ISSN: 2602-3563

ACTA INFOLOGICA

(ACIN)

JUNE, 2018

Volume: 2 | Issue: 1

ISTANBUL UNIVERSITY, INFORMATICS DEPARTMENT

dergipark.gov.tr/acin

informatics.istanbul.edu.tr

Acta INFOLOGICA (ACIN) | ISSN: 2602-3563
Haziran (June) 2018
Cilt (Volume): 2 | Sayı (Issue): 1

Dergi Sahibi
(Owner)
İstanbul Üniversitesi Enformatik Bölümü
adına Dr. Sevinç Gülseçen

Sorumlu Yazı İşleri Müdürü
(Managing Editor)
Dr. Sevinç Gülseçen

Editörler
(Editors)
Dr. Sevinç Gülseçen
Baş Editör (Editor in Chief)
Dr. Çiğdem Erol
Dr. Serra Çelik
Dr. Emre Akadal
Dr. Fatma Önay Koçoğlu

İletişim
(Contact)
İstanbul Üniversitesi Enformatik Bölümü
Kalenderhane Mah. 16 Mart Şehitleri Cad. No: 8
Vezneciler Fatih İstanbul Türkiye
+90 212 440 00 00 External: 10037
enformatikdergi@istanbul.edu.tr

Yayın Dili
(Publication Language)
Türkçe (Turkish)
İngilizce (English)

Yayın Sıklığı
(Publication Period)
Yılda 2 sayı (Haziran ve Aralık)
Biannual (June and December)

EDİTÖRDEN

Acta Infologica, Enformatik disiplininin diğer tüm disiplinlerle olan arakesitindeki uygulamaları, başka bir ifade ile Enformatiğin kullanımı ve uygulanması ile ortaya çıkan sonuçları ve Enformatiğin temel teorileri ile ilgili çalışmaları kapsayan yayınları araştırmacılarla ve okuyucularla paylaşmak amacı ile Haziran 2017'de yayın hayatına başlamıştır. O günden beri dergimizin yenilenme çalışmaları devam etmekte olup kurullarımız güncellenmiş, bu sayıdan sonra TR dizine başvuru yapılması planlanmış ve ulusal ile uluslararası endekslerde taranma gibi çok önemli bir hedef konmuştur.

Bilgi teknolojilerinin hemen her alanda ve gündelik yaşamda yaygın bir şekilde kullanımı, bu kullanım esnasında ve sonucunda gelişen olaylarda insanın alacağı tavır ve duruş, Enformatik disiplininin araştırma eksenlerini oluşturmaktadır. Bu iki eksende yerini alabilecek tüm çalışmalara dergimiz açıktır. Bugüne kadar değerli makaleleriyle katkı sağlamış yazarlarımıza ve değerlendirme süreçlerinde bizlere yardımcı olan hakemlerimize içtenlikle teşekkür ederiz.

Bir derginin yayın hayatına başlaması gerekli bir adım olmakla birlikte yeterli değildir, önemli olan derginin sürdürülebilir olmasıdır. Bu zorlu süreçte, dergi yayın ekibine, editörler kuruluna ve hakemlerimize çok iş düşmektedir. Bu vesile ile dergi yayın ekibini oluşturan Dr. Çiğdem Erol, Dr. Serra Çelik, Dr. Emre Akadal, Dr. Fatma Önay Koçoğlu'na, editörler kurulunda yer alan sevgili araştırmacılarımıza ve değerli hakemlerimize minnettar olduğumuzu ifade etmek isterim.

"Bilge" dergimizin başarılı bir yayın hayatı olsun!

Acta INFOLOGICA Yayın Ekibi Adına

Dr. Sevinç Gülseçen

CONTENT

Research Articles

*** Corresponding Author / Sorumlu Yazar**

ACTA INFOLOGICA, 2018; 2(1): 1-11
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

Bir Güvenlik Trendi: Bal Küpü

Süleyman Muhammed Arıkan*, Recep Benzer

ÖZ

Bu çalışma, bilişim dünyasında güvenliği sağlamak adına kullanılmakta olan ve her geçen gün popülerliği artan bal küpleri hakkında bir araştırma niteliğindedir. Honeypot olarak kullanılan bu terim dilimize bal küpü olarak geçmiştir. Balküpleri, izinsiz yapılmış girişlerin veya zararlı aktivitelerin tespitinde kullanılmasına ek olarak asıl amaçları sadırganın veya zararlı aktivitenin kullanmış olduğu metod ve araçlar hakkında bilgi edinmektir. Bal küplerinin birçok çeşidi vardır. Çalışmamızda bu çeşitleri açıklamaya odaklanılmıştır. Dinamik bal küpü dizaynı, bal küpü sistemleri için saldırgan web uygulaması, ssh ataklarının analizi ve görselleştirilmesi gibi çalışmalar incelenmiştir. Ayrıca Glastopf ile sql enjeksiyonu saldırısına ve yapılan saldırının kayıtlardan görüntülenmesine örnek bir uygulama yapılmıştır. Gelişen modern dünyada her geçen gün farklı alanlar oluşmakta ve gereksinimler değişmektedir. Bu gereksinimlere bal küpü sistemlerinin kendini adapte etmesi her ne kadar hızlıda olsa bazı konularda eksikleri devam etmektedir. Ancak esnek yapıları ve herhangi bir sisteme kolay adapte olmaları sayesinde yeni moduller geliştirilerek eksikliklerin büyük kısmı kullanıcılar tarafından giderilebilmektedir.

Anahtar Kelimeler: Bal küpü, malware, sql injection, ransomware, worm.

New Security Trend: Honeypot

ABSTRACT

This study is a research about the honeypots which is being used to provide security in the world of information and increasing in popularity day by day. This term used as a honeypot is passed down to our slice as honey cube. In addition to being used in the detection of unauthorized entries or harmful activities, it is also necessary to obtain information about the methods and means by which the original purpose is to use the malicious or harmful activity. There are many kinds of honeypots. Our work focuses on explaining these species. Dynamic honeypot design, aggressive web application for honeypot systems, analysis and visualization of ssh attacks have been studied. In addition, Glastopf, sql injection attack and the attack is displayed on the records of a sample application has been done. In the developing modern world, different fields are formed every day and requirements are changing. Although these requirements are fast for the honeypot systems to adapt themselves but some are still missing. However, due to their flexible structure and easy adaptation to any system, new modules can be developed and most of the deficiencies can be eliminated by the users.

Keywords: Honeypot, malware, sql injection, ransomware, worm.

Information of Author(s):

Süleyman Muhammed Arıkan
ORCID: 0000-0003-1526-2970
suleyman.arikan@tubitak.gov.tr
Gazi University, Information Institute

Recep Benzer
ORCID: 0000-0002-5339-0554
rbenzer@kho.edu.tr
National Defense University, KHO

DOI: 10.30801/acin.356815

Submit Date:	21.11.2017
Accept Date:	02.05.2018
Publish Date:	26.06.2018

(*) Contact Author
Address: National defense University, Department of Computer Engineering, Ankara, Turkey
Telephone Number: +90 419 75 58 /5453

ACTA INFOLOGICA
ISSN: 2602-3563

1. GİRİŞ

Günümüz modern dünyasında güvenliği sağlamak amacıyla birçok teknik ve yöntem kullanılmaktadır. Her geçen gün inanılmaz bir hızla arttığı görülen saldırı tekniklerine nazaran savunma teknikleri için aynı şeyi söylemek oldukça güçtür. Fakat her geçen gün saldırılar hakkında yeni teknikler keşfedilmekte ve karşı önlemler alınmaya çalışılmaktadır. Tüm bunlara ek olarak sıfırıncı gün atakları (zero-day attacks) ile tanımlanan ataklar günlerce hatta yıllarca gizemini korumaktadırlar. Keşfedilmemiş bu atakları, modern çözümlemelerin birçoğu engelleyememekte hatta birçok durumda saldırıya uğradığını dahi fark edememektedirler. Çalışmamızın devam eden kısmında günümüz güvenlik tekniklerine yüzeysel olarak bir bakış atılacaktır.

İstemci tarafında güvenliği sağlamak adına en çok kullanılan teknolojilerin başında şüphesizki antivirüs yazılımları gelir. Antivirüsler, sistem üzerindeki zararlı yazılımları tespit edip amaçlanan işlemin gerçekleşmesini engelleyerek zararlı yazılımı sistemden kaldırmayı hedefleyen programlardır. Modern antivirüsler; solucanlar (worm), truva atları (trojan horses), arka kapılar (backdoor), fidye yazılımlar (ransomware), tarayıcı ele geçirme (browser hijacking), casus yazılım (spayware), çevrimiçi bankacılık atakları ve sosyal mühendislik gibi birçok saldırıya karşı kullanıcıyı ve sistemi koruma yetenekleri geliştirmiştir (Anonymous, 2017). İstemci taraflı saldırılarda son kullanıcıyı korumada antivirüslerin faydası göz ardı edilemez. Fakat tüm saldırılar yalnızca son kullanıcıya yönelik değildir. Ev ortamında antivirüs kullanımı yeterli bir güvenlik önlemi olarak sayılabilmesine karşın kurumsal iş ortamlarında tek başına bir yeterliliği olduğunu söylemek güçtür. Ayrıca, antivirüslerin son kullanıcı nezninde sistem kaynaklarını fazla kullanma gibi bazı olumsuzlukların olduğuda ortadadır.

Güvenliği sağlamak adına kullanılan bir diğer teknoloji olarak ise güvenlik duvarı (firewall) sayılabilir. Güvenlik duvarı olmayan bir iş veya kurumsal ağ bulmak neredeyse imkansızdır. Güvenlik duvarı, paketler üzerinde incelemeler yaparak filtrelemeye olanak sağlar (Yang ve ark., 2011). Filtrelemeler kural tabanlıdır. Kurala uyan / uymayan paketler düşürülür. Güvenlik duvarlarını yapılarına göre yazılımsal ve donanımsal olarak ikiye ayırabiliriz. Donanım tabanlı güvenlik duvarları oldukça performanslı çalışır. Üzerindeki donanım, barındırdığı yazılıma özgü hazırlandığından yetenekleri haliyle daha da gelişmiştir. Donanımsal güvenlik duvarlarında hedef, ağların birbirinden izole edilmesidir. Kurulumları ve konfigürasyonları kolay değildir. Ayrıca satış ücretleri oldukça yüksektir. Yazılımsal güvenlik duvarlarında ise maliyet düşüktür. Zaten hali hazırda, işletim sistemleri ve antivirüslerin büyük bir çoğunluğu içerisinde güvenlik duvarı barındırmaktadır. Farklı bir yazılım ile koruma amaçlanırsa, internet üzerinden ücretsiz güvenlik duvarı edinebilmenin yanı sıra cüzi miktarlar ile lisanslı yazılımsal güvenlik duvarları da satın alınabilir. Yazılımsal güvenlik duvarları, bilgisayar ile ağ arasında gerçekleşen trafiğin filtrelenmesine olanak sağlar. Her türlü sisteme uyum sağlayabilirler. Yazılımsal olduklarından dolayı sistem kaynaklarını kullanılar. Bu yüzden çoğu zaman performansları düşüktür.

Ağ içerisinde gerçekleşen atakların tespiti ve engellenmesi için IDS ve IPS araçları kullanılabilir. Ağ trafiğinin izlenerek yetkisiz giriş ve aktivitelerin tespitinde, dilimize saldırı tespit sistemi olarak çevrilmiş IDS (Intrusion Detection Systems) kullanılır. Ağ trafiğini inceleyerek saldırıları engellemek adına ise IPS (Intrusion Prevention Systems) kullanılır. Dilimize saldırı engelleme sistemi olarak çevrilmiştir. IDS ve IPS arasındaki fark oldukça basittir. IDS, saldırıları tespit ederek ağ yöneticisine durumu raporlar ve bir alarm oluşturur. IPS ise, aynı IDS gibi çalışmasına ek olarak aynı zamanda bu saldırıyı engellemeye çalışır. ID/PS araçları iki basit tipe ayrılır: Anasistem tabanlı (Host-based) ve Ağ tabanlı (Network-based). Anasistem tabanlı ID/PS araçları, bireysel bilgisayarlar üzerindeki verileri inceler. İçeriden kaynaklı kötüye kullanımların tespitinde ve engellemesinde oldukça verimlidir. Anasistem tabanlı IDS araçlarına Windows NT/2000 Security Event Logs ve UNIX Syslog örnek olarak gösterilebilirken anasistem tabanlı IPS araçlarına Cisco Security Agent (CSA) örneği verilebilir (Kuwatly ve ark., 2004). Diğer taraftan ağ tabanlı ID/PS araçları, mevcut ağ üzerinde taşınan tüm paketleri analiz eder. Çoğu zaman paketler deneysel bir veri ile kıyaslanarak incelenir. Ağ tabanlı IDS araçlarına açık kaynak kodlu Snort örnek olarak verilebilir. Snort, gerçek zamanlı ağ tarafiği üzerinde paketleri inceleyerek arabellek taşması (buffer overflow) ve gizli port taraması (stealth port scan) gibi daha birçok geniş çeşitlilikteki atakları tespit eder. Ayrıca Snort ‘a ek olarak Suricata, Bro Ids, OpenWPS-ng ve Security Onion gibi yine

ISSN: 2602-3563

ücretsiz araçlar örnek gösterilebilir. Ağ tabanlı IPS araçlarına ise McAfee's Network Protection System örneği verilebilir.

