Elektronik Sosyal Bilimler Dergisi www.esosder.org
Electronic Journal of Social Sciences ISSN:1304-0278
Yaz -2017 Cilt:16 Sayı:62 (1111-1122) Summer-2017 Volume:16 Issue:62

VERİ MADENCİLİĞİ ÇALIŞMALARI ÜZERİNE BİR ANALİZ: TÜRKİYE ADRESLİ YAYINLAR

AN ANALYSIS ON THE DATA MINING STUDIES: PUBLICATIONS ADDRESSING TURKEY

Alptekin DURMUŞOĞLU[1]

Öz

Veri Madenciliği (VM); ham ve yığınlar halinde olan veri setlerinde, doğrudan tespit edilemeyen, gizli ve anlamlı bilgilerin keşfedilmesine yönelik olarak bir dizi yöntemin kullanıldığı bir araştırma alanıdır. Son yıllarda, veri depolama ve veri toplama teknolojilerindeki ucuzlamaya paralel olarak artan veri miktarıyla VM uygulamaları tüm dünyada giderek yaygınlaşmıştır. Uygulamadaki sayıca artış, konuyla alakalı bilimsel yazındaki makalelerin de sayıca artışını beraberinde getirmiştir. Bahsi geçen yayınların, bilimsel alandaki genel gidişatını belirlemek, gerek araştırmacılar gerekse uygulayıcılar için çok çeşitli faydalar sağlayacaktır. Bu doğrultuda; bu çalışmada, son on yılda indeksli dergilerde (Science Citation Index-SCI, Science Citation Index-expanded, Social Sciences Citation Index-SSCI ve Arts and Humanities Citation Index-AHCI) ve konferanslarda yayınlanan (Conference Proceedings Citation Index) VM çalışmaları ele alınmakta, bu çalışmaların; ülke, yazar, yöntem vb. dağılımları ortaya konmaktadır. Ayrıca Türkiye adresli VM çalışmaları da ayrı bir başlık altında ele alınmış ve bu çalışmaların nitelik ve nicelikleri irdelenmiştir.

Anahtar kelimeler: Veri Madenciliği, bibliyometrik analiz, Türk araştırmacılar

Abstract

Data Mining (DM) is a research field where a sequence of methods are applied to discover hidden and significant knowledge that cannot be directly detected inside bulk and raw data sets. In the recent years with the increasing amount of data in parallel to the advances in data storage and collection technologies, DM applications have become widespread. Increase in the number of applications has entailed the increase in the number of scientific publications. Determining the general state of affairs can provide several benefits both for the implementers and the researchers. In this direction, the DM studies published in indexed journals (Science Citation Index-SCI, Science Citation Index-expanded, Social Sciences Citation Index-SSCI ve Arts and Humanities Citation Index-AHCI) and conferences (Conference Proceedings Citation Index) on the last ten years has been covered and the country, author , method etc. distribution of those studies has been put forward. In addition to that, the DM studies addressing Turkey has been handled under a separate title and qualifications and quantifications have been examined.

Keywords: Data mining, bibliometric Analysis, Turkish researchers

[1] Yrd.Doç.Dr., Gaziantep Üniversitesi, durmusoglu@gantep.edu.tr

1. GİRİŞ

Veri depolama ve toplama teknolojilerinin çeşitlerindeki (ör: mobil ve çevrim içi uygulamalar, Radyo Frekanslı Tanımlama teknolojileri, barkodlar, Optik Karakter Tanımlama teknolojileri vb…) ve kapasitelerindeki artışla, veri tabanlarında depolanan ham veri miktarı zamanla artmıştır. Ancak hacimce fazla ve karmaşık durumdaki ham veri (işlenmemiş, analize tabi tutulmamış) çoğu zaman kendi başına yeterince bilgi verici (enformasyon) nitelikte olmamaktadır. Hâlbuki ham verinin içerisinde yer alan bilgi: karar desteğinin sağlanması ve ilgilenilen fenomenin daha iyi anlaşılması konusunda oldukça faydalı olabilme potansiyeline sahiptir (U. M. Fayyad, 1997). Bilgi bir amaca yönelik işlenmiş veridir (Savaş, Topaloğlu, & Yılmaz, 2012). Bu bağlamda; veri içerisinde var olan ilişkilerin ve kalıpların keşfedilmesi için birbirini izleyen bilgisayar destekli faaliyetlerin bir bütünsellik içerisinde uygulanması veri madenciliği (VM) olarak tanımlanmaktadır (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

VM çok kısa bir sürede dikkatleri çekebilmiş, sürekli gelişmekte olan bir alandır (Glymour, Madigan, Pregibon, & Smyth, 1997). VM alanında yürütülen akademik çalışmaların da oldukça dikkat çekici boyutta olduğu söylenebilir. Öyle ki 2006-2015 yıllarını kapsayan on yıllık dönemde VM'yi konu edinen sadece Thomson Reuters Web of Science ("Thomson Reuters, New York, NY, ABD," 2016) veri tabanında 30.683 adet bilimsel dergi ve konferans makalesi yer almıştır. Veri madenciliği ve bilgi keşfi alanında gerçekleşen bilimsel gelişmeler; yöntemsel gelişmeler ve uygulama alanlarındaki gelişmeler olarak iki ana kategoride incelenebilir. Yöntemsel ilerlemeler, bu alanın önemli parçaları olan istatistik, veri tabanları, makine öğrenme ve yapay zekâ alanlarındaki ilerlemelerin bir araya gelmesiyle ortaya çıkmaktadır (U. Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). Veri madenciliği uygulama alanlarındaki ilerlemeler ise geniş bir yelpazede sınıflandırılabilir. Bu uygulama alanlarının başlıcaları şu şekilde sıralanabilir (Perner, 2010):

