26 elemandan oluşmaktadır.
- Eflatun küme ise "Büyük Veri Analiz Yöntemleri" üzerine yoğunlaşmış ve 20 elemanı bulunmaktadır.
- Diğer kalan 6 küme içerisindeki kelime sayısı doğru bir analiz yapmak için yeterli değildir.



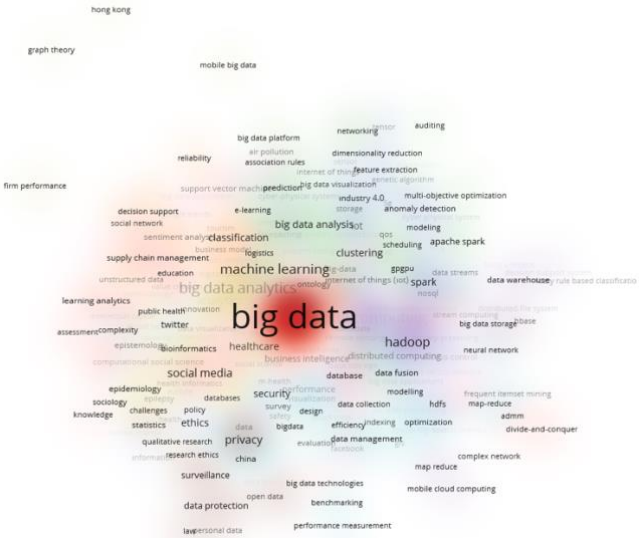
Şekil 4. Anahtar kelimeye göre co-occurence analizi

Şekil 5'te anahtar kelimelerin yoğunluk haritasına yer verilmiştir. Yoğunluk haritası sayesinde kelimelerin hem yoğunlukları daha net anlaşılmakta hem de birbirleri ile yakınlık ilişkilileri de görülebilmektedir.

Şekil incelendiğinde, Machine Learning, Social Media, Hadoop, Health Care, Security, Clasification, Clustering gibi kelimelerin net bir şekilde alanı temsil ettiği görülmektedir.

Alanı tasvir etmede bu görselleştirmenin de kullanılması ileride yapılacak çalışmalara daha fazla yön verebilir.

Elde edilen şekillerden de anlaşılacağı üzere, Büyük Veri kavramı birçok kavram ile iç içe yer almaktadır. Özellikle, Veri Analizi, Endüstri 4.0, Optimizasyon, Paralel Hesaplama, Hadoop Mimarisi ve MapReduce Algoritması, Kümeleme, Sınıflama, Birliktelik Kuralları ve İstatistiksel Yöntemler ile çok sık anılmakta ve birlikte kullanılmaktadır.



Şekil 5. Anahtar kelimelerin yoğunluk haritası

Bunların dışında bizlere uygulama alanlarında da önemli ipuçları vermektedir. Büyük verinin en fazla kullanıldığı alan eldeki verilere göre Sağlık alanı olarak ortaya çıkmaktadır. Fakat, pazarlama, parakendecilik, güvenlik gibi farklı alanlarda da kullanıldığı görülmektedir.

Özellikle Sosyal Medya analizinde de önemli bir yeri bulunmakta ve twitter verilerinin analizinde kullanılmaktadır. Bu alanda Sentiment Analysis yani Duygu Analizi yöntemi ise ilk sırada yerini almaktadır.

Unutulmaması gereken önemli husus ise yapılan tüm bu işlemlerin aslında karar vermeyi kolaylaştırmak için yapıldığıdır. Bu yüzden, karar alma, karar mekanizmaları, karar ağaçları, karar destek sistemleri gibi bir çok kavram da ilgili haritalarda yer almaktadır.

5 Sonuçlar

Bu çalışmada büyük veri kavramının öne çıkanları bilim insanlarının yapmış oldukları makale çalışmaları kullanılarak elde edilmeye çalışılmıştır. Bu kapsamda Web of Science Core Collection veri tabanında bulunan SCI-Exp, SSCI ve ESCI indexlerinden yararlanılmıştır. Bu indexler içerisinde bulunan dergilerdeki makalelerin başlıklarının içerisinde "Big Data" kavramı aratılmıştır. Arama alt kriteri olarak ise son beş yıl içerisinde yayınlanmış ingilizce makaleler belirlenmiştir. Toplamda 3868 sonuç elde edilmiştir ve bu makalelerin tüm verileri "plain text" olarak veribanından indirilmiştir.