2. YÖNTEM

Bal küpü terimi için esin kaynağı, ayılar için kurulan tuzaklara yerleştirilen içi bal dolu kaplardır (Zakaria ve Kiah, 2012). Bal küpleri ile ayıların tuzağa gelmesi sağlanmaktaydı. Günümüz bilişim dünyasında ise bal küpleri, saldırganları üzerine çekmek için kullanılmaktadırlar. Bal küpü sistemleri zararlı aktiviteyi ya da saldırganı durdurmakta aktif görev almazlar. Balküpleri, izinsiz yapılmış girişlerin veya zararlı aktivitelerin tespitinde kullanılmasına ek olarak asıl amaçları sadırganın veya zararlı aktivitenin kullanmış olduğu metod ve araçlar hakkında bilgi edinmektir (Song ve ark., 2012). Edinilen bu bilgiler ışığında istenmeyen faaliyetlerin engellenmesi adına çalışmalar yapılır. Dolayısı ile bal küpü sistemlerini, saldırganın veya zararlı aktivitenin kullandığı teknikler ve araçlar hakkında çalışmamıza olanak sağlayan platformlar olarak tanımlayabiliriz (Chauhan ve Shiwani, 2017).

Bal küpü sistemlerine yönlendirilmiş hiçbir trafik yoktur. Buna ek olarak bal küpü sistemleri herhangi bir üretim değeri bulundurmaz. Dolayısı ile balküpleri üzerinde hiçbir trafik görülmemelidir. Eğer bal küpü sistemleri ile bir etkileşim söz konusu ise, etkileşime geçen kaynak yüksek olasılıkla saldırgan ya da zararlı aktivitedir.
Bal küpü sistemleri bir ağa kurulduğunda, ağdaki diğer makineler gibi gözükmesine dikkat edilmelidir. Bulunduğu ağın özelliklerine göre servisler seçilmelidir. Bilinen açıklıklar bırakılarak saldırganların dikkatleri bal küpü üzerine çekilerek gerçek sistem ve servisler korunur. Birçok saldırı tespit ve engelleme sisteminin aksine yanlış alarm (false positive) oranı oldukça düşüktür. Bal küpü sistemleri kurulum için büyük sistem kaynaklarına gerek duymaz.

İlk bal küpü yazılımı 1998 yılında yazılmış Cybercop Sting adındaki yazılımdır. Aynı zamanda Decoy Server olarak da bilinmektedir. Decoy Server, Telnet ve SMTP gibi servisleri simüle edebilmekteydi. Oldukça limitli bir kullanıma ve kayıtlama yeteneğine sahip olsa da saldırgan ataklarını analiz etmede oldukça faydalıydı (Sharma, 2016).

Günümüz balküpleri etkileşim seviyelerine göre Düşük etkileşimli (Low-Interaction) ve Yüksek Etkileşimli (High-Interaction) olarak ikiye ayrılır (Singh ve Joshi, 2011). Buradaki etkişim ile kastedilmek istenilen, aktivite düzeyi veya faaliyet alanıdır.

Düşük etkilişimli bal küpü sistemlerinin faaliyet alanı oldukça kısıtlıdır. Bu kısıtlılık taklitten kaynaklanır. Çünkü düşük etkileşimli bal küpü sistemlerinde servisler ve servislerin üzerinde çalıştığı işletim sistemi tamamen simüle edilmektedir. Bu simülasyondan dolayı saldırganın bal küpü sistemi üzerinde gerçekleştirebileceği işlemler limitlidir. Bir düşük etkileşimli bal küpü sisteminde ftp protokolü simüle edilmek istendiğini düşünelim. Bu durumda 21. port dinleniyormuşcasına cevaplar döndürülür. Bu portta login komutlarının çalışmasına olanak sağlanır. Buna ek olarak çeşitli ftp komutlarıda çalıştırılır. Saldırgan veya zararlı aktivite burada ftp servisinin çalıştığını düşünür. Fakat yapılan tüm işlemler taklitten ibarettir. Bunlara ek olarak; saldırganın, taklit servisin desteklemediği bir komutu çalıştırması sonucu sistemin bir bal küpü olduğu anlaşılabilir. Düşük etkileşimli balküplerinden sınırlı bilgi edinilir. Saldırganın ve zararlı aktivitenin gerçek bir sistemde yapacağı tüm işlemleri, taklit sistemde uygulayamayacağı ihtimalinden dolayı her konuda bilgi edinilemeyebilir. Buna karşın kurulumları ve bakımları oldukça basittir. İçerisinde barındırdıkları sistemler taklit olduklarından dolayı ağ için herhangi bir risk oluşturmazlar. Bilinen aktivitelerin yakalanmasında oldukça etkilidirler. En çok kullanılan düşük etkileşimli bal küpü sistemleri olarak Honeyd, Specter, KFSensor ve Dionaea örnek verilebilir. KFSensor, bir windows tabanlı bal küpü sistemidir. Servisleri windows işletim sistemi üzerinde çalışıyormuş gibi taklit eder. Servisler OSI katmanlarından uygulama katmanında taklit edilir. Dolayısı ile çoğunlukla yeni güvenlik duvarı kuralları oluşturmak için veya yeni IDS imzaları yazmak için kullanılmaktadır (Singh ve Joshi, 2011). Honeyd en çok kullanılan düşük etkileşimli bal küpü sistemlerinin başında gelir. Bir ağda sanal sistemler oluşturan uygulamadır. Oluşturulan sanal sistemler uzaktan kontrol edilebilir. Konfigürasyonlarına göre işletim sistemi ve servisler taklit edilir. İşletim sistemi ve servis çeşitliliği geniştir. Ayrıca tek bir sisteme birden fazla ip adresi atanmasının mümkün olması önemli bir özelliktir. GNU

ISSN: 2602-3563

lisansı altında açık kaynak kodlu bir yazılımdır. Specter, honeyd gibi herhangi bir işletim sistemini üzerinde belirlirlediğimiz servisler çalışıyormuşcasına simüle eder. Honeyd'den ayıran en büyük özelliği, içerisinde tuzak uygulamalar barındırmasıdır. Bu sayede saldırgana ait bilgiler edinmeye çalışmaktadır.

Yüksek etkileşimli bal küpü sistemleri, düşük etkileşimli bal küpü sistemlerine nazaran daha karmaşık yapıya sahiptirler. Kurulumları ve bakımları zaman almakla birlikte zor olduğu söylenebilir. Gerçek işletim sistemi ve uygulamalar kullanılır. Dolayısı ile saldırgan veya zararlı aktivite hakkında çok geniş çapta bilgiler edinilebilir. Sıfırıncı gün ataklarının tespitinde veya saldırganın kullandığı teknikler üzerinde çalışmak için idealdirler. Symantec Decoy Server ve Sebek bu balküplerine örnek olarak gösterilebilir (Djanali ve ark., 2014). Sebek, iki modül barındırır. Birinci modül yüksek etkileşimli bal küpü üzerinde çalışarak büyük çapta kayıt toplar. Toplanan bu kayıtlar ikinci modül olan sunucuya gönderilir. Bu sunucu genellikle bal küpü sistemlerinin önünde duran ağ geçididir. Fakat bağımsız olarakta kullanılabilir. Sunucu sayesinde tüm kayıtlar merkezi bir noktada toplanır. Sebek sunucu modülü ve bal küpü modülü aynı makine üzerine kurulamamaktadır.

Yüksek etkileşimli bal küpü sistemlerine birçok kaynakta ayrıca honeynet örneğine rastlamak mümkündür. Honeynet kavramı, iki ya da daha fazla bal küpü sistemlerinin bir arada kullanılmasına verilen addır (Zakaria ve Kiah, 2012).

3. BULGULAR

Bal küpü istemleri ile birçok konuda örnek çalışmalar yapılmıştır. Makalemizde bu çalışmalardan bazılarına yer verilmiştir. İlgili çalışmalar incelenmiş ve yararlı olacağı düşünülen araştırmalar okuyucuya aktarılmıştır.

Dinamik bal küpü dizaynı (Kuwatly ve ark., 2004), bal küpü sistemleri için saldırgan web uygulaması (Djanali ve ark., 2014), ssh ataklarının analizi ve görselleştirilmesi (Koniaris ve ark., 2013) gibi çalışmalar incelenerek okuyucuya aktarılması hedeflenmiştir. Bu bölümde ayrıca Glastopf ile sql enjeksiyonu saldırısına ve yapılan saldırının kayıtlardan görüntülenmesine örnek oluşturmak adına tarafımca yapılmış uygulama okuyucuya sunulacaktır.

3.1. Dinamik Bal Küpü Dizaynı

Bal küpü sistemleri; genellikle yönetici tarafından atanmış ip adreslerinde oluşturulmaktadır. Statik atamanın aksine, dinamik bal küpü ile kullanılmayan ip adresleri tespit edilerek bu iplerde sanal balküpleri oluşturulması hedeflenmiştir. Bu sayede saldırganların gözden kaçması minimuma indirilecektir. Dinamik bal küpü sistemi aynı zamanda, ip kullanımına göre kendini güncelleyecektir. Bir ip boşa çıktığında ilgili adrese bir sanal bal küpü kurulacak, ip kullanılmaya başladığında ise ilgili bal küpü sistemden kaldırılacaktır.

Dinamik bal küpü yaklaşımının uygulanması için şu bileşenlere ihtiyaç duyulmaktadır: Aktif tarama aracı, pasif tarama aracı, düşük etkileşimli bal küpü, yüksek etkileşimli fiziksel bal küpü, veri tabanı, dinamik bal küpü motoru.

Aktif tarama, kullanılan işletim sistemleri ve portları bulmak amacıyla kullanılacaktır. Nmap uygulaması seçilmiştir. Nmap içerisinde işletim sistemi ve servislerin imzalarını barındırmaktadır. Hedef makineye paketler göndererek cevap ister. Aldığı cevapları veritabanındaki imzalarla karşılaştırarak, hedef sistemin ne olduğuna karar verir.

Pasif tarama, ağdaki paketlerin dinlenemesi ile hedef sistemin bilgilerini bulma işlemine dayanır. Aktif tarama ile pasif tarama arasındaki fark, pasif taramada aktif taramada olduğu gibi hedef ile etkileşimin olmamasıdır. Pasif tarama araçlarına örnek olarak p0f ve Snort gösterilebilir.

p0f, ağda yakaladığı SYN paketlerini içinde barındırdığı veri tabanındaki imzalarla karşılaştırarak işletim sistemine karar verir. Bu karar için paketlerin içerisinde bulunan TCP/IP alanlarına bakar. Tamamıyla sessizdir.

ACTA INFOLOGICA
ISSN: 2602-3563

Pasif tarama ağda çok az bir trafiğe sebep olur. Buna nazaran aktif tarama ağ içerisinde çok fazla faaliyete sebebiyet verir. Aktif taramada sonuçlar birer anlık fotoğraf gibidir. Pasif taramada ise sonuçlar gerçek zamanlı olarak sürekli toplanmaktadır. Aktif tarama tüm açık port ve servisleri bulabilmesine karşın, pasif tarama yalnızca trafik oluşturan portları bulabilir.

Düşük etkileşimli bal küpü olarak ise honeyd kullanılmıştır. Çünkü honeyd bal küpü sistemi hem uzaktan oluşturulabilmekte hemde üzerinde konfigürasyon değişiklikleri yine uzaktan yapılabilmektedir. Honeyd balküpleri, oluşturulduktan sonra bir etkileşimle karşılaştıklarında bu trafiği doğruca fiziksel yüksek etkileşimli bal küpü sistemi olan Sebek bal küpüne yönlendirir. Böylece saldırgan yüksek etkileşimli bal küpü sistemi ile karşılaşarak hakkında daha fazla bilgi toplanmasına olanak sağlar.

Yüksek etkileşimli bal küpü sistemi olarak Sebek kullanılmıştır. Sebek bal küpü modülleri, honeyd balküplerinden yönelendirilen trafik ile girdiği etkileşimlerin kaydını tutarak Sebek sunucusuna tutulan kayıtları yollar. Sunucuda merkezi olarak tutulan bu kayıtlar, belirli periyotlarla sunucu tarafından incelenerek risk farkedildiği takdirde SMS ile yöneticiyi bilgilendirmektedir.

Veri tabanı içerisinde; IPInformaiton, Ports, OperatingSystems, ServicesScripts, HoneydLogs ve SebekLogs olmak üzere altı tablo vardır. IPInformation tablosu içerisinde belirli ip adresine ait; işletim sistemi bilgisi, verinin edinildiği zaman ve ip adresinin gerçek bir sistem veya sanal bir sistem tarafından kullanıldığına dair kayıtları içerir. Ayrıca bu bilgilerin hangi tarama aracı ile edinildiği bilgiside içerilmektedir. Çünkü tarama araçları bazı durumlarda diğerlerine göre daha güvenilir sonuçlar vermektedir. Ports tablosu içerisinde belirli ip adresine ait sistemdeki açık tüm portlar ve bu portlarda çalışan servisler tutulur. Yine bu tablodada veriye hangi tarama aracı ile ulaşıldığının yanı sıra zaman bilgiside bulunmaktadır. OperatingSystems tablosunda her işletim sistemine benzersiz bir numara verilerek açıklaması yazılmıştır. ServicesScripts tablosunda ise honeyd tarafından simüle edilen servisler ve bu servislerin simüle edilmesi için gerekli senaryolar tutulmaktadır. HoneydLog tablosunda honeyd tarafından tutulmuş kayıtlar bulunmaktadır. SebekLogs tablosunda ise Sebek sunucusunda toplanmış kayıtlar tutulmaktadır.