- Multimedya veri madenciliği
- Pazarlama alanında veri madenciliği
- Endüstriyel süreçlerde veri madenciliği
- Tıpta veri madenciliği
- Tarımda veri madenciliği
- Web madenciliği
- Finansta veri madenciliği

Bu çalışmayla amaçlanan; son on yılda indeksli (SCI-SCI Expanded-SSCI-AHCI) dergilerde ve seçkin konferanslarda yayınlanan veri madenciliği çalışmalarını niceliksel olarak ele almak, bu çalışmaların; ülke, yazar, yöntem vb. dağılımları ortaya koymaktır. Ayrıca Türkiye adresli veri madenciliği çalışmaları da ayrı bir başlık altında ele alınmış ve bu çalışmaların nitelik ve nicelikleri irdelenmiştir.

2. BİBLİOMETRİK ANALİZ

Günümüzde, araştırma denetimi ve değerlendirmesi, hem hükümetler hem de araştırma finansörleri için çok daha önemli bir hal almış ve bibliyometrik çalışmalar bilim politikası yetkilileri açısından bu değerlendirmenin önemli bir aracı olmuştur (Forsman, 2015). Bibliyometrik çalışmalar, yayınlara ait geçmiş kayıtların organize bir şekilde incelenmesi ve mevcut durumun tespitini içermektedir. Bu çalışmalar, araştırma eğilimlerinin belirlenmesi, henüz çalışılmamış alanların tespiti ve araştırma performansının ölçümlenmesi amacıyla kullanılabilmektedir.

Bibliyometrik çalışmalar kapsamında en yaygın kullanılan analizler; yazar analizi, kavram haritaları, küme-faktör analizleri, atıf ve karşılıklı atıf analizleri olmuştur (Daim,

Rueda, Martin, & Gerdsri, 2006). Bugüne değin birçok akademik çalışma alanında bibliyometrik çalışma yapılmış olmasına karşın (ör:(D. Chen, Liu, Luo, Webber, & Chen, 2016; H.-Q. Chen et al., 2016; Durmuşoğlu, 2016; Garousi & Mäntylä, 2016; Merigó, Mas-Tur, Roig-Tierno, & Ribeiro-Soriano, 2015)); VM alanında bu tip çalışmaların oldukça kısıtlı olduğu bilinmektedir. VM alanında benzer bir çalışma, 2013 yılı öncesindeki 5 yılda yayınlanmış olan 11.577 makalenin incelenmesiyle gerçekleştirilmiştir (Durmuşoğlu & Dereli, 2013). Ancak ilgili bu çalışmada değerlendirme 5 yıl ile sınırlı kalmış ve konferans makaleleri kapsam dışı tutulmuştur. Bu çalışma ile kapsam genişletilmiş; incelenen unsurlar artırılmış ve sonuçların 5 yıllık performansa göre nicel ve nitel olarak farklılaştığı gözlemlenmiştir.

Bu amaçla, bu çalışmada 2006-2015 yıllarını kapsayan on yıllık dönemde VM'yi konu edinen 30.683 adet bilimsel dergi veya konferans makalesi/bildirisi bibliyometrik analize tabi tutularak incelenmiştir. Çalışmaya esas teşkil eden makale verileri Thomson Reuters-Web of Science ("Thomson Reuters, New York, NY, ABD," 2016) veri tabanı kullanarak derlenmiştir. Bilindiği üzere WoS disiplinler arası bir atıf indeksi veri tabanıdır ve yeni dergilerin kapsama alınması/performansı düşen dergilerin sistemden çıkarılması şeklinde bir güncelleme mekanizmasına sahiptir. Bu çalışma için WoS veri tabanı sorgusu konu (topic) alanına "veri madenciliği" ("data mining") girilerek gerçekleştirilmiştir. İlgili veri tabanının "konu" alanında sorgulama yapılması; tüm makalelerin; başlığı, özeti ve anahtar kelimeleri taranmasını sağlamaktadır.

3. VERİ MADENCİLİĞİ YAYINLARI TEMEL İSTATİSTİKLER

Bu çalışmada, son 10 yıl boyunca (2006-2015) dergi ve konferans bildiri kitaplarında yer alan çalışmalar incelenmiş ve çeşitli istatistiki analizlere tabi tutulmuştur. Çalışma boyunca; incelenen 30.683 yayın; tür (bildiri/dergi makalesi/inceleme/düzeltme/editör notları), yıl, ülke, yayınlandığı dergi ve konferanslar gibi alt başlıklar altında incelenmiştir.