İkinci aşamada ise verilerin birleştirilmesi aşaması gelmektedir. Birleştirme işlemi için "BibExcel" programı kullanılmış ve tek bir dosya elde edilmiştir. Verinin

temizlenmesi ve parçalara ayırma işlemleri de "BibExcel" programı yardımı ile yapılmıştır.

Üçüncü aşamada ise verinin ilişkilerinin kurulması ve bu ilişkilerin görselleştirilmesi gelmektedir. Bu bölümde "Vosviewer" programının 1.6.10 sürümünden yararlanılmıştır. Tüm görselleştirme işlemleri bu program vasıtasıyla yapılmıştır. Çalışma esnasında kümelerin ilişkilerinin daha belirgin olmasını sağlamak için "Resolution" değeri 0.80 olarak kabul edilmiştir. Her uygulama içerisinde de kendisine özgü olarak alt kriterler uygulanmış ve uygulanan bu alt kriter değerlerin "Bulgular" içerisindeki ilgili kısımlarda değinilmiştir.

Ülkeler bazında yapılan araştırma sonucuna göre, Amerika ve Çin alanda öncü ülkeler olarak görülmekte ve diğer ülkeler ile aralarında büyük bir farkın olduğu görülmektedir. Türkiye ise kendisine 31. sırada yer bulabilmiş ve bu alanda gerilerde kalmıştır. Görselleştirme haritaları incelendiğinde ise Amerika ve Çin'in ayrı kümelerde bulunduğu görülmüş ve alandaki güçlü ülkelerin bu iki ülke ile yüksek ilişki içerisinde olduğu anlaşılmaktadır. Türkiye ise bu iki ülkenin bulunduğu kümede yer almayıp, üçüncü farklı bir kümede kendisine yer bulmuştur. Türkiyenin ortak çalıştığı ülkelerin Mısır, Norveç ve Macaristan gibi ülkeler olduğu görülmektedir.

Çalışmanın ikinci aşamasında yazarların birliktelik ilişkileri incelenmiştir. Yazarlar ile ilgili ilk analiz aldıkları atıflar üzerinden yapılmıştır. Bu çalışma sonucunda, Rameshwar Dubey'in aldığı 328 atıf ilk sırada yer aldığı görülmektedir. Ayrıca yayın başına en fazla atıf alan yazarın da aynı yazar olduğu görülmektedir. Yayın sayısı bakımından bakıldığı zaman ise Anand Paul, 20 yayın ile öne plana çıkmaktadır. Türkiyeden bu listeye girebilmiş bir akademisyen bulunmamaktadır.

Yazarlar ile ilgili yapılan ikinci analiz ise elde edilen makalelerde verilmiş olan atıflardan birliktelik atıfı analizi ile yapılan yazar ilişkileri araştırmasıdır. Elde edilen makalelerde en fazla atıf yapılan yazar J. Dean'dir. 3868 makalenin 439 tanesinde bu yazarın yaptığı çalışmalara yer verilmiştir. Dean, J.'yi 292 atıf ile Manyika, J., 291 atıf ile Boyd, D., 287 atıf ile 275 atıf ile Davenport, Th. izlemektedir. Toplam kuvvet bağı en yüksek olan yazar tekrardan J. Dean'dir onu Davenport Th. izlemektedir. Yazarların birlikltelik haritası incelendiğinde 5 farklı kümenin oluştuğu görülmüş ve en güçlü kümenin merkezinde Dean. J.'nin yer aldığı tespit edilmiştir. En yüksek bağlantı gücüne sahip yazar J. Dean'dir. Mavi küme ise güçlü yazarların birlikte bulunduğu bir küme olarak dikkat çekmektedir. Davenport, Th., Mcafee, A., Gandomi, A., Chen. Hc. ve Manyika, J. gibi alanın öncülerinin bu kümede olduğu ve birliktelik ilişkilerinin yüksek olduğu görülmektedir.