Dinamik bal küpü sistemi motor aracının şu işlemlerde kullanılması amaçlanmıştır: yerleştirilecek sanal balküpü sayısının hesaplanması, konfigürasyon komutlarının gönderilmesi, kayıtların toplanması ve çıktı oluşturulması ile hesaplamalarda yanlışlık oluşmasını engellemek adına veri tabanında eski kayıtların silinerek güncel kalmasını sağlamak.

Sonuç olarak, kullanılmayan ip adreslerine düşük etkileşimli bal küpü sistemleri kurularak saldırganın ağda yaptığı tüm hareketlerin kayıt altına alınması hedeflenmiştir. Honeyd ile etkileşime geçen her trafik Sebek bal küpüne yönlendirilmiştir. Sebek bal küpü ise kayıtları merkezi sunucuda toplayarak belli zamanlarda bu kayıtları incelemiş ve yüksek risk gördüğünde durumu yöneticiye bildirmiştir. Bu sayede izinsiz giriş yapan saldırgan veya ağda bulunan zararlı aktiviteler kısa sürede farkedilmiştir. Sanal balküplerinin ağdaki makinelerin karakteristik özellikleri ile aynı oranda kurulması ile sistemdeki varlıkları gizlenmiştir. Bu sayede saldırgana durum farkettirilmeden gerçek ve taklit sistemler paralel olarak çalıştırılmıştır.

3.2. Bal Küpü Sistemleri için Saldırgan Web Uygulaması

Web uygulamalarında en ciddi açıklıkların başında şüphesiz ki Sql Enjeksiyonu (Sql Injection) ve XSS (Cross – Site Scripting) gelir.

Sql Enjeksiyonu, sayfadaki ilgili alana girilen meta karakterler (‘, ; vb.) ile arka planda çalışacak olan sql sorugusuna eklemeler yaparak sorguyu değiştirme işlemine denir. Çoğunlukla web uygulamalarında kullanılır. Fakat sql sorgusunun olduğu heryerde bu saldırı tipi ile karşılaşabiliriz. Komut-1 üzerinde görülen sorgu cümlesi arka planda kulanılan sql cümleciği olarak kabul edelim.

ISSN: 2602-3563

```
sql = “SELECT * FROM users WHERE name=’ ”+kullanici+” ’ ;”
calistir(sql)
```

Komut 1. Arka plan Sql Sorgusu

Bu sorgu ile kullanici adi kısmına girilen değere eşit kullanıcılar veri tabanından çekilmek istenmektedir. Sql enjeksiyonu yapılarak tüm kullanıcıların veri tabanından çekilmesi sağlanabilir. Kullanıcı adı alanına Komut-2 üzerindeki gibi kullanıcı tarafından giriş yapıldığını kabul edelim.

```
1’ or ‘1’ = ‘1
```

Komut 2. Sql Enjeksiyonu İçin Girdi Değeri

Bu durumda arka planda çalıştırılacak olan sql sorgumuz Komut-3 üzerinde görüldüğü gibi bir sorguya dönüşmüştür.

```
SELECT * FROM users WHERE name=’1’ or ‘1’=’1’
```

Komut 2. Sql Enjeksiyonu Yapılmış Sorgu

İkinci şart doğru olduğundan dolayı sorgu çalışacak ve veri tabanındaki tüm kullanıcı adlarını döndürecektir. Böylece veri tabanından tüm kullanıcı adlarının çekilmesi sağlanmıştır.

XSS, çapraz betik saldırıları (Cross-Site Scripting) kelimelerinin baş harfleridir. Web sayfasında kullanıcıdan giriş alınacak alanlarda; herhangi bir kontrol mekanizması olmaması durumunda, html ya da javascript gibi kodların bu alana girilerek ilgili sayfanın kodları yorumlamasını sağlamaktır. Kullanımına göre 3 farklı tipi mevcuttur. Bunlar Reflected XSS, Stored XSS ve DOM XSS olarak listelenebilir (Anonymous, 2013). Bunlar arasından DOM, belge nesnesi modeli (Document Object Model) kelimelerinin baş harflerinden oluşmaktadır. XSS saldırıları içinde en tehlikeli olanı kabul edilmektedir. Web sayfaları belge olarak kabul edilerek, bu belge içerisindeki her eleman (resim, video vb.) nesne olarak görülmektedir. Javascript kodları ile bu nesneler üzerinde değişiklik yapılması mümkündür. İşte bu XSS saldırı yöntemi kullanılarak sayfa içersindeki hemen hemen herşey değiştirilebilir.

Tüm bu saldırıların tespit edilmesi ve saldırgan hakkında bilgi edinmek adına bir bal küpü sistemi geliştirilecektir. Öncelikle bal küpü üzerinde kullanıcılacak sayfanın XSS ve Sql enjeksiyonu açıklıklarını içeren gerçek bir web sayfası gibi görünmesi gerekmektedir. Bu amaç doğrultusunda web sitesi 3 sayfadan oluşturulmuştur. Bunlardan birincisi anasayfadır. Anasayfa saldırganın bir kuruluşun web sitesi olduğuna inanması için sahte bilgilerle doldurulmuştur. Diğer sayfalardan ilki XSS açıklığı diğeri ise Sql enjeksiyonu açıklığı bulunduran sayfalardır. Eğer saldırgan XSS veya Sql enjeksiyonu saldırılarında bulunursa, ilgili sayfa bu saldırıları emüle edecek ve başarılı olduğu izlenimi bırakacaktır. Fakat bunların yanında kullanılmış farklı saldırılar varsa XSS ve Sql enjeksiyonu dışındaki saldırılar simüle edilemez. XSS saldırısının simüle edileceği sayfada, 3 adet sahte makale bulunmaktadır. Sayfada gerekli alana girilecek birden üçe kadar olan sayılarla ilgi makale sayfada gösterilmektedir. Sayfa içerisinden istek yapıldığında, bal küpü isteğin beklendiği gibi olması durumunda makaleleri kullanıcıya sunar. Fakat farklı bir istek görüldüğünde tarayıcıda çalıştırılacak ancak saldırganın kimliğine ilişkin veri toplayan javascript kodu sayfaya eklenecektir.

Öncelikle sayfaya gelen http isteği incelenir. Bu isteğin Sql enjeksiyon açıklığı bulunduran sayfadan ya da XSS açıklığı bulunduran sayfadan geldiğine karar verilir. İlgili sayfanın açıklığının sömürülmeye çalışıldığına dikkat edilir. Eğer bahsi geçen açıklık sömürülmeye çalışılıyorsa bal küpü tarafından başarılı olduğuna dair cevap döndürülür. Cevap yollanmadan önce içerisinde javascript kodu eklenir. Javascript kodu, bir önlem olarak

ISSN: 2602-3563

karartılmıştır. Bu sayede saldırganın geri dönen cevabı incelemesi halinde kolaylıkla amacımızı anlaması engellenmek istenmiştir.

3.3. SSH Ataklarının Analizi ve Görselleştirilmesi

Saldırganlar sürekli olarak internette kötü amaçları için kullanabilecekleri sunucu aramaktadırlar. En belirgin hedeflerinin yöneticisinin uzaktan erişim için yapılandırdığı sunucular olduğu açıktır. Bu tip uzaktan erişim ile kullanılmak istenilen sunucularda çoğunlukla SSH (Secure Shell) servisi kullanılır.

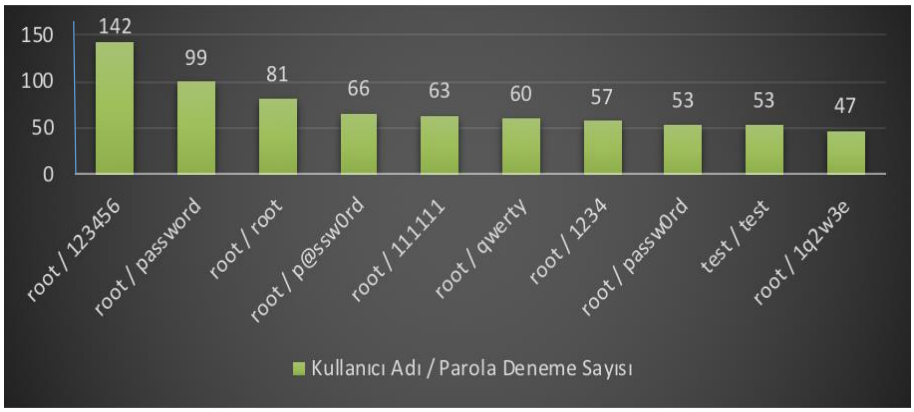
SSH şifreli bir uzak bağlantı mekanizmasıdır. Genellikle Linux ve Unix tabanlı işletim sistemlerinde kullanılır. Bağlantı sağlandığında ilgi işletim sisteminin konsoluna düşülür. İstenilen komutlar böylece uzaktan çalıştırılabilmektedir. Varsayılan olarak 22. portu kullanır.

SSH servisi, zayıf parola ile korunduğu taktirde kolaylıkla saldırganlar tarafından kullanılabilmektedir. Herhangi bir şekilde bahsi geçen ssh servisi bulan saldırganlar, sürekli olarak kullanıcı adı ve parola çiftleri deneyerek sunucuya erişim sağlamayı hedefler. Eğer bu kullanıcı adı – parola çiftlerinden herhangi biri erişim sağlamada başarılı olursa; sunucu, saldırgan tarafından kötü amaçları adına kullanılmaya başlar.

SSH servisi açık bir bal küpü sistemine statik ip adresi verilerek internete açılır. Bal küpü sistemi olarak Kippo SSH bal küpü kullanılmıştır. Kippo SSH, python programlama dili ile yazılmış bir bal küpüdür. Varsayılan 22. portta bir ssh servisi simüle etme kabiliyetine sahip olduğundan dolayı tercih edilmiştir. Bu port ile gerçekleştirilen her etkileşim kayıt altına alınır. Bu kayıtlar ayrıca MySQL veri tabanına kaydedilebilmektedir. Bu özelliğin kullanılması için sisteme ek olarak MySQL veri tabanı kurulmuştur. Kippo bu özelliklerinin yanı sıra kullanıcı adı – parola listesi alarak, bu listedeki kayıtlar ile eşleşen erişim isteklerine başarılı olduğuna dair cevap döndürür. Sisteme girdiklerini düşünen saldırganlar, işletim sisteminin simüle edildiğinden habersiz komutlar çalıştırır. Bu komutlarda kayıt altına alınarak, başarılı bir saldırının ardından yapılacak işlemler üzerinde analizlerde bulunmamıza imkân sağlanmaktadır.

Tüm bu kayıt toplama işlemlerinin ardından daha iyi analizler elde etmek adına sonuçların görselleştirilmesi amaçlanmıştır. Görselleştirme aracı olarak MySQL de tutulan veriler ile uygun olarak çalışabilen Kippo Graph uygulaması seçilmiştir. Kippo Graph, php programlama dili ile yazılmıştır. Oluşturulacak grafikler Libchart kütüphanesi kullanılarak çizdirilir.

Kippo SSH bal küpü, dört ay boyunca statik ip adresi ile internete bağlı olarak kalmıştır. Dört ay sonunda, saldırılarda 298 ip kullanılarak, SSH servisi üzerinde toplamda 23.271 adet deneme yapıldığı görülmüştür.



Şekil 1. En Sık Kullanılmış K. Adı / Parola Çifti [10]

ACTA INFOLOGICA
ISSN: 2602-3563

En çok erişim sağlanmaya çalışılan kullanıcı adı – parola çifti "root" ve "123456" olarak tespit edilmiştir. Zararlı aktiviteleri gerçekleştirmek için sisteme en yüksek haklara sahip olan kullanıcı olarak bağlanmak istemeleri şaşırtıcı değildir. Sisteme bağlanmak için; test, user ve guest gibi geçici kullanıcı adlarınında denendiği görülmüştür. En çok denenen diğer kullanıcılar ise adını kullanılacak olan servisden alan kullanıcılardır. Bunlara örnek olarak oracle, postgres veya tomcat verilebilir. Bu tip kullanıcıların yanında michael, alex veya amanda gibi sık kullanılan isimlerde denenmiştir. Şekil-1 'de kullanılmış kullanıcı adı-parola çiftleri incelenmiştir. Bal küpü, 12.269 benzersiz kullanıcı adı-parola çifti yakalamıştır. Bunların başında 142 deneme ile "root/123456" gelmektedir.

Kippo SSH bal küpüne bir kullanıcı adı – parola listesi verilerek bu kimlik bilgilerine erişimi başarılı olarak göstermesi amaçlamıştı. Bu kapsamda listeye "root/123456" çifti doğru kullanıcı adı – parola olarak verilmiştir. Fakat 23.271 denemeden 152 adet başarılı giriş olduğu görülmüştür. Ancak "root/123456" çiftini kullanan 142 adet deneme olduğu hali hazırda Şekil 1 üzerinde görülmüştü. Buradaki 10 adet fazlalık, doğru kimlik bilgilerini tespit eden saldırganların parolayı değiştirdiklerini göstermektedir.