Çalışma kapsamında, veri madenciliği (VM) ile ilgili son on yılda indeksli dergilerde yapılmış yayınlar araştırıldığında 30.683 yayının mevcut olduğu tespit edilmiştir. Bu çalışmaların yayın kategorilerine göre dağılımı Tablo 1'de sunulmaktadır. Tablo 1'de de gösterildiği üzere son 10 yıl itibariyle, bu alanda yayınlanan bildiri makaleleri, dergi makalelerinden fazla olmuştur.

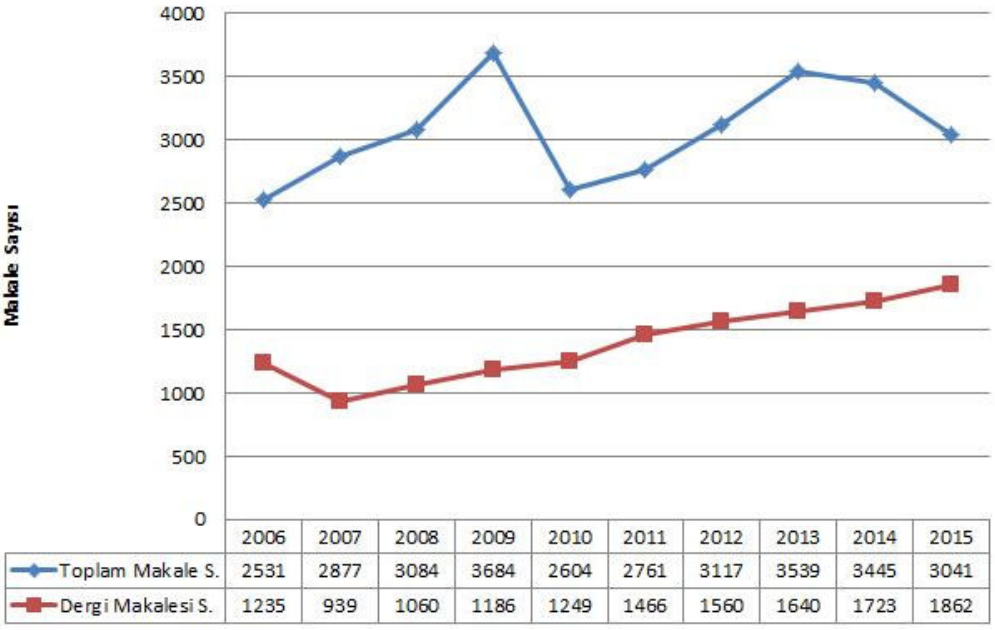
Tablo 1: Son 10 yılda gerçekleştirilen yayınların kategorik dağılımı*

Yayın Türü	Yayın Sayısı	Yüzde
Bildiri Makalesi (Tam metin+Özet)	16639	54.23 %
Dergi Makalesi	13920	45.37 %
Tarama/İnceleme Makalesi	671	2.19 %
Editör Notları	277	0.90 %
Kitap İnceleme	40	0.13 %
Düzeltme Makalesi	33	0.11 %

(*Bir makale birden fazla kategoride yer alabildiğinden yüzde sütununda yer alan yüzdelerin toplamları 100 değildir).

Kapsam dâhilindeki yayınların (konferans ve dergi) yıllara göre dağılımları ise Şekil 1'de gösterilmektedir. Yayın sayıları konferans bildirileriyle birlikte düşünüldüğünde; 2006 yılında 2.531 iken 2015 yılında 3.041 olmuştur. Yayınların en fazla sayıda olduğu yıl ise 3.684 ile 2009 yılı olmuştur. Toplam makale sayısında düzenli bir artış veya azalış olmadığı

ve dalgalı bir seyir olduğu görülmektedir. Beri yandan; sadece indeksli dergilerde yayınlanan makale sayıları incelendiğinde (Şekil 1); 2007 yılındaki düşüş haricinde istikrarlı bir artışın olduğu görülmektedir. Bu nedenle, toplam makale sayısındaki dalgalanmanın konferans bildirilerinden kaynaklandığı anlaşılmaktadır. Konferans katılımları ülkelerin bilim politikalarından yanı sıra bireysel tercihler de belirleyici olmaktadır. Uluslararası konferanslara katılım sebeplerinin araştırıldığı bir çalışmada (MHM & PhD, 2000); seyahat fırsatının niteliği, iş amaçlı veya politik faaliyetler, ortam değişimi, bağlantı kurma ve eğitim fırsatlarının katılımda en önemli faktörler olduğu belirlenmiştir. Bu nedenle konferans katılımıyla dergilerde makale yayınlamanın farklı motivasyonları olması, niceliksel olarak benzer artış ve azalışlar olmamasının gerekçesi olarak kabul edilebilir.