Çalışmanın son kısmında ise anahtar kelimelere göre metin madenciliği analizi yapılmış ve alanda öne çıkan kelimeler tespit edilmiştir. Daha sonra ise co-occurence analizi sayesinde bu terimlerin birlikte kullanımları ve ilişkileri resmedilmiş ve en son olarak da yoğunluk haritası ile kelimelerin yoğunlukları ve birbirlerine yakınlıkları gösterilmiştir.

Yazarların vermiş oldukları anahtar kelimelere göre terim analizi yapıldığında ilk sırada "Big Data" teriminin olduğu görülmektedir. Bunun nedeni yazarların yapmış oldukları büyük veri çalışmalarında anahtar kelimelerin içerisinde ilk sıraya bu terimi yerleştirmiş olduklarıdır. Bu da aslında çalışmanın doğru yürütüldüğünün bir göstergesidir. Özellikle Data Analytics terimi dikkat çekmekte ve farklı türlerde yazıldıkları da görülmektedir. Büyük veri kavramı veri analizi olmadan bir anlam taşımayacaktır. Bu yüzden farklı şekillerde yazılmış olan bu terimin önem derecesi de yüksektir diyebiliriz. Bunun dışındaki diğer önemli terimler ise; Cloud Computing, Machine Learning, Mapreduce, Data Mining, Hadoop, Internet of Things, Social Media ve Privacy'dir.

Bağlantı güçleri bakımından analiz edildiğinde ise "Big Data" teriminden sonra "Cloud Computing" gelmektedir. Bu veri bize sadece tekrar sayısı bakımından bakmanın yanlış olacağını ayrıca bağlantı gücünün de incelenmesi gerektiğini göstermektedir. Machine Learning, Mapreduce, Hadoop, Big Data Analytics ve Data Mining kavramları Cloud Computing'i izlemektedir. Büyük veri denildiği zaman akla Mapreduce algoritması ile Hadoop mimarisi gelmektedir. Bu yüzden bu iki terimin en çok bağlantı gücüne sahip ilk 5 terim arasında olması şaşırtıcı değil aksine beklenen bir durumdur.

Görselleştirme aşamasında 15 faklı küme bulunmuştur ve bu kümeler şu şekilde isimlendirilmiştir; "Veri", "Optimizasyon", "Endüstri 4.0", "Sağlık", "Hadoop Mimarisi ve Map Reduce Algoritması", "Paralel İşleme", "İşletme Analiği ve Zekası", "İstatistik" ve "Büyük Veri Analiz Yöntemleri". 6 kümeye ise içlerinde bulunan terim sayısı yetersiz olduğu ve anlamlandırlamadığı için isim verilmemiştir. Yukarıdaki bu kavramlar aslında Büyük Veri kavramının ilişki haritasını net bir şekilde ortaya koymakta ve bizlere yol göstermektedir.

Bunlara ek olarak, yapılan bu çalışma ile aslında Büyük Verinin en fazla kullanıldığı alanlarında farkına varılmıştır. Yapılan incelemede, "Sağlık" alanında çok fazla kullanıldığı ve bu alanı sırasıyla pazarlama, parakendecilik ve güvenliğin izlediği görülmüştür.

Ayrıca, "Sosyal Medya" üzerinden yapılan Duygu analizininde önemli bir yer edindiği ve en çok twitter verilerinin kullanıldığı görülmektedir.

Büyük Veri, Veri Madenciği ve Veri Analitiği gibi konuların genel amacı karar vermeyi kolaylaştırmak ve doğru karar almak için yöneticilere ya da kişilere yol göstermektir. Bu bağlamda, ilişki haritalarında dikkat çeken diğer terimler ise karar alma, karar mekanizmaları, karar ağaçları ve karar destek sistemleridir.