Bağlanan kullanıcıların sistemde yaptıkları işlemlere ait kayıtlar elde edilmiştir. En çok çalıştırılmış komut, sistem bilgilerini geri döndüren "w" komutudur. Tahmin edildiği üzere, diğer en sık kullanılan komut ise ilgili klasördeki dosyaları listelemek için kullanılan "ls" komutudur. Birçok defa "-a" parametresi ile gizli dosyalarıda listelemeyi amaçladıkları görülmüştür. En sık kullanılan bir diğer komut ise "chmod +x *" komutudur. İlgili klasördeki tüm dosyalara çalıştırma izni verir. Bunlara ek olarak dosya indirmek için kullanılan "wget" ve parola değiştirmek için kullanılan "passwd" gibi birçok komutta çalıştırılanlar arasındadır.

3.4. Glashtopf Bal Küpü ile Sql Enjeksiyonu Uygulaması

Glastopf web uygulamaları için kullanılan bal küpü sistemlerinde sıklıkla tercih edilmektedir. Düşük etkileşimli bir bal küpü olduğu için kolaylıkla kurulmakta ve konfigüre edilebilmektedir. Python programlama dili ile yazılmıştır. Birçok bal küpü sisteminin aksine Glastopf, saldırıları simüle etmekte ve uygun bir cevabı üretmektedir. Bu sayede saldırganlar bir bal küpü ile etkileşimde olduklarını kolaylıkla farkedemezler. Böylece saldırganın teknikleri ve yöntemleri hakkında birçok bilgi edinilebilir. Uygulama olarak sql enjeksiyonu yapılacaktır.

Windows 10 işletim sistemi yüklü makine üzerinde bulunan sanallaştırma uygulaması ile Ubuntu 12.04 işletim sistemi kurulmuştur. Glastopf, Ubuntu içerisinde yüklenmiş ve konfigürasyonu yapılmıştır. Yine sanallaştırılmış Ubuntu üzerinden yerel olarak saldırı geçekleştirilecektir. Yapılan saldırılara ait kayıtlar sqlite veri tabanı kullanılarak Windows 10 üzerinde incelenmiştir. Gerekli konfigürasyonlar tamamlandığında Glastopf balküpü Şekil 2 üzerinde görüldüğü gibi başlatılır.

Şekil 2. Glastopf Başlatma Komutu Ve Çalıştırılması

Glastopf başlatıldıktan sonra konfigürasyonunda belirtilen çalışma dizinine gider. İlgili dizindeki db klasörü altına kayıtları tutacağı glastopf.db veri tabanını oluşturur. Eğer bu veri tabanı daha önceden mevcutsa kaldığı yerden devam ederek eklemeler yapar. Sqlite kullanıldığı için ve veri tabanı dizininin db klasörü olduğu belirtildiğinden dolayı ilgili dosya burada oluşturulmuştur.

Şekil 2 üzerinde görüldüğü gibi Glastopf hazır olarak beklemektedir. Aynı işletim sisteminde bulunan web tarayıcısı ile yerel ip adresimizin konfigürasyonda belirtilen portuna bağlanarak bal küpü ile etkileşime geçebiliriz. Bu durumda yerel ip adresimiz 192.168.5.136 olarak görülmüş ve bal küpüne konfigürasyonda 8080 portu atandığı için adres satırına http://192.168.5.136:8080/ yazarak bal küpüne Şekil 3'de görüldüğü gibi bağlanılmıştır.

ACTA INFOLOGICA
ISSN: 2602-3563

Şekil 3. Glastopf Web Arayüzü

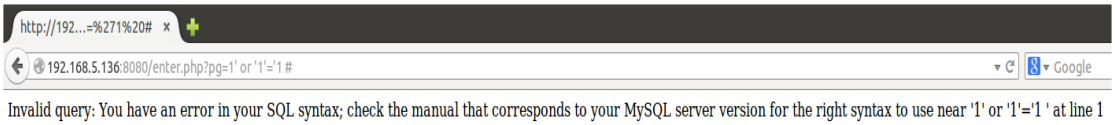
Web arayüzü oldukça basit olarak dizayn edilmiştir. Fakat istenildiği gibi üzerinde değişiklikler yapılabilmektedir. Görüldüğü gibi üzerinde giriş formu ile bir yazı ve ayrıca sayfa sonunda yorum yazılabilecek bir alan bulunmaktadır.

```
root@ubuntu:/opt/balkupuGlastopf# glastopf-runner
                    (glastopf.glastopf) Initializing Glastopf 3.1.2 using "/opt/balkupuGlastopf" as work directory.
                    (glastopf.glastopf) Connecting to main database with: sqlite:///db/glastopf.db
                    (glastopf.glastopf) Glastopf started and privileges dropped.
                    (glastopf.glastopf) 192.168.5.136 requested GET / on ubuntu.local:8080
                    (glastopf.glastopf) 192.168.5.136 requested GET /style.css on ubuntu.local:8080
```

Şekil 4. Glastopf Etkileşim Kaydı

Bal küpü ile etkileşime geçildiği anda bağlanan tarafın ip adresi ve yapılan istek kayıt altına alınmaktadır. Şekil-2 üzerinde görüldüğü gibi hazır olarak bekleyen Glastopf, Şekil 3'deki bağlanma işlemi ile Şekil 4'da görüldüğü gibi bir kayıt oluşturmuştur. Yerel işletim sisteminden saldırıyı gerçekleştirdiğimiz için ip adresimiz etkileşime geçen ip adresi olarak görülmektedir.

Glastopf, yapılan istek ve saldırıları emüle ettiğinden dolayı olmayan bir sayfanın olmayan bir pg değişkenini varmış gibi simüle etmekte ve yapılan sql enjeksiyonunu emüle ederek geriye Şekil-5 da görüldüğü gibi cevap döndürmektedir. Bu saldırının ardından Şekil 6 üzerinde görülebileceği gibi yapılan saldırıya ait kayıt düşmüştür.



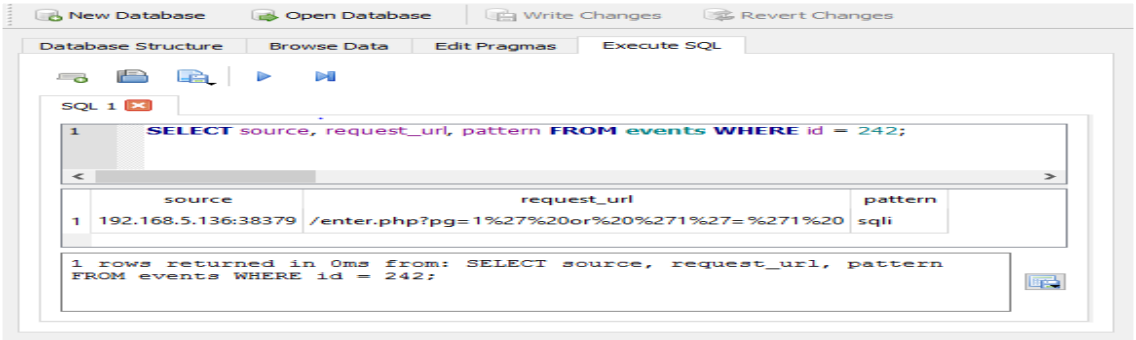
Şekil 5. Glastopf Sql Enjeksiyonu Cevabı

```
root@ubuntu:/opt/balkupuGlastopf# glastopf-runner
                    (glastopf.glastopf) Initializing Glastopf 3.1.2 using "/opt/balkupuGlastopf" as work directory.
                    (glastopf.glastopf) Connecting to main database with: sqlite:///db/glastopf.db
                    (glastopf.glastopf) Glastopf started and privileges dropped.
                    (glastopf.glastopf) 192.168.5.136 requested GET / on ubuntu.local:8080
                    (glastopf.glastopf) 192.168.5.136 requested GET /style.css on ubuntu.local:8080
                    (glastopf.glastopf) 192.168.5.136 requested GET /enter.php?pg=1%27%20or%20%271%27=%271%20 on ubuntu.local:8080
```

Şekil 6. Glastopf Sql Enjeksiyonu Kaydı

ISSN: 2602-3563

Windows 10 işletim sistemi üzerinde bulunan sqlite ile glastopf.db açılmış ve yalnızca ilgili kaydın görüntülenmesi adına select sorgusu yapılmıştır. Şekil 7 üzerinde görüldüğü gibi yapılan isteğimiz sql enjeksiyonu olarak tanımlanmış ve "sqli" etiketi ile kayıt altına alınmıştır.



Şekil 7. Glastopf Veri Tabanı Sql Enjeksiyonu Kaydı

4. TARTIŞMA VE SONUÇ

IDS, IPS ve ACL (erişim denetleme listeleri) gibi birçok pasif korunma yöntemleri, bilinen saldırı teknikleri üzerinde verimli bir şekilde çalışmasına karşın henüz keşfedilmemiş yöntemleri kullanan zararlı aktivitelerinin tespiti ve engellemesinde zayıf kalmaktadırlar. Bal küpü sistemleri bu eksikliği gidererek bilinmeyen yöntemleri kullanan saldırıları yakalamayı ve üzerinde analizler yapmayı olanaklı hale getirmiştir. Hemen hemen her alanda kullanımlarını sağlamak adına birçok bal küpü sistemi geliştirilmiştir. Gelişen modern dünyada her geçen gün farklı alanlar oluşmakta ve gereksinimler değişmektedir. Bu gereksinimlere bal küpü sistemlerinin kendini adapte etmesi her ne kadar hızlıda olsa bazı konularda eksikleri devam etmektedir. Ancak esnek yapıları ve herhangi bir sisteme kolay adapte olmaları sayesinde yeni moduller geliştirilerek eksikliklerin büyük kısmı kullanıcılar tarafından giderilebilmektedir.

KAYNAKLAR

Anonymous (2013). On Security Issues in Web Applications through Cross Site Scripting (XSS). In 2013 20th Asia-Pacific Software Engineering Conference (APSEC). IEEE.

Anonymous (2017). Antivirus software. En.wikipedia.org. Retrieved 1 February 2017, http://en.wikipedia.org/wiki/Antivirus_software.

Chauhan, S., & Shiwani, S. (2017). A honeypots based anti-phishing framework. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE.

Djanali, S., Arunanto, F., Pratomo, B., Baihaqi, A., Studiawan, H., & Shiddiqi, A. (2014). Aggressive web application honeypot for exposing attacker's identity. In 2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering. IEEE.

Koniaris, I., Papadimitriou, G., & Nicopolitidis, P. (2013). Analysis and visualization of SSH attacks using honeypots. In Eurocon 2013. IEEE.

Kuwatly, I., Sraj, M., & Al Masri, Z. (2004). A Dynamic Honeypot Design for Intrusion Detection. In The IEEE/ACS International Conference on Pervasive Services. IEEE.

Sharma, S. (2016). Detection and analysis of network & application layer attacks using Maya Honeypot. In 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). IEEE.

Singh, A., & Joshi, R. (2011). A honeypot system for efficient capture and analysis of network attack traffic. In 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies. IEEE.

Song, Y., Zhu, X., Hong, Y., Zhang, H., & Tan, H. (2012). A Mobile Communication Honeypot Observing System. In 2012 Fourth International Conference on Multimedia Information Networking and Security. IEEE.

ISSN: 2602-3563

Yang, Y., Yang, H., & Mi, J. (2011). Design of distributed honeypot system based on intrusion tracking. In 2011 IEEE 3rd International Conference on Communication Software and Networks. IEEE.

Zakaria, W., & Kiah, M. (2012). A review on artificial intelligence techniques for developing intelligent honeypot. In 2012 8th International Conference on Computing Technology and Information Management (NCM and ICNIT). IEEE.

Bu çalışma 4. Uluslararası Yönetim Bilişim Sistemleri konferansında sunularak, özeti konferans özet kitabında yayınlanmıştır.

ACTA INFOLOGICA, 2018; 2(1): 12-21
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

İdeal Steganografi Senaryosu: Taşıyıcı Resimlerin Kapasitelerinin Hesaplanması, Frekans Tabanlı Steganografide OPA Yöntemi

Ferdi Sönmez*, Faruk Takaoğlu, Oğuz Kaynar

ÖZ

Bu çalışmada steganografi'nin bir dalı olan dijital resim steganografisinden ve onunda bir alt dalı olan frekans tabanlı steganografi yöntemlerinden olan AKD (Ayrık Kosinüs Dönüşümü) ve ADD (Ayrık Dalgacık Dönüşümü)'nden bahsedilmiştir. Steganografik yöntemlerin performans hesaplama parametreleri olan OHK(Ortalama Hataların Karesi) ve TSGO (Tepe Sinyali Gürültü Oranı), gibi yöntemler açıklanmış ve bu parametrelerin değerlerinin arttırılması için resim kapasitesi hesaplama yöntemleri olan Kullback-Leibler Iraksaması, Jensen-Shannon Iraksaması ve DAS (Dörtlü Ağaç Segmentasyonu)'ndan bahsedilmiştir. Sonuç olarak resimlerdeki var olan kapasitenin daha da arttırılmasını sağlayan OPAİ (Optimal Piksel Ayarlama İşlemi) yönteminden bahsedilmiş ve ideal bir steganografi senaryosu belirtilmiştir. Çalışmamıza ek olarak bu senaryo denemesi gerçekleştirilmiş ve sonuç olarak DAS'na göre daha yüksek veri gizleme kapasitesi olan resimlerin daha yüksek TSGO değerleri verdiği sonucuna ulaşılıştır.