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Toplam Makale S.	2531	2877	3084	3684	2604	2761	3117	3539	3445	3041
Dergi Makalesi S.	1235	939	1060	1186	1249	1466	1560	1640	1723	1862

Şekil 1. VM makalelerin yıllar içindeki dağılımı

Thomson and Reuters, konferanslara ait bildirileri indekslerken/tararken (konferans bildirileri atıf indeksi), bir dizi değerlendirme yapmakta ve ancak belirlenen eşik değerin üzerindeki konferansları kapsama almaktadır. Seriler halinde düzenlenen konferanslarda, bir yıl kapsam dâhilinde olunması bir sonraki yıl düzenlenecek konferans bildirilerinin taranacağını garanti etmemekte, değerlendirme dinamik olarak her sene tekrarlanmaktadır. Bu nedenle bu çalışma kapsamında ele alınan konferans makalelerinin/bildirilerinin belirli bir eşit değeri aşan, nitelikli makaleler olduğu ifade edilebilir. VM konusunda bildiri makalelerine en fazla yer veren (Thomson and Reuters tarafından taranan- Conference Proceedings Citation Index) konferanslar şunlardır:

- IEEE International Conference on Data Mining,
- IEEE International Conference on Systems Man and Cybernetics
- International Conference on Machine Learning and Cybernetics
- International Conference on Advanced Data Mining and Applications
- Pacific-Asia Conference on Knowledge Discovery and Data Mining
- International Conference on Fuzzy Systems and Knowledge Discovery

Belirtilen konferansların VM alanının önemli konferansları olduğu ifade edilebilir. Öyle ki; "IEEE International Conference on Data Mining" bildiriler kitabında yayınlan makaleler yıllık ortalama 100 civarında atıf almaktadır. Bir yılda 60-80 arasında bildirinin

kabul edildiği bu konferansta bu her yayın için yıllık bir civarı bir atıf ortalaması anlamına gelmektedir.

Bilindiği üzere internet teknolojisindeki gelişmelere paralel olarak; bilimsel dergiler de çevrim içi makale başvuru ve hakemlik/değerlendirme sistemleriyle süreçleri kolaylaştırma çabasına girmişlerdir. Benzer şekilde; araştırmacılar da kaynaklara daha kolay erişebilir bir hale gelmiş ve makale başvuru sayılarında/yayınlanan makale sayılarında ciddi artışlar yaşanmıştır (Vinkler, 2013). Bu nedenle; bir bilimsel alandaki gelişmeleri düzenli olarak takip edebilmek oldukça zorlu bir hal almıştır. Bu nedenle "tarama/inceleme" makaleleri önemli bir işlev görmeye başlamış ve seçilmiş bir bilim alanı için gelişmeleri ana hatlarıyla araştırmacılara sunmuşlardır. Özellikle; çalışma alanı arayışlarında; boşlukların ve eksiklerin belirlenmesinde bu tip yayınların önemli katkıları olduğu ifade edilebilir. Benzer şekilde; VM alanının çok-disiplinli bir çalışma alanı olmasına ve mevcut yayın sayılarının oldukça fazla olmasına bağlı olarak "tarama/inceleme" makalesi sayısı da dikkat çekici rakamlara (671-Tüm yayınların %2.19'u) ulaşmıştır.

VM alanındaki makalelerinin %45.37'sini oluşturan 13.920 dergi makalesinin yayınlandığı dergilerin isimleri ve bu dergilere ait uluslar-arası Bilimsel Yayınları Teşvik (UBYT) Programı puanları Tablo 2'de sunulmuştur. İndeksli dergilerde yayınlanan VM makalelerinin, %5'inin; uzman sistemler ve yapay zekâ konularında yayınlara yer veren "Expert System with Applications" (ESWA) isimli dergide yayınlandığı görülmektedir. İkinci sırada yer alan "IEEE Transactions On Knowledge & Data Engineering" dergisi; birinci sırada yer alan ESWA dergisinin yarısından daha az sayıda VM konusunda yayına yer vermiştir. Bu durum ESWA dergisinin yayın politikasıyla yakından ilişkilendirilebilir. ESWA dergisinin yıllar itibariyle yer verdiği toplam makale sayıları ve makalelerin VM ile ilgili olanlarının sayıları Tablo 3'de sunulmuştur (editoryal notlar kapsam dışında tutulmuştur). VM konusunda bu dergide basılan 701 makalenin (20 editör notu çıkarılmıştır) 426 âdetinin 2009-2012 yıllarında basıldığı anlaşılmaktadır. Aynı yıllarda derginin; her yıl 1000'den fazla yayına yer verdiği görülmektedir. ESWA'da basılan makale sayısı yıllar itibariyle değişiklik gösterse de, yayınlanan makalelerin her yıl en az %7'sinin VM alanında olduğu anlaşılmaktadır.