Bu gibi çalışmalarda karşılaşılan en büyük sorun dil birliğine varılamaması ve yazarların anahtar kelimeleri isimlendirirken farklı yapıları kullanmasıdır. Bu da analiz yaparken tekrarlarla karşılaşılmasına neden olmakta ve aynı terimin farklı yapılarının olduğu görülmektedir. Buna bir örnek verecek olursak; yazarlar bazen kısaltma kullanırken (IOT gibi) bazen de anahtar kelimenin uzun formatını (Internet of Things gibi) kullanabilmektedir. Bu da veri tekrarına yol açmaktadır. Bu veriler birleştirilebilirse uygulama programlarında daha verimli sonuçlar elde edilebilir.

6 Kaynaklar

[1] İ. Özdemir, Ş. Sağırlıoğlu, "Denetimlerde Büyük Veri Kullanımı Ve Üzerine Bir Değerlendirme". GU J Sci, Part C, 6(2), 470-480, 2018.

[2] Ş. Işıklı,"Büyük Veri, Epistemoloji ve Etik Tartışmalar", Academic Journal of Information Technology, 9, 2014.

[3] N. Gürsakal, Büyük Veri, Dora Yayıncılık, 2013.

[4] J. Gantz, D. Reinsel, "Extracting value from chaos", IDC iview, 1142, 1-12, 2011.

[5] D. Tellan, "Büyük Veri Türbülansını Yönetmek", Türkiye Bilişim Derneği 31. Ulusal Bilişim Kurultayı, ISBN: 978-9944-5291-8-1, 41-42, 2014.
[6] Y. Gahi, M. Guennoun, H. T. Mouftah, "Big Data Analytics: Security and Privacy Challenges", 2016 IEEE Symposium on Computers and Communication (ISCC), Messina, Italy, 952-957, 2016.
[7] E. Aktan, ''Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu''. Bilgi Yönetimi Dergisi, 3-4, 2018.
[8] M. Doğan, Büyük Verinin Kişiler Ve Kurumlar Üzerinde Etkileri, Yüksek Lisans Tezi, İstanbul Bilgi Üniversitesi, Sosyal Bilimler Enstitüsü, 2014.
[9] B. Cyganek, M. Grafia, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, M. Woźniak, "A Survey of Big Data Issues in Electronic Health Record Analysis", Applied Artificial Intelligence, 30, 2016.
[10] A., Gandomi, M. Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics", International Journal of Information Management, 137-144, (2015).
[11] Minelli, M. Chambers, A. Dhiraj, Big Data Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Hoboken, NJ, U.S.A. Wiley CIO Series, John Wiley & Sons, 2013.
[12] F. Ohlhorst, Big Data Analytics Turning Big Data into Big Money, Hoboken, NJ, U.S.A. J. Wiley and SAS Business Series, John Wiley & Sons, 2013.
[13] S. Chandra, S. Ray, R. Goswami, Big Data Security: Survey on Frameworks and Algorithms. 2017 IEEE 7th International Advance Computing Conference (IACC),, (s. s.48-54.). Hyderabad, India 2017.
[14] Internet: A. Garip, Büyük Veri Kriterleri, https://atacangarip.wordpress.com/2015/08/24/buyuk-veri-bilesenleri/, 10.01.2019.
[15] Özköse, H., Arı, E. S., & Gencer, C. (2015). Yesterday, today and tomorrow of big data. Procedia-Social and Behavioral Sciences, 195, 1042-1050.
[16] R. Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, New York USA, 2007.
[17] S. Sumathi, S. N. Sivanandam, Introduction to Data Mining and Its Applications, Berlin: Springer-Verlag, 2006.
[18] K. Seçkin, Metin Madenciliğinde Kullanılan Yöntemlerin Karşılaştırılması: Siyasi Parti Liderlerinin Grup Genel Toplantı Konuşmaları İle Bir Uygulama, Yüksek Lisans Tezi, Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, 2011.
[19] C. Melek, Metin Madenciliği Teknikleri İle Şirketlerin Vizyon İfadelerinin Analizi, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, 2012.
[20] A. Güven, Ö. Ö. Bozkurt, O. Kalıpsız, ''Veri Madenciliğinin Geleceği'', Akademik Bilişim, Dumlupınar Üniversitesi, Kütahya