Anahtar Kelimeler: Ayrık Kosinüs Dönüşümü, Ayrık Dalgacık Dönüşümü, Kullback-Leibler Iraksaması, Jensen-Shannon Iraksaması, Bilgi Teknolojileri Güvenliği, Veri Güvenliği, Veri Gizleme, Frekans Tabanlı Steganografi.

Ideal Steganography Scenario: Calculation of Capacities of Carrier Images, OPA Method in Frequency-Based Steganography

ABSTRACT

In this study, digital image steganography, a branch of steganography, and DCT (Discrete Cosine Transform) and DWT (Discrete Wavelet Transform), frequency-based steganography methods that are a sub-branch of it, are mentioned. Methods such as MSE (Mean Square Error), PSNR (Peak Signal to Noise Ratio) which are performance calculation parameters of steganographic methods are explained and the methods of calculating image capacity like Kullback-Leibler Divergence, Jensen-Shannon Divergence and QTS (Quard Tree Segmentation) for increasing the values of these parameters are mentioned. This study explains the OPAP (Optimal Pixel Adjustment Process) method, which allows the existing capacity in the pictures to be further increased, in detail and provides an ideal steganography scenario. Additionally, this scenario has been tried and consequently reached the result that the images with higher data concealment capacity than QTS have higher PSNR values.

Keywords: Discrete Cosine Transform, Discrete Wavelet Transform, Kullback-Leibler Divergence, Jensen-Shannon Divergence, Information Technology Security, Data Security, Data Hiding, Frequency Based Steganography.

Information of Author(s):

Ferdi Sönmez
ORCID: 0000-0002-5761-3866
ferdisonmez@hotmail.com

Faruk Takaoğlu
ORCID: -
faruktakaoglu@stu.aydin.edu.tr
İstanbul Aydın Üniversitesit

Oğuz Kaynar
ORCID: 0000-0003-2387-4053
okaynar@cumhuriyet.edu.tr
Cumhuriyet Üniversitesi, İİBF, YBS Bölümü

DOI: 10.30801/acin.358076

Submit Date: 26.11.2017
Accept Date: 14.06.2018
Publish Date: 26.06.2018

(*) Contact Author
Address: İstanbul Arel Üniversitesi, İstanbul, Türkiye ● **Telephone Number:** +90 850 27 35 - 1299

ACTA INFOLOGICA
ISSN: 2602-3563

1. GİRİŞ

Steganografi eski çağlardan bu yana farklı formlarda ve uygulama alanlarında gördüğümüz bir bilimdir [1]. Şahısların sahip oldukları değerli bilgileri saklamak için kullandıkları bu bilim günümüzde dijital formlar üzerinden veri transferi ve iletişimi sağlamak için kullanılmaktadır. Steganografi veri güvenliği biliminin alt dallarından biri olsa da asıl icraatı verinin içeriğinin korunmasından çok verinin içeriğinin gizlenmesidir [2]. Bu da aslında steganografinin verinin korunması ile ilgilenen kriptoloji bilimine karşı avantajıdır. Kriptoloji bilimi var olan verinin içeriğinin korunmasını ve başka şahıslarca içeriğinin okunamaması veya çözülememesini amaçlar ancak steganografinin asıl amacı verinin başka şahıslarca görünmemesini sağlamaktır. Görünen ancak içeriği belli olmayan bir bilgi elbet yetenekli kişilerce ve ilerleyen teknolojilerce çözülebilir. Ancak, verinin varlığından haberdar olunamaması durumunda, kişilerin yetenekleri ve sahip oldukları teknolojiler bir anlam ifade etmez [1]. Diğer bir steganografi benzeri uygulama ise Watermarking (filigran)'dır [3]. Multimedya unsurlarının aitliklerinin ispatlanması için kullanılan yöntemdir. Bu yöntem çok benzer sistemler kullanarak var olan ve üzerinden kazanç elde edilen veya edilebilecek multimedya unsurlarının aitliklerinin ispatlanması için bu unsurların içerisine bazı imza niteliğinde bozulmalar "signature" eklemektir. Üçüncü şahısların gerekli izin veya bedeli ödemeden görsel unsurları kullanmasını engellemek için yapılmıştır. Çoğu zaman imzalar flu veya gölgeli biçimde multimedya unsurlarının üzerinde gözükür halde bulunur. Bu tarz bir uygulanış hem uygulama hemde mantık olarak steganografiye aykırıdır. Watermarking'de amaç görseldeki imza niteliğindeki bozulmayı göstermek ve resmin korunduğunu ve hatta kime ait olduğunu ispatlamakken, steganografi'de amaç veri gizlenmesinden kaynaklanan bozulmaları minimize etmek, olabildiğince göstermemektir. Steganografi bu bahsettiğimiz yönlerinden diğer benzer bilimlerden ayrılmakla birlikte, beraber de kullanılabilir. Steganografide kullanılacak gizli mesajların önce kriptolanarak sonrasında veri gizleme işlemine tabii tutulduğu uygulamalar çokça mevcuttur[2]. Bu tarz melez uygulamalar güvenliği ve gizliliği arttırıcı yöntemlerdir. Tüm bu bilgilerden sonra steganografinin kullanıldığı bazı terimler şunlardır [4]:

- Taşıyıcı Unsur / Masum Obje: Veri gömülümü yapılacak ve gizli mesajı taşıyacak çoklu ortam unsuruna verilen addır. Çalışmamız resim steganografisi olduğu için buradaki taşıyıcı objemiz resim unsurlarıdır.
- Gizli Mesaj: 3.Şahıslardan saklamak istediğimiz ve değerli olan bilgilerdir. Bu bilgiler taşıyıcı unsurların içerisine saklanır.
- Stego Key / Steganografik Anahtar: Veri gizlemesi bazen karşılıklı anlaşılarak bazen de algoritma ve yöntemler kullanılarak belirli bir düzen içerisinde yapılır. Böyle durumlarda karşılıklı olarak anlaşma sağlanmadığı için her iki tarafta veri gizlenmesini çözecek bir anahtar üzerinde anlaşırlar.
- StegObje / Steganografik Obje: Taşıyıcı resim içerisine gizli mesajın saklanmasından sonra oluşan taşıyıcı objeye çok benzeyen mültimedya unsurlarına verilen ad.
- Steganografik Analiz / StegAnaliz: Steganografik unsurların denetlenmesi ve içlerindeki gizlenmiş verilen bulunmaya çalışılması işlemine verilen addır.
- StegAnalist: Steganografik analiz işlemini gerçekleştiren uzmana verilen addır.

Güvenilir bir steganografi sistemi inşa ederken dikkat edilmesi gereken unsurlar şunlardır [5]:

- Görünmezlik: Steganografik sistemin insanlar tarafından (insan gözüyle) farkına varılamaz olmasıdır. Steganografik mesajın taşıyıcı unsur üzerine gömülümü işleminden sonra resimde meydana gelecek değişimler insan çıplak gözüyle farkına varılamaz olmalıdır.
- Güvenlik: Saldırgan taşıyıcı obje üzerinde gizli mesajın varlığını farketse bile mesajı ortaya çıkarmasının imkansıza yakın olması durumudur. TSGO (Tepe Sinyalinin Gürültüye Oranı) ölçü birimi ne kadar yüksek değerli olursa sistemimiz o kadar güvenli demektir.
- Kapasite: Önemli mesajın kapasitesinin taşıyıcı mesajın kapasitesinden fazla olmaması durumudur. Bu yaklaşık olarak maksimum %51 'lik kısımı geçmemesi tavsiye edilir aksi durumda, steganografinin unsurlarından olan görünmezlik unsuru delinmiş olur.
- Sağlamlık: Steganografinin taşıyıcı unsuru resim, video vb. üzerinde yapılan filtreleme, kesme-kırpma, yön değiştirme ve sıkıştırma gibi manipülasyonlara karşı dayanıklı olması durumudur.

ACTA INFOLOGICA
ISSN: 2602-3563

1.1. İlgili Çalışmalar

Önceki ve ilerleyen bölümlerde yer yer belirtildiği gibi watermarking filigran yazılımları ve yöntemleri steganografi ile aynı branşta kabul edilse de, yaptıkları işler ve sistemsel bakış açılarından farklılık gösteririler. Yaghmaee ve Jamzad, çalışmalarında resim kapasitelerinin hesaplanması konusunda eğitici bir yayın çıkarıldığı görüşmektedir [3]. İlk olarak resimde kapasiteyi belirleyen faktörler hakkında geniş ve kronolojik bir bilgi serisi sunulmuş daha sonrasında resimlerin kapasitelerinin ölçülmesini sağlayan yöntemler anlatılmıştır.

N. Verma'nın çalışmasında; EAB (En Az Ağırlıklı Bit), ADD (Ayrık Dalgacık Dönüşümü) ve DS (Dalgacık Steganografi) yöntemleri, geniş resim verileri üzerinde uygulanmış, avantaj ve dezavantaj'ları açısından karşılaştırılmıştır [6]. Bu 3 popüler yöntem aynı materyaller üzerinde uygulanmış ve analiz aşamasında SGO, (Sinyal Gürültü Oranı) ve Histogram analizinden faydalanılmıştır. Sonuç olarak ADD'nün daha başarılı olduğu görülmüştür. Makale renkli resimler üzerinde çalışılarak geliştirilebilir. Steganografi hakkında bilgi sahibi olmak isteyenler için uygun bir çalışmadır.

Steganografi önemli olduğu kadar onun analizinde kullanılan steganaliz metodları da araştırılmıştır [2]. İlgili bir çalışmada, steganografi platformu olan yazılımların performans analizi yapılmakta ve steganografi algoritmaları hakkında bilgilendirilme yapılmaktadır [7]. NÇK, (Normalize Çapraz Korelasyon) yönteminin steganografik sistemlerin güvenilirliğini ve algoritmalarının sağlamlığını ölçmekte kullanıldığını ve önceki çalışmalarda TSGO, dikkate alınarak yüksek değerli güçlü sinyaller içeren mesajlarda, gizli mesaj gömebilmek için alanın daha da arttığı gözlemlenmiştir. Bir steganografi işleminin kabul edilebilir olması İGS (İnsan Görüş Sistemi) tarafından anlaşılamamasına bağlıdır. Bunun için OHK (Ortalama Hataların Karesi) gibi yöntemler kullanılarak taşıyıcı ve steganografik resim arasındaki farklılıklar ölçülür ve belli değerlerin üzerinde olmamasına dikkat edilir. Bu makalede tüm bu sistemler göz önünde bulundurularak, steganografi uygulamaları belirli veri blokları üzerinde test edilmiş ve sonuç olarak "Invisible Secrets 4" ve "S-Tools" en verimli programlar olarak bulunmuştur.

Hemalatha ve arkadaşlarının 2012 yılında gerçekleştirdikleri çalışmada renkli resimlerde steganografi uygulaması yapılmıştır [8]. İnsan gözü parlaklık değişimlerini algılamada kuvvetli iken renk tonlarının değişiminde zayıftır. Bu bilgiden yola çıkarak araştırmacı renkli steganografi de "YCbCr" parlaklık değeri tutan 'Y' değerinin yerine 'Cb' mavi renk ton değerleri ve 'Cr' kırmızı renk ton değerleri üzerinde veri gizleme işlemi gerçekleştirmiştir. Veri gizleme işlemi dönüşüm formülleri kullanılarak renklerin sayısal değerleri üzerinde gerçekleştirilmiştir. 256x256 boyutunda bir renkli resim içerisine 128x128 boyutunda bir siyah beyaz resim gizlenmiştir. ADD kullanılarak resimler alt band'lara ayrılmıştır. En düşük seviyeli band olan LL bandında veri gizleme işlemi yapılmıştır. EAB, bir piksel tabanlı dönüşüm sağlayan yöntemdir.

Suvarna ve Chandel çalışmalarında TSGO, değerlerine göre beş Wavelet Dalgacık türünün performans analizini yapmışlardır [9]. Buradaki önemli olan bilgi yeni başlayanlar için Matlab platformundaki Wavelettools'un yöntemleri olduğu gibi dalgacık türlerinin de kendi içinde yöntemleri olduğudur. Bu türler yapılacak işlemlerin veya işlenecek sinyallerin özelliklerine göre farklı performans gösterebilirler. Bu yüzden performanslı işlem yapacakların bu türlere dikkat etmesi gerekmektedir. Çalışmada, tüm Matlab-Wavelettools türleri anlatılmış ve steganografi işlemi adımları ile açıklanmıştır. Testlerin sonucuna göre ise Dört Seviyeli Haar ADD diğer Dalgacık türlerine göre daha başarılı bir TSGO değeri vermiştir.

Dhawale ve arkadaşları, steganografinin uygulama platformlarından biri olan dijital resim steganografisi alanında tanıtıcı yönü ağır basan özet makaleye benzer bir çalışma yapmışlardır [10]. Bu makalede piksel tabanlı veri gizleme yöntemlerinden olan EAB, TDA (Tekil Değer Ayrışımı), YS (Yayılı Spektrum) yöntemleri kullanılmış, frekans tabanlı steganografi yöntemleri olarak; ADD, AKD, AFD (Ayrık Fourier Dönüşümü) ve TDD (Tamsayı Dalgacık Dönüşümü) yöntemleri ele alınmıştır. Çalışmanın amacı bir masum taşıyıcı resim verisi içerisine bir papatya resmini gizlemek ve yukarıda bahsi geçen yöntemlerin performans parametrelerince test edilerek hangisinin daha başarılı olduğunun bulunmasıdır. Çalışmada steganografi bilimi hakkında yeni bilgi edilecekler için çokca bilgi mevcut iken bu alanda çalışma sahibi olanlar için sığ denilebilecek veriler içermektedir. Çalışmada kullanılan materyal çeşitliliği arttırılmalı ve güçsüz oldukları bir bedahet olan yöntemler rastgelelik veya sinyal düzeltmeyi sağlayan hamming kod gibi uygulamalarla güçlendirilerek frekans tabanlı olan ve başarı oranları bilinen yöntemlerle tekrar kıyaslanmalıdır.