Tablo 2: 2005-2015 yılları arası veri madenciliği üzerine yapılan çalışmaların yer aldığı ilk 20 dergi

Dergi İsmi	Makale Sayısı	Yüzde	UBYT Dergi Puanı
Expert Systems with Applications	721	5.00 %	33,43
IEEE Transactions on Knowledge & Data Engineering	274	1.95 %	67,84
Information Sciences	189	1.35 %	50,59
Knowledge and Information Systems	184	1.31 %	35,57
Knowledge Based Systems	179	1.28 %	34,99
Plos One	165	1.18 %	20,86
Data Mining and Knowledge Discovery	128	0.91 %	82,87
BMC Bioinformatics	124	0.88 %	50,33
Applied Soft Computing	123	0.88 %	46,63
Intelligent Data Analysis	120	0.86 %	12,92
International Journal of Data Mining & Bioinformatics	102	0.73 %	3,11

Neurocomputing	96	0.68 %	26,05
Data & Knowledge Engineering	94	0.67 %	38,82
Decision Support Systems	90	0.64 %	48,90
European Journal of Operational Research	87	0.62 %	57,64
Nucleic Acids Research	75	0.53 %	68,49
BMC Genomics	70	0.50 %	37,61
Journal of Intelligent Information Systems	66	0.47 %	13,42
Bioinformatics	65	0.46 %	100,00
Journal of Biomedicial Informatics	60	0.43 %	77,90

Tablo 3: 2006-2015 yılları arasında "Expert Systems with Applications Dergisinde" basılan makale sayıları

Yıl	Tüm Makaleler	VM Makaleleri	Oran
2006	161	19	11,80%
2007	221	21	9,50%
2008	514	58	11,28%
2009	1362	129	9,47%
2010	1003	77	7,68%
2011	1699	119	7,00%
2012	1338	101	7,55%
2013	706	62	8,78%
2014	685	57	8,32%
2015	772	58	7,51%

VM alanında, dergilerde ve konferans bildiri kitaplarında yer alan makalelerde yazarların belirtmiş oldukları adreslerdeki ülkelere yönelik istatistikler Tablo 4'de sunulmuştur. Tablo 4'de belirtildiği üzere dergi makalelerinde; ABD, Çin Halk Cumhuriyeti ve Tayvan adresli VM makaleleri ilk üç sıradadır. Tablo 4'ün en sağında yer alan "tüm alanlar sıra" sütununda belirtilen rakamlar; ülkelerin tüm bilimsel faaliyetlerine göre sıralandığı "SCImago Journal & Country Rank" platformunun web sayfasından alınmıştır ((University of Granada, n.d.)). Bu sıralama; 1996-2015 yıllarında yayınlanmış ve Scopus veritabanı tarafından listelenen yayınlarına göre oluşturulmaktadır. Tüm alanlar sıralamasında 17. sırada yer alan Tayvan'ın VM alanında 3. olması oldukça dikkat çekicidir. Benzer bir durum Fransa için de geçerlidir. Tüm alanlarda 6. sırada olan Fransa; VM alanında ancak 11. olabilmiştir. Konferans makaleleri dikkat alındığında, Çin adresli makalelerin ABD adreslilerden daha fazla olduğu görülmektedir. İngiliz yazarlar konferanslar da 10. sırada iken, dergi makalelerinde 4. sıradadır. Bu durum İngiltere'yi adres veren yazarların konferanslar yerine dergileri tercih ettiği şeklinde yorumlanabilir. Bilindiği üzere; dergilere sunulan araştırmalar, konferanslara kıyasla daha zorlayıcı bir hakemlik sürecinden geçmektedir. Beri yandan; dergi makaleleri Türkiye ve benzeri ülkelerde akademik performans açısından daha değerli olarak değerlendirilmektedir. Genel kanı; konferanslarda sunulan bilimsel araştırmaların; diğer katılımcılardan da alınan geri beslemeler sonucu; olgunlaştırılarak ileri de dergi makalesine dönüşebileceği yönündedir. Bu durumda; konferans makalelerinde ön sıralarda yer alan ancak dergi yayınlarında görece düşük performans gösteren (VM alanı için Japonya bu duruma

uygun bir örnek olarak göze çarpmaktadır) ülkelerde "araştırma verimliliği" irdelenmeye değer bir araştırma konusu olarak durmaktadır.

Tablo 4: VM makalelerinin yazarlarının adreslerinde belirtikleri ülke göre dağılımı

Ülke	Dergi Makalesi	Yüzde	Konferans Makalesi	Yüzde	Dergi Makalesi	Konferans Makalesi	Tüm Alanlar
ABD	3665	26,33%	2107	12,66%	1	2	1
Çin Halk Cumh.	1983	14,25%	5046	30,33%	2	1	2
Tayvan	1053	7,56%	575	3,46%	3	5	17
İngiltere	752	5,40%	392	2,36%	4	10	3
İspanya	670	4,81%	465	2,79%	5	7	10
Almanya	665	4,78%	473	2,84%	6	6	4
Kanada	587	4,22%	421	2,53%	7	8	7
Avustralya	540	3,88%	392	2,36%	8	11	11
İtalya	538	3,86%	398	2,39%	9	9	8
G. Kore	503	3,61%	293	1,76%	10	14	12
Fransa	482	3,46%	377	2,27%	11	12	6
Japonya	388	2,79%	609	3,66%	12	4	5
Hindistan	386	2,77%	1182	7,10%	13	3	9
Brezilya	303	2,18%	291	1,75%	14	15	15
İran	272	1,95%	205	1,23%	15	18	22
Türkiye	249	1,79%	144	0,87%	16	22	20
Belçika	235	1,69%	98	0,59%	17	27	21
Hollanda	205	1,47%	101	0,61%	18	26	14
Polonya	200	1,44%	320	1,92%	19	13	19
Singapur	191	1,37%	114	0,69%	20	24	32

Yayın sayıları kadar önemli bir husus da bahsi geçen dergilerin itibar/etki düzeyleridir. ULAKBİM, 30 Nisan 2013 tarihinden başlayarak, dergiler için A, B, C sınıflamasından vazgeçmiş ve puanlama sistemine geçmiştir (Asan, 2013). UBYT Programının yeni uygulama esasları ile, Institute for Scientific Information (ISI) Citation Index Veri Tabanlarınca taranan hakemli ve sürekli dergilerde yayımlanmış uluslararası yayınlara verilecek teşvik miktarları, dergilerin 5 Yıllık Etki Faktörü (5-Year Impact Factor) değerlerinin yanı sıra Atıf Yarı Yaşı (Cited Half-Life) değerlerini de dikkate alan yeni bir formüle göre hesaplanmaktadır (TÜBİTAK-ULAKBİM, n.d.). Bahsi geçen VM makalelerinin sıklıkla yer aldığı dergiler arasında en yüksek UBYT puanına Bioinformatics dergisinin sahip olduğu görülmektedir. Biyoinformatik, biyolojik problemlerin çözümünde bilişim teknolojilerinin kullanılması esasına dayanan ve biyolojik olayların moleküler düzeyde açıklanmasına yardımcı olmakta olan görece yeni bir bilim dalıdır (Polat & Karahan, 2009). Bilgisayar bilimlerinde, son zamanlarda gelişmeye başlayan alanlarından biri olan veri madenciliğinin biyoinformatikte kullanılması gün geçtikçe artmaktadır (Polat & Karahan, 2009).

2. en yüksek puan ise "Data Mining and Knowledge Discovery" dergisine aittir. Bu dergi, Mart 1997 yılında ilk sayısını yayınlamış ve günümüze değin toplam 580 makaleye yer vermiştir ("Data Mining and Knowledge Discovery - Springer," n.d.). Dergi, yayın politikası gereği; bir sayıda ortalama 5 ila 7 makaleye yer vermektedir.

VM'nin disiplinler arası bir çalışma alanı olduğu ve farklı alanlarda çalışan araştırmacılar tarafından çalışıldığı bilinmektedir. Bu hususun doğruluğu, VM'yi konu edilen makalelerin hangi çalışma alanına girdiğinin özetinin verildiği Tablo 5'den de kolaylıkla anlaşılmaktadır. Belirtilen alanlarda "Bilgisayar Bilimleri"nin baskın olduğu görülmektedir. Çeşitli mühendislik disiplinlerinin de bu alanda belirgin bir yoğunlaşmaya sahip olduğu anlaşılmaktadır.

Tablo 5:VM makalelerinin çalışma alanlarına göre dağılımı*

(* Listelenen alanlar Thomson Reuters'in belirlediği çalışma alanlarıdır)

Sıra	Alan	Yayın Sayısı	Yüzde*
1	Bilgisayar Bilimleri	6707	48,18%
2	Mühendislik	3374	24,24%
3	Yöneylem Araştırması ve Yönetim Bilimleri	1270	9,12%
4	Matematik	935	6,72%
5	Biokimya ve moleküler biyoloji	813	5,84%
6	Matematiksel Hesaplamalı Biyoloji	605	4,35%
7	İş Ekonomisi	462	3,32%
8	Bilim Teknoloji ve Diğer Konular	454	3,26%
8	Kimya	429	3,08%
9	Bioteknoloji Uygulamalı Mikrobiyoloji	406	2,92%
10	Medikal Enformatik	342	2,46%
11	Otomasyon kontrol sistemleri	328	2,36%
12	Çevre Bilimleri Ekolojisi	315	2,26%
13	Farmakoloji	308	2,21%
14	Genetik	266	1,91%
15	Enformasyon ve kütüphane bilimleri	259	1,86%
16	Telekomünikasyon	232	1,67%
17	Tarım	186	1,34%
18	Fizik	214	1,54%
19	Sağlık hizmetleri ve bilimleri	186	1,34%
20	Jeoloji	178	1,28%

(*Bir makale birden fazla kategoride yer alabildiğinden yüzde sütununda yer alan yüzdelerin toplamları 100 değildir).

4. KULLANILAN YÖNTEMLER

Veri madenciliği; çok çeşitli algoritmaların, yapay zekâ uygulamalarının, matematik ve istatistik bilgisinin ve tüm bu yöntemlerin çeşitli birleşimlerinin kullanıldığı bir alandır. İndeksli yayınların tam metinlerinde yaklaşık 300'e yakın bilinen VM yöntemi ismi aranmış,

100 den fazla makaleye konu olan yöntemler Tablo 6'da sunulmuştur. Son 10 yılın en popüler VM yöntemi "Sinir Ağları" olmuştur. En çok kullanılan keşif yöntemleri açısından "Sinir ağlarını" sırasıyla; "İlişkilendirme kuralları" ve "k-ortalamalar" izlemiştir.