ACTA INFOLOGICA
ISSN: 2602-3563

2. YÖNTEM

Resimlerde veri gizlemeye uygun olan bölgelerin ve bu bölgelerin kenar noktalarının iyileştirilmesi üzerinde durulmuş, matrisler kullanılarak gruplar halinde veri gizleme işlemlerinin yapılmasının rastgele veri gizleme işlemi yapılmasından daha etkili olduğu keşfedilmiştir. Frekanslar üzerinden işlemler yaparken elimizdeki gizlenecek verinin büyüklüğüne göre hangi frekans band'larında veri gizlemesi yapılması gerektiği (çok verinin düşük band'larda, az veri yüksek band'larda saklanmalıdır) belirlenmiştir. Steganografi yapılırken kullanılan resimlerin dokusal özelliklerinin veri gizleme kapasitesini etkilediği anlaşılmıştır. Frekans tabanlı işlemlerde steganografinin anlaşılmaması için bazı durumlarda gürültü ekleme veya gürültü temizleme işlemleri yapılarak steganografi uygulanmış ve elde edilen sonuçlar not edilmiştir. Veri madenciliği yöntemleri kullanılarak steganaliz çalışmaları da yapılmıştır [11]. Resimlerin olası enerji seviyeleri hesaplanarak eşik değerleri belirlenmiş ve bu değerlerden yola çıkarak steganaliz yöntemleri geliştirilmiştir.

2.1. Frekans Tabanlı Steganografi

Frekans tabanlı steganografi yöntemleri daha karmaşık ve anlaşılması zor yöntemler olduklarından daha güvenli ancak daha az tercih edilen yöntemlerdir [12]. Bu yöntemlerde diğer piksel bazlı değişim yöntemlerinde olduğu gibi resim içerisinde yer alan ancak değişiminin resmin bütününde çokça bir fark oluşturmayacağı bit'lerin bulunması ve bunların üzerinde değişim yapılmasını hedefler. Kısacası mantık olarak diğer yöntemlerle aynıdır ancak izlediği yollar daha karmaşık ve resimde daha az etki oluşturur [11]. AKD yönteminde resim 8x8 boyutunda matris'lere bölünür ve her bir matris bloğundan 64 adet AKD katsayısı elde edilir. EAB mantığında olduğu gibi burada da resimlerin parlaklık değerlerini içermeyen ve minimum bozulmaya sebebiyet verecek olan bit'ler artık bit'lerdir ve bunların üzerinde değişim yapılabilir. Genel bilgi olması açısından resim uzantıları aslında resimlerin sıkıştırılma yöntemlerini sembolize eder ve JPEG uzantısında olduğu gibi bir grup veya kuruluşun önderliğinde yapılmaktadır. GIF uzantılı resimlerde veri saklamada kullanılan platformlardan biri olsada GIF resimlerin renk histogramında denge içermesi ve veri gizlemenin bozulmalara yol açmasıyla kolaylıkla bulunabilmesi söz konusudur bu yüzden JPEG veri saklama yöntemlerinde çokca kullanılan bir resim uzantısıdır. Çünkü en performanslı sıkıştırma yöntemine sahiptir. Bu yöntemde [13], sıkıştırılması yapılmış resimlerin değerleri üzerinde oynamak, resim üzerinde bozulmalara elbette sebebiyet verecektir. Ancak en az etkiye sahip değerler üzeriden değişim yapıldığında, resimlerin parlaklık değerlerinden ziyade tonlamaları değişebileceğinden fark edilmesi zor olacaktır. Steganaliz yöntemlerinden olan ve resimlerin bölgesel olarak histogram ve entropi analizleri yapıldığında buradaki resimlerin bozulmaları takip edilerek veri gizlenmesi anlaşılabilir. ADD yönteminde de AKD mantığıyla resim değerleri frekans değerlerine çevrilmekte ve farklı frekans bandlarının farklı özelliklerine uygun veri gömülümü yapılabilmektedir [14][11][12]. Örneğin, yüksek frekans bandında az, düşük frekans bandında çok veri gizlemek.

i) AKD - Ayrık Kosinüs Dönüşümü

$$F(u,v)= \frac{c(u)c(v)}{4}\sum_{i=0}^{7}\sum_{j=0}^{7}\cos\left(\frac{(2i+1)u\pi}{16}\right)\cos\left(\frac{(2j+1)u\pi}{16}\right)f(i,j)$$

$$c(\acute{\epsilon}) = \begin{cases} 1^{1/\sqrt{2}} & \acute{\epsilon} = 0 \text{ ise} \end{cases}$$

Formüldeki;

- F(u,v) fonksiyonu bir AKD'nün (u,v) kordinatındaki,
- f(i,j) fonksiyonu bir AKD'nün (i,j) kordinatındaki piksel değerlerini göstermektedir.

AKD'nde [10], kosinüs sinyallerini kullanan ve resimleri uzaysal tabandan frekans tabanlı matris yapısına kosinüs dönüştürücüsü ile dönüştüren bir yapıdır.

ii) ADD - Ayrık Dalgacık Dönüşümü

$$W_{j,k}(t) = 2^{-j/2}\, W(2^{-j}t - k)$$

W, sürekli bir fonksiyon / j, skala parametresi / k, öteleme parametresidir [15].

Bu yöntem dalgacık olarak tabir ettiğimiz ana sinyalin matematiksel, zaman ve frekans bandında ufak dalgalara böler ve bu bandlarda işlem yapar. Bu dalgacıkların diğer yöntemlere kıyasla üstünlüğü daha ufak zaman dilimlerinde meydana gelen ufak ama sonucu etkileyebilecek dalgalanmaları inceleyebilmemize olanak sağlamasıdır. AFD'nde, olduğu gibi sinyali sadece tek bir frekans bandında incelemez, daha ufak ve ayrıntılı

ACTA INFOLOGICA
ISSN: 2602-3563

dalgacıklarda incelenerek daha iyi gözlemler yapabilmemize olanak sağlar. AKD'nde olduğu gibi ADD uygulanırken taşıyıcı resim ve mesaj frekans boyutuna dönüştürülmektedir [16][17].

2.2. Steganografide Performans Parametreleri

Steganografik algoritmaların performansları, gizli mesajı içerisine sakladığımız steganografik resim ve bu resmin içerisine mesaj iletisi eklenmeden önceki hali olan taşıyıcı resmin karşılaştırılması ile belirlenmektedir [12][11]. Bu karşılaştırmaya dayalı analiz ise bazı matematiksel parametrelerin bulunması ile tamamlanmış olur. Bu parametrik değerler; TSGO, OHK, NÇK, OF (Ortalama Farkı), Yİ (Yapısal İçerik), MF (Maksimum Fark) ve NMH (Normalleştirilmiş Mutlak Hata)'dır. Bu parametreler taşıyıcı obje ile steganografik obje arasındaki farkı değerlendirmemizi sağlayan matematiksel fonksiyonlar içermektedir [18].

i) OHK - Ortalama Hataların Karesi

OHK, genellikle sinyallerde iki sinyalin birbirilerine olan benzerliklerini ölçmek için kullanılan bir yöntemdir. Steganografide buna benzer olarak taşıyıcı resim ile steganografik resmin benzerliklerini ölçmek için kullanılır. Aşağıdaki formüle göre OHK, benzerlik bulmaya çalışmaktadır [10].

$$\text{OHK} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

Formüldeki;

- I(i,j) değeri orijinal taşıyıcı resmi temsil etmektedir.
- K(i,j) değeri steganografik resmi temsil etmektedir.
- m,n değerleri ise resmin boyutlarını göstermektedir.

Formülün sonucunda eğer OHK değeri düşük ise bu benzerliğin az olduğunu ve algoritmanın başarılı olduğunu göstermektedir. Tersi durumlarda ise algoritmamız başarısız sayılacaktır.

ii) TSGO - Tepe Sinyalinin Gürültüye Oranı

TSGO, logaritmik desibel ölçütü ile tanımlanır, ölçümlendirilir. Steganografik resmin görüntüsünün bozulumuna sebebiyet veren en üst seviye sinyal ile bozuluma sebebiyet veren gürültü değerinin arasındaki orana TSGO denir. Düşük TSGO ölçümü görsel kalitede düşüklük ve bilgi sıkıştırmada kalitesizlik anlamına gelir. Tersine durumda yani TSGO'nun yüksek ölçüldüğü durumda, resim kalitesi, sıkıştırması ve yeniden yapılandırılmasının kaliteli ve başarılı olduğu anlaşılır. TSGO değeri aşağıdaki formül ile hesaplanmaktadır [10].

$$\text{TSGO} = \log_{10}\left(\frac{MAX_1^2}{MSE}\right)$$

TSGO formülü görüldüğü üzere başka bir ölçüm parametresi olan OHK değerine bağlı olarak hesaplanır. MAXı değeri var olan en yüksek piksel değeridir.

2.3. Resimlerin Veri Gizleme Kapasiteleri

Dijital resimlerin kapasiteleri, çözünürlükleri ve hafızada kapladıkları alanla doğru orantılı olarak değişse de steganografide bunların haricinde renk, bit derinliği ve dinamik değer aralığı değerleri de önem kazanmaktadır. Bir resmin çözünülürlüğü Uzaysal Çözünürlük ve Tonal Çözünürlük olarak iki kısımda incelenir. Uzaysal Çözünürlük .dpi (dots per inch, inç başına nokta) değerinde var olan nokta sayısı veya .ppi (pixel per inch, piksel hassasiyeti) değeri gibi yöntemlerle ölçülebilir. Her iki yöntemde resimlerde detayı yakalayabilmek olarak adlandırılan görüntü kalitesinin ölçümü için gereklidir. Steganografide bu detay parametresi önemli olduğu gibi aynı zamanda Tonal Çözünürlük denilen renk çeşitliliği ve geniş bir yelpaze içermesi ve tüm bu değerlerin yüksek değerlikli olması hem resim için kalite unsurunu arttırıcı hem de steganografi için veri gömülümü kapasitesini arttırıcı özellik içerir. Resmin kendi çözünürlük değerleri dpi gibi parametreler ile resim boyutunun orantısının düzgün kurulmasına bağlı olduğundan (4800x6000 piksel boyutu 8x10 inç'lik baskı boyutunda 600 dpi çözünürlük içerip 28.8 MB (mega-bayt) boyuta sahipken, yine aynı piksel boyutunun 4x5'lik baskı boyutunda 1200 dpi çözünürlük oluşturması ve 86.4 MB'lık boyuta sahip olması gibi.) önemli olanın uygun çözünülürlükte uygun boyut değerleri ile seçilim yapılmalı ve steganografi için kapasiteli resimler oluşturulmalıdır [19]. Aksi durumda, boyutu düzgün olmayan resimlerde uygulanan steganografi görsel bozulmalar göstereceğinden kolayca anlaşılabilmektedir.

ACTA INFOLOGICA
ISSN: 2602-3563

i) Kullback-Leibler Iraksama Yöntemi

Yapacağımız bu çalışmada resimlerin veri saklama kapasitesini ölçmek için KL-Iraksama yöntemini kullanacağız. Bu yöntem, Liu ve arkadaşları çalışmalarında [12] aktarıldığı gibi kapasite ölçme yöntemleri ile ters orantı göstermektedir. Yani resimlerde ölçüm yapılırken, OPA (Optimal Piksel Ayarı) veya GKM (Gauss Karışımı Modeli) ve FBM (Fisher Bilgi Matrisi)'nden elde edilen değerlerin yüksek çıkması beklenirken, KL-Iraksama değerlerinin olabildiğince düşük çıkması istenilmektedir. KL-Iraksama değeri en az olan resimler bize resim saklama kapasitesinin en yüksek olduğu resimleri gösterecektir. Böylelikle yüksek kapasiteli resimlere ulaşılabilecek ve amacımız doğrultusunda resim saklama kapasitesi hesaplanmamış resimlerle Stegnografi yöntemleri üzerindeki etkileri ölçülebilecektir. Aşağıda KL-Iraksama yönteminin formülü yer almaktadır:

$$D_{KL}(P||Q) = \int_{\mathbf{X}} P \, ln\left(\frac{P}{Q}\right) d\mathbf{X}$$

Bu denklemde P(X|O) Masum X objesinin dağılım olasılığı değeri, Q(Y|O) değeri ise Steganografik resim olan Y'nin dağılım olasılığı değeridir. KL-Iraksama yöntemi bu iki dağılım/serpilme olasılıklarının farkının hesaplanmasıdır.