Tablo 6: Son 10 yılda yayınlanan makalelerde en çok kullanılan keşif yöntemleri

Yöntem	Arama Terimi	Yayın
Sinir ağları	Neural networks	1573
İlişkilendirme kuralları	Association Rule	1536
K-Ortalamalar	K-means	933
Karar ağaçları	Decision trees	908
Destek vektör makinaları	Support vector machines	824
Lojistik regresyon	Logistic regression	514
Genetik algoritma	Genetic algorithms	425
Temel bileşenler analizi	Principal component analysis	422
Yaygın kalıplar	Frequent pattern	419
Apriori algoritması	Apriori algorithm	406
Naive Bayes	Naive Bayes	401
C4.5	C4.5	363
Yaygın nitelikler	Frequent item set	324
Hiyerarşik kümeleme	Hierarchical clustering	262
CART	CART	249
Sıralı kalıp madenciliği	Sequential pattern mining	240
K-en yakın komşu	K-nearest neighbor	211
Regresyon analizi	Regression analysis	186
Regresyon ağacı	Regression tree	171
FP-büyüme algoritması	FP-growth algorithm	152
FP ağaç	FP-Tree	147
Vaka tabanlı gerekçelendirme	Case based reasoning	145
HITS	HITS	131
Faktör Analizi	Factor analysis	126
Matris çarpanlarına ayırma	Matrix factorization	115
DBSCAN	DBSCAN	111

5. KULLANILAN YAZILIMLAR

Veri Madenciliği uygulamalarını gerçekleştirmek için çeşitli yazılımlardan faydalanmak mümkündür. Bu kapsamda, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB ticari ve RapidMiner (YALE), WEKA, R, C4.5, Orange, KNIME açık kaynak olmak üzere birçok yazılım geliştirilmiştir (Dener, Dörterler, & Orman, 2009). İndeksli yayınların tam metinlerinde yaklaşık 40'a yakın bilinen VM yazılımı ismi taranmış, ilk 5'e giren yazılımlar Tablo 7'de sunulmuştur. Yazılım kullanan araştırmacıların ise en yaygın tercihi "WEKA" isimli yazılım olmuştur. Waikato Üniversitesi tarafından java

platformu üzerinde açık kaynak kodlu olarak geliştirilen ve devamlı güncellenen bir yazılımdır (Tekerek, 2011). Bu yazılımın kullanımına yönelik çok sayıda ücretsiz kaynağa ulaşmak mümkündür. Matlab'ın yaygın kullanımı; yapay sinir ağları için kullanıcılara sunulan kolay kullanımlı araç paketinin (toolbox) olmasıyla ilişkilendirmek yanlış olmayacaktır. Mevzu bahis bu yazılımlar daha çok bilinen/mevcut yöntemleri uygulamak için kullanıldığından; araştırmacılar çoğunlukla kendi kodlarını yazmayı tercih edebilmektedir.

Tablo 7: Son 10 yılda yayınlanan makalelerde en çok kullanılan veri madenciliği yazılımları

Sıra	Yazılım	Yayın Sayısı
1	WEKA	286
2	Matlab	97
3	SAS	73
4	R	64
5	SPSS	63

6. TÜRKİYE'DE VERİ MADENCİLİĞİ ÇALIŞMALARI

Türkiye'yi adres gösteren 249 VM makalesinin, %50'sinden fazlası (137 âdeti) alanda en çok yayını olan 50 araştırmacı tarafından yapılmıştır. Türkiye adresli makaleler yıllık ortalama 0.72 atıf almıştır.

Bu makalelerin en fazla yayınlandığı dergi (34'ü) ESWA olmuştur. Alanın en yüksek UBYT puanına sahip (100 puan) dergisi Bioinformatics Dergisi'nde Türkiye adresli makale yer almamıştır. İkinci en yüksek UBYT puanına sahip dergisi "Data Mining and Knowledge Discovery" de ise Türkiye adresli 3 makale bulunmaktadır.

Bu alanda en fazla yayına sahip yazarın (Dr. Lale Özbakır, Erciyes Üniversitesi) 12 makalesi olduğu görülmektedir. 12 makale, Türkiye adresli VM yayınlarının yaklaşık %4'üne tekabül etmektedir. Türkiye'yi adres gösteren ve VM alanında 2. en fazla yayın yapan yazar 10 makaleye sahiptir (Dr. Yücel Saygın, Sabancı Üniversitesi).

Çalışma kapsamında ele alınan Türkiye adresli yayınlardan en fazla atıf alan çalışma (250 atıf) 2007 yılında; "Data & Knowledge Engineering" dergisinde yayınlanan "ST-DBSCAN: An algorithm for clustering spatial-temp oral data (Birant & Kut, 2007)" (Derya BİRANT ve Alp KUT tarafından yazılmıştır) olmuştur.