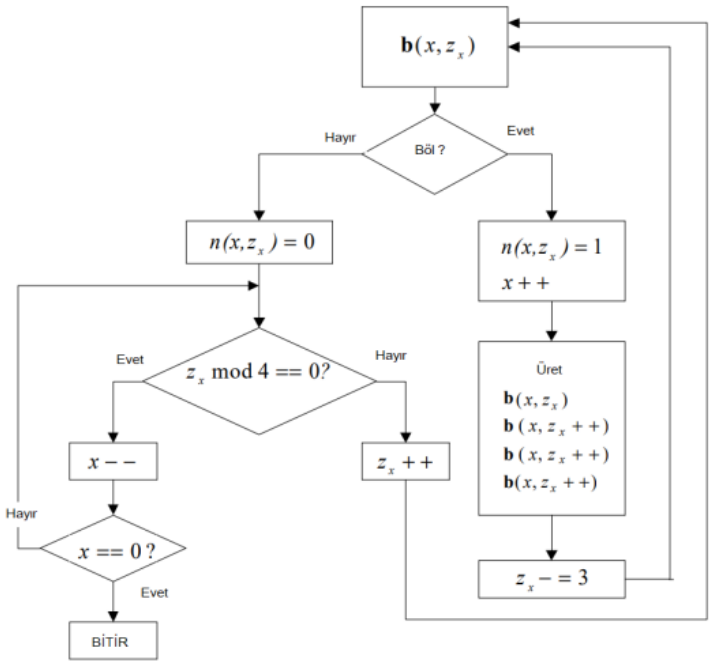
ii) Jensen-Shannon Iraksama Yöntemi

$$\mathrm{ICC}(X) = H(X) - \sum_{S=1}^{R} \frac{n_s}{N} H(I_s),$$

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i,$$

Burada, N = Tüm piksel sayısı, X = Orijinal Resim, R = Toplam Segment Sayısı, n_s = Segment'deki piksel sayısı, Is = Segmentle bağıntılı histogram değerinin rastgele yoğunluk değeri, H = Entropi Fonksiyonu

Görsellik açısından farkına varılabilirliğin zor olması için heterojenlik değeri resim içerisinde sağlanmalıdır. Böylelikle, bozulmaların farkına varılması zor olacaktır [7]. Resim içindeki her bir segment de JS-Iraksama formülü rastgele dağılım ile bu hesaplamayı yapmaya çalışmaktadır.



Şekil 1. Dörtlü Ağaç Segmentasyonu Algoritması

ACTA INFOLOGICA
ISSN: 2602-3563

iii) DAS - Dörtlü Ağaç Segmentasyonu

Bu yöntemin amacı resim içerisindeki parlaklık "kontrast" değerlerinin düşük olduğu pikselleri bulmaktır. Bunun tercih edilmesinin sebebi [20] 1.1.İlgili Çalışmalar bölümünde bahsedilen İGS, insan gözünün parlaklık değerlerine karşı duyarlı olmasıdır. Yüksek kontrast değerlerine veri gizleme işlemi yapıldığında oluşan bozulmalar insan gözüyle farkına varılabilir. Bu yüzden bu kısımlara fazla veri gizlenemez. Ancak parlaklık "kontrast" değerleri daha düşük olan piksellere veri gizlenmesi daha uygundur. Buna bağlı olarak, bu yöntem resimleri piksel gruplarına bölerek her bir bölgede düşük parlaklık değerli pikselleri işaretler. Bu işaretlenmiş yerlerin sayısı resimdeki veri gizleme kapasitesini belirtmektedir [21]. Şekil 1'de DAS yönteminin bir algoritması verilmiştir.

2.4. OPAS - Optimal Piksel Ayarlama Süreci

OPAS (Optimal Piksel Ayarlama Süreci) veya sadece OPA, steganografide mesaj gömülümü gerçekleştikten sonra uygulanan bir yöntemdir. Asıl amacı steganografik resim ile orijinal resim arasındaki farklılıkları veya hataları gidermek, en aza indirmektir [22]. Örneğin elimizde taşıyıcı orijinal mesajın bir bölümünün binary değeri olsun. Bu değerin 10000 olduğunu farz edelim (10'luk sayı tabanındaki değeri 16). Sonrasında bu noktaya gizlemek istediğimiz gizli mesajın değerinin 1111 (10'luk tabanda değer karşılığı 15) olduğunu farz edelim. Bu iki değerin steganografi sonucunda birleşmesi ile elde edilecek değer 11111'dir (10'luk sayı tabanındaki değeri 31). Sonuç olarak 10'luk sayı tabanında hesaplandığında eski orijinal resim ile steganografik resim arasında 16 fark olduğu gözlemlenir. OPA algoritması bu 11111 sayı bloğunun EAB değerini değiştirerek 01111 yani 15'e çeker ve sonuç olarak taşıyıcı medya/orijinal resimle arasındaki farkı 1'e indirir. Bu uygulama yapılarak resimlere daha fazla gizli mesaj yüklenebilir ve dolayısıyla kapasiteler arttırılmış olur [23].

3. BULGULAR: DENKLEM ÖRNEĞİ

Bu çalışmadaki amacımız ideal bir steganografi senaryosu veya prosedürü belirlemektir. Şekil 2'de senaryonun adımları gözükmektedir.

Şekil 2. İdeal Steganografi Senaryosu

ACTA INFOLOGICA
ISSN: 2602-3563

Burada,

*1= Seçilen resime ADD "haar" yönteminin uygulanması ve LL "low" bandın seçilimi –değerlerin bulunması,

*2= Low band değerlerine AKD yöntemiyle öznitelik katsayılarının 8x8 matris bloklarından çıkarılması öznitelik tablosunun oluşumu,

*3= EAB mantığıyla en az değer içeren bit'lerin yani artık bit'lerin bulunması,

*4= Gizli mesaj ve gizli mesaj uzunluğunun birleştirilip eş sayı tabanına çevrilerek artık bit'lere eklenmesi ve gizlenmiş veri içeren öznitelik tablosunun oluşumu,

*5= Gizlenmiş öznitelik tablosundan OPA algoritmasının uygulanarak hataların minimize edilmesi ve kapasitenin arttırılması,

*6= Kalan masum değerleri içeren öznitelik değerleri ile gizli veri içeren öznitelik değerlerinin birleşimi,

*7= Ters ADD uygulanarak özniteliklerin frekans bandına, sonrasında da steganografik resime dönüşümü,

*8= Orijinal resim ile steganografik resim kullanılarak bulunan KL-Iraksama, JS-Iraksama yöntemlerinin veya DAS yönteminin uygulanması.

Yukarıdaki adımlarda ideal bir steganografi senaryosu verilmiştir. 8. adımda KL/JS-Iraksama yöntemlerinden çıkan sonuçlarda en ufak değerlerin aslında en yüksek veri saklama kapasitesi resimleri olduğunun işaretidir. Sonuç olarak bu işlem aynı anda birden fazla resme uygulanıp ve en sonunda bu hesaplama yöntemleri test edilirse minimum değerli resim en yüksek kapasiteli resim olarak anlaşılır ve sonuç olarak gizlenmiş verinin aktarımında o resim tercih edilebilir. Diğer bir yöntem olan DAS seçilerek steganografi işlemi yapılmadan önce resimlerin veri gizlemeye uygun olan blokları seçilerek bu bloklara veri gizlenmesi için özel bir kodlama işlemi yapılabilir veya doğrudan hangi resimde daha fazla uygun bloğu varsa o resme veri gizlemesi uygulanarak daha yüksek TSGO değerleri elde edilebilir. Aşağıdaki resimlere DAS uygulanmış ve sonuç olarak ADD steganografi sisteminden sonraki elde edilen TSGO değerleri verilmiştir.

Şekil 3. Taşıyıcı Resim 1

Şekil 4. Taşıyıcı Resim 2

Şekil 5. Taşıyıcı Resim 3

Tablo 1. Taşıyıcı Resimlerin DAS Uygun Blok Sayıları

DENEMELER	ADD-TSGO
Taşıyıcı Resim 1	16.252
Taşıyıcı Resim 2	16.096
Taşıyıcı Resim 3	16.072

Tablo 2. Taşıyıcı Resimlerin ADD sonrası TSGO değerleri

DENEMELER	ADD-TSGO
Taşıyıcı Resim 1	50,3908 dB
Taşıyıcı Resim 2	47,0555 dB
Taşıyıcı Resim 3	46,0858 dB

ACTA INFOLOGICA
ISSN: 2602-3563

Elde edilen tablolardan anlaşılacağı gibi DAS yöntemine göre veri gizlemeye uygun olarak belirtilen veri bloklarının sayıca fazla olduğu resimlerde daha fazla TSGO değerleri elde edildiği görülmektedir. Bazı araştırmacılar [24][25][26] çalışmalarında benzer yöntemler ve metotlar kullanmıştır. Çalışmamızda daha fazla sayıda veri gizlenmesine rağmen bu çalışmalara yakın TSGO değerleri elde edilmiştir.

4. TARTIŞMA VE SONUÇ

Bu çalışmada, AKD ve ADD frekans tabanlı Steganografi yöntemlerinin birlikte kullanımını ele alınmıştır. Çalışmamızda ADD'nün alçak frekanslı ve veri gizlemeye müsait bandları çıkarma özelliği ve bu bantlarda AKD'nin öznitelik katsayılarını elde ederek EAB yöntemini uygulamasından faydalanılmıştır. Bu hibrid yöntemler ek olarak diğer çalışmalardan farklı olması için gizli mesaja uygulanan veri sıkıştırma yöntemleri kullanılmadan, OPA algoritması tercih edilmiş ve taşıyıcı unsur üzerindeki değişimleri minimize etme yoluna gidilmiştir. Çalışmamızda son olarak resimlerin kapasitelerinin ölçülmesi alanına değinilmiştir. Steganografi alanında resimlerin gizli veri taşıma kapasiteleri üzerinde halen tam bir görüş birliği bulunmamaktadır. Bu alanda denenen yöntemler olan Iraksama Yöntemlerinin yerine DAS yöntemi kullanılmış, taşıyıcı resimlerinin başarılı TSGO değerlerinin DAS yönteminden çıkan taşıma kapasiteleri ile doğru orantılı olduğu görülmüş ve kapasitesi yüksek taşıyıcı resimlere veri gizleme işlemi uygulandığında daha başarılı TSGO değerleri alınacağı anlaşılmıştır. Bazı araştırmacılar çalışmalarında benzer yöntemler ve metotlar kullanmıştır. Çalışmamızda daha fazla sayıda veri gizlenmesine rağmen bu çalışmalara yakın TSGO değerleri elde edilmiştir.

Çalışmamızda, AKD ve ADD beraber kullanımı ile hibrid bir yöntem sunması ve ayrıca OPA, DAS gibi yöntemlerle bağdaşması güvenilirliğini ve sağlamlığını arttırmaya yönelik avantajlar olarak gözükse de sadece siyah beyaz resimlerde çalışması ve renkli resimlerde uygulanmaya başlandıkça uygulama alanında karmaşıklık ve hataların artması açısından dezavantajlar göstereceğini düşünmekteyiz. Bu dezavantajların önüne geçmek ileriki çalışmalar için ayrı bir motivasyon ve araştırma konusu olarak görülmektedir.

KAYNAKLAR

[1] Holub, V., Fridrich, J., Denemark, T. (2014). Universal distortion function for steganography in an arbitrary domain, EURASIP Journal on Information Security, 2014(1), ss.11-19.

[2] Sajedi, H. (2016). Steganalysis based on steganography pattern discovery. Journal of Information Security and Applications, 30, 3-14.

[3] Yaghmaee, F., Jamzad M.(2010). Estimating watermarking capacity in gray scale ımages based on ımage complexity , EURASIP Journal on Advances in Signal Processing, ss. 20102010:851920, doi: 10.1155/2010/851920.

[4] Subhedar, M.S., Mankar, V.H. (2014). Current status and key issues in image steganography: A survey, Computer Science Review, 13, ss. 95-113.

[5] Challita, K., Farhat, H. (2011). Combining steganography and cryptography: new directions. International Journal on New Computer Architectures and Their Applications, 1(1), ss.199-208.

[6] Verma, N.(2011). Review of steganography techniques, International Conference and Workshop on Emerging Trends in Technology (ICWET 2011)–TCET, Mumbai, India

[7] Zeki, A.M., Ibrahim, A.A., Manaf, A.A.(2012). Steganographic software:analysis and implementation, International Journal Of Computers And Communications, 6(1).

[8] Hemalatha, S., Acharya U. D., Renuka, A., Kamath, P.R.(2012). A Novel color image steganography using discrete wavelet transform, CCSEIT-12, October 26-28, 2012, Coimbatore, India.

[9] Suvarna P., Chandel, G.S.(2013). Performance analysis of steganography based on 5-wavelet families by 4 levels -dwt suvarna, International Journal of Advance Research in Computer Science and Management Studies, 1(7), ss. 20-33.

[10] Dhawale, C. A., Hegadi, R., Jambhekar, N.D.(2014). Performance analysis of digital ımage steganographic algorithm. ICTCS '14, Kasım 14 – 16, 2014, Udaipur, Rajasthan, India, ACM 978-1-4503-3216-3/14/11.

ACTA INFOLOGICA
ISSN: 2602-3563

[11] Sujatha, P., Purushothaman, S., Rajeswari, R. (2014). Performance study of combined artificial neural network algorithms for ımage steganalysis, In Proceedings of International Conference on Internet Computing and Information Communications, ss. 441-451, Springer India.

[12] Liu, Y., Liu, Y., Wu, S., Zhong, S. (2015). What Makes the Stego Image Undetectable? ICIMCS '15, Ağustos 19-21, 2015, Zhangjiajie, Hunan, China, ACM. ISBN 978-1-4503-3528-7/15/08.