7. SONUÇLAR VE ÖNERİLER

Veri Madenciliği (VM), bu çalışmada da gösterilmeye çalışıldığı üzere; akademik alanın yıl geçtikçe daha fazla ilgilendiği bir çalışma alanı olmuştur. Türkiye adresli yayınlar ele alındığında; ülkeler arası sıralamamızın tüm alanlara kıyasla daha iyi durumda olduğu görülmesine karşın, VM alanında gerçekleştirilen çalışmaların geniş bir kitle tarafından değil, az sayıdaki akademisyenler tarafından hazırlandığı anlaşılmaktadır. Beri yandan; alanın en yüksek etki skoruna sahip dergilerinde Türkiye adresli yayınların oldukça sınırlı olduğu anlaşılmaktadır. Türkiye adresli yayınların ise daha çok mühendislik kökenli araştırmacılarca hazırlandığı ve diğer alanların katkılarının sınırlı olduğu görülmüştür. Çok disiplinli bir çalışma alanı olan VM; farklı bilimsel disiplinlerden araştırmacıların bir araya gelmesi ve uygulamaya dönük yeni çalışmalar yapılmasıyla önemli bir ivme kazanabilecektir. Türkiye nüfusunun büyüklüğü ve son yıllarda veri depolama kapasitesindeki artışa bağlı olarak birçok yeni VM uygulaması potansiyel oluşturduğu ifade edilebilir. Waikato Üniversitesi tarafından

geliştirilen açık kaynak kodlu WEKA isimli yazılımın, eğitimleri yaygınlaştırılarak, bu yazılımın ücretsiz olmasından faydalanabilir.

Bu çalışmada tespit edilen hususların, hem araştırmacılara hem de uygulayıcılara ileriki çalışmalarda/uygulamalarda yol gösterici olması amaçlanmaktadır.

KAYNAKÇA

Asan, A. (2013). Türk Dergilerinin Web of Science'teki Yeri, İmpakt Faktör (Etki Faktörü) ve h indeks. 16. Ulusal Halk Sağlığı Kongresi, 61–79.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221.

Chen, D., Liu, Z., Luo, Z., Webber, M., & Chen, J. (2016). Bibliometric and visualized analysis of emergy research. *Ecological Engineering*, *90*, 285–293.

Chen, H.-Q., Wang, X., He, L., Chen, P., Wan, Y., Yang, L., & Jiang, S. (2016). Chinese energy and fuels research priorities and trend: A bibliometric analysis. *Renewable and Sustainable Energy Reviews*, *58*, 966–975.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*(8), 981–1012.

Data Mining and Knowledge Discovery, Springer, 2016, http://link.springer.com/journal/10618

Dener, M., Dörterler, M., & Orman, A. (2009). Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama. XI. Akademik Bilişim Konferansı Bildirileri, 787-96.

Durmuşoğlu, A. (2016). A pre-assessment of past research on the topic of environmental-friendly electronics. *Journal of Cleaner Production*, *129*, 305–314.

Durmuşoğlu, A., & Dereli, T. (2013). Veri madenciliği alanında gerçekleştirilen çalışmalar üzerine bir inceleme ÜAS 2013: Üretim Araştırmaları Sempozyumu, 389–395.

Fayyad, U. M. (1997). Editorial. *Data Mining and Knowledge Discovery*, *1*(1), 5–10.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, *39*(11), 27–34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press.

Forsman, M. (2015). Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, Blaise Cronin, Cassidy R. Sugimoto (Eds.). MIT Press, Cambridge, MA (2014). *Library & Information Science Research*, *37*(4), 383.

Garousi, V., & Mäntylä, M. V. (2016). Citations, research topics and active countries in software engineering: A bibliometrics study. *Computer Science Review*, *19*, 56–77.

Glymour, C., Madigan, D., Pregibon, D., & Smyth, P. (1997). Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery*, *1*(1), 11–28.

Merigó, J. M., Mas-Tur, A., Roig-Tierno, N., & Ribeiro-Soriano, D. (2015). A bibliometric overview of the Journal of Business Research between 1973 and 2014. *Journal of Business Research*, *68*(12), 2645–2653.

MHM, B. N., & PhD, J. B. B. M. (2000). A Pilot Study of Motivations, Inhibitors, and Facilitators of Association Members in Attending International Conferences. *Journal of Convention & Exhibition Management*, *2*(2–3), 97–111.

Perner, P. (Ed.). (2010). *Advances in Data Mining. Applications and Theoretical Aspects* (Vol. 6171). Berlin, Heidelberg: Springer Berlin Heidelberg.

Polat, M., & Karahan, A. G. (2009). Multidisipliner Yeni Bir Bilim Dalı: Biyoinformatik ve Tıpta Uygulamaları. *SDÜ Tıp Fakültesi Dergisi*, *16*(3).

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri Madenciliği ve Türkiyede'ki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *11*(21), 1–23.

Tekerek, A. (2011). Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları. XIII. Akademik Bilişim Konferansı Bildirileri, 161-169.

Thomson Reuters, New York, NY, ABD. Nisan 20, 2016, http://apps.webofknowledge.com/

TÜBİTAK-ULAKBİM. UBYT Programı Uygulama Esasları. http://cabim.ulakbim.gov.tr/wp-content/uploads/sites/4/2015/09/2016-Y%C4%B1l%C4%B1-UBYT-Program%C4%B1-Uygulama-Esaslar%C4%B1.pdf

University of Granada,. (n.d.). SCImago Journal & Country Rank [SCImago Journal & Country Rank]. http://www.scimagojr.com/countryrank.php

Vinkler, P. (2013). Would it be possible to increase the Hirsch-index, π-index or CDS-index by increasing the number of publications or citations only by unity? *Journal of Informetrics*, *7*(1), 72–83.