[13] Hemalatha, S., Acharya, U.D., Renuka, A.(2015). Wavelet transform based steganography technique to hide audio signals in image, Procedia Computer Science, 47, ss.272-281.

[14] Provos, N., Honeyman, P. (2001). Detecting steganographic content on the internet, Center for Information Technology Integration, NDSS 2002, San Diego.

[15] Haşiloğlu, A. (2001). Dalgacık dönüşümü ve yapay sinir ağları ile döndürmeye duyarsız doku analizi ve sınıflandırma, Turk J. Engin Environ Sci, 25, ss.405-413.

[16] Dalvi, A., Kamathe, R.S. (2015). Color ımage steganography by using dual wavelet transform (DWT SWT). International Journal of Scientific Engineering and Research (IJSER), 3(7), ss.25-41.

[17] Bera, S., Dewangan, U., Sharma, M. (2013). Development and analysis of stego ımage using discrete wavelet transform, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[18] Cheddad A., J. Condell, K., Kevitt, P.(2010). Digital ımage steganography: survey and analysis of current methods, Signal Processing, 90(3), ss.727-752.

[19] Peterson, A.K. (2005). Introduction to Basic Measures of a Digital Image for Pictorial Collections, Prints & Photographs Division, Library of Congress, Washington, DC, ss. 20540-4720.

[20] Muhsin, Z. F., Rehman, A., Altameem, A., Saba, T., & Uddin, M. (2014). Improved quadtree image segmentation approach to region information. The Imaging Science Journal, 62(1), ss. 56-62.

[21] Lin,Y-C, Li, T-S (2011). Reversible image data hiding using quad-tree segmentation and histogram shifting, Journal of Multimedia, 6 (4), ss. 349-358.

[22] Kaur, S., Goel, N. (2015). Segmentation and block based image steganography using optimal pixel adjustment process and identical approach, 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS), Chandigarh, 2015, ss. 1-5.

[23] Nithya , R. K., Nehru, C.P., Ubramanıam, T.B.(2014). Optimal pixel adjustment based reversible steganography, (IJITR) International Journal Of Innovatıve Technology And Research, 2 (3), 2014, ss. 963–966.

[24] Demirci, B. (2016). Görüntü steganografi metotları ve performanslarının karşılaştırılması (Doktora Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü).

[25] Kaya, H. V. (2015). Watermarking in medical images by using DWT, DCT, DFT and lSB algorithms (Doktora Tezi, Çankaya Üniversitesi).

[26] Şahin A. (2007). Görüntü Steganografide Kullanılan Yeni Metodlar ve Bu Metodların Güvenilirlikleri. (Doktora Tezi, Selçuk Üniversitesi).

Bu çalışma 4. Uluslararası Yönetim Bilişim Sistemleri konferansında sunularak, özeti konferans özet kitabında yayınlanmıştır.

ACTA INFOLOGICA, 2018; 2(1): 22-32
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

ANFIS Analysis of Wireless Sensor Data with FPGA

Ahmed Khazal, Tuncay Ercan*

ABSTRACT

Applications related with WSNs may include thousands of separate sensor nodes, production and control data for different industrial sectors. It is important to manage these applications, monitor the network and reprogram the nodes to avoid operational problems. In this study, we propose a smart wireless sensor network using a reconfigurable embedded system of Field-Programmable Gate Arrays (FPGAs) with a soft-core processor. This software-based processor can be programmed dynamically and synthesized to implement the pre-processing of sensed data by ensemble Hybrid Neuro-Fuzzy algorithms such as Adaptive Neuro-Fuzzy Inference System (ANFIS). The first part of the proposed work is based on Matlab software to develop and train the ANFIS algorithm. Two different types of data sets (temperature and humidity) downloaded from Internet have been used in order to make a comparison between the Matlab Toolbox and modified ANFIS algorithm with momentum factor. The results obtained in this study have shown that the modified ANFIS algorithm is the convenient choice in terms of speed and accuracy.

Keywords: ANFIS, Neuro-Fuzzy System, FPGA, Sensor nodes, Smart node.

FPGA ile Kablosuz Sensör Verilerinin ANFIS Analizi

ÖZ

Kablosuz algılayıcı ağlarla ilgili uygulamalar binlerce ayrık algılayıcı cihazı, farklı endüstriyel sektörlerdeki üretim ve kontrol verilerini içerebilir. Bu uygulamaları yönetmek, bilgisayar ağını izlemek ve her bir cihazı tekrar programlayarak işletme problemlerinden kurtulmak çok önemlidir. Bu çalışmada yazılım tabanlı işlemci özelliğiyle yeniden yapılandırılabilir bir FPGA (Alan Programlanabilir Kapı Dizini) cihazı kullanan akıllı bir kablosuz algılayıcı ağı mimarisi önerilmiştir. Bu yazılım tabanlı işlemci, dinamik olarak tekrar programlanabilir ve ANFIS (Uyarlamalı Sinir-Bulanık Çıkarsama Sistemi) gibi bir toplu hibrit Sinir-Bulanık algoritması ile algılanmış verileri önişlemden geçirebilir. Önerilen çalışmanın ilk bölümü ANFIS algoritmasını geliştirmek ve eğitmek için Matlab yazılımı kullanmaya dayanır. Bu çalışmada internet ortamından tedarik edilmiş, sıcaklık ve nem verileri içeren iki veri seti ile Matlab ortamında hazır bulunan ve çalışma kapsamında momentum faktörü ile değiştirilerek geliştirilmiş ANFIS algoritmalarının karşılaştırması yapılmıştır. Bu çalışma değiştirilmiş ANFIS algoritmasının hız ve doğruluk açısından daha uygun olduğunu göstermektedir.

Anahtar Kelimeler: ANFIS, Bulanık Sinir Sistemi, FPGA, Algılayıcı düğüm, Akıllı düğüm.

Information of Author(s):

Ahmed Khazal
ORCID: 0000-0001-8171-5582
ahmedkhazal@gmail.com
Yasar University, Department of Computer Engineering

Tuncay Ercan
ORCID: 0000-0003-0014-5106
tuncay.ercan@yasar.edu.tr
Yasar University, Department of Computer Engineering

DOI: 10.30801/acin.357635

Submit Date: 24.11.2017
Accept Date: 08.06.2018
Publish Date: 26.06.2018

(*) Contact Author
Address: Yasar University, Department of Computer Engineering, İzmir, Turkey ● **Telephone Number:** +90 232 570 82 41

ACTA INFOLOGICA
ISSN: 2602-3563

1. INTRODUCTION

It is now a well-known fact that the use of Internet with the capabilities like communication, sharing of information and mutual interaction among people has changed our daily life significantly. The Internet of Things (IoT) is a new technological concept that the intelligent devices communicate with each other and form an intelligent and autonomous network in order to access Internet. Real-time continuous stream of data coming from the sensing devices and industrial equipment in the environment are aggregated, modified and transferred into the Information Systems (Storage, database, application services) provided by Cloud Service Providers on the internet. This kind of Industrial IoT (IIoT) information flow will result in changes that can positively contribute in our daily life, business life and industrial production systems.

Traditional internet communication infrastructure connects different communication systems and end users into each other. With the new developments in technology, many sensor devices can be integrated into the internet environment through WSN (Wireless Sensor Networks) systems [2][16] in order to collect the information and monitor the areas like the status of patients and elderly people, transportation, traffic flow and security and then take the appropriate decision for each case.

The requirements for wireless sensor networks like energy consumption, cost and processing capabilities, lead and motivate science and engineering disciplines to think in a different way. Applications related with WSNs may include huge number of nodes. Thus, it is important to be able to manage these applications, monitor the network and reprogram the nodes to avoid existing problems. In general, hardware part of sensor networks include processing elements (Microcontrollers), sensors, battery and wireless communication elements which sends data from sensors to sink node that routes the data to the central management system like a gateway [2][17]. Limited battery power, storage and computation capabilities, and security problems are limitation and disadvantages of traditional sensor nodes.

Due to all of these limitations, smart sensors have been developed to handle these challenges. In most cases, the smart sensor represents a system on chip (SoC) when integrated with FPGA devices. In such cases, the smart sensors can implement many intelligent functions such as pre-processing data, testing and performing complex algorithms, in addition to some communication tasks. FPGAs support reprogrammable technology which can be used for a reconfigurable sensor system [17]. They improved the efficiency of sensor systems by their architectural flexibility, re-configurability, processing capabilities, interfaces. They also perform many simple and complex functions [14][15]. These types of sensors are usually used for high-speed applications that need online measurements.

In many monitoring applications, sensor nodes route the aggregated data to a few number of sink nodes. In this case, transmitting unnecessary redundant data will cause a waste of communication channel bandwidth and energy consumption. The performance of data collection in a wireless sensor network will decrease when every node sends all collected data to the sink node. Moreover, this will increase collisions in the transmission channels. In order to increase the power of computation of sensor nodes, Field-Programmable Gate Arrays (FPGAs) and soft computation strategy such as Neuro-Fuzzy algorithms can be used within node architecture to overcome the limitation of traditional microcontrollers.

In our work, we present a new methodology to build and develop multiple soft-core processors of the FPGA set that perform ensemble soft computing algorithms in an IoT environment based on wireless sensor network. Ensemble technique has been proposed to design and build parallel embedded soft core FPGA microcontrollers within the reconfigurable sensor hardware part that can implement different types of soft computing algorithms. The ensemble of soft computation algorithms such as ANFIS algorithms offer more successful and highly efficient techniques rather than separate algorithms. Our work offers multi-functions and multi-sensing data to measure and deal with more than one physical signal simultaneously like temperature, humidity. Researchers recently use ensemble techniques to perform many tasks such as, control systems, self-testing error correction, weather forecasting [15], data classification, and prediction of energy consumption.

ACTA INFOLOGICA
ISSN: 2602-3563

2. ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM

In this approach, artificial neural networks are used to train some parameters of the fuzzy control system in an iterative way. This technique refers to a hybrid neuro-fuzzy system. In this technique, artificial neural network and control fuzzy system are performed together by keeping their different functions. This type of merging is suitable for control application, decision making, pattern recognition applications and classification. The idea of such hybrid model is to use a gradient-based learning algorithm in neural networks to determine its parameters through input and output patterns, so this will enable us to have advantages of learning through different patterns [11]. One of the most common examples that can be used in many different applications of Hybrid Neuro-Fuzzy systems is Adaptive-Neuro-Fuzzy Inference System (ANFIS).

ANFIS is one of the most important schemes that integrates the advantages of both neural and fuzzy systems in one model. In ANFIS, all parameters associated with linear and non-linear equations, number of rules and architecture of fuzzy system can be adapted by neural network algorithms. ANFIS algorithm has a very high learning speed and gives high accuracy in testing phase. Also, it has the ability to use human expressions (linguistic terms) when using to solve real life complex problems. It is simple to realize and can be used to control different application fields [12].

Normally ANFIS is used to map input pattern to membership functions (MFs), output of MF to a set of "if-then rules", rules to a set of output sets, output sets to output of linear MFs, and the output MFs to a single valued output or a decision associated with the output. A typical ANFIS structure is depicted in Figure 1.

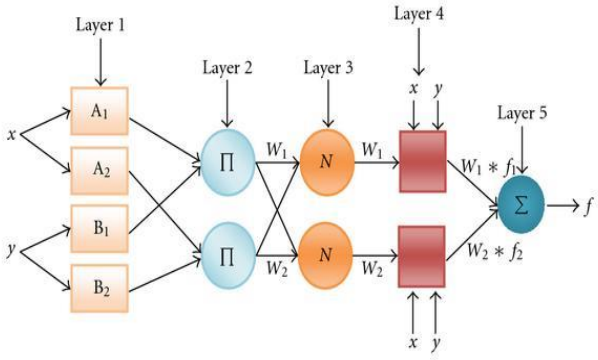


Figure 1. Adaptive Neuro-Fuzzy Inference System

According to figure 1, ANFIS structure normally has 5 layers of neurons [12]:

Layer 1: Each node generates the membership degree of linguistic terms that are given by the following Eq. 1 and Eq. 2:

$$Q_{1,i} = \mu_{Ai}(x), \quad i = 1,2,3, \qquad (1)$$

$$Q_{2,i} = \mu_{Bi}(y), \quad i = 1,2,3, \qquad (2)$$

Where "x" and "y" are the inputs, "I" is number of the nodes, and A and B are the linguistic terms. Eq. 3 is an example of a membership function that is bell-shaped function:

$$\mu_{Ai}(x) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2bi}}, \quad i = 1,2,3, \qquad (3)$$

As we can see, the bell-shaped function has three parameters a, b, and c that are called premise parameters.

ACTA INFOLOGICA
ISSN: 2602-3563

Layer 2: Each node calculates the firing strength of each rule using the "min" or "prod" operator. Normally, AND operation is the most commonly used operation in fuzzy systems. The output of this layer can be as the following Eq. 4:

$$Q_{2,i} = w_i = \mu_{Ai}(x) * \mu_{Bi}(y), \quad i = 1,2,3, \qquad (4)$$

Layer 3: In this layer, normalized process is done. Every node in this layer calculates the proportion of each rule to the aggregate of all conditioning rules in previous layer. The number of the normalization rules is equal to the number of rules in layer 2. In general the formula in Eq. 5 can be used to represent the output of this layer.

$$\boldsymbol{Q_{3,i} = \bar{w}_i = \frac{w_i}{w_1+w_2}}, \quad i = 1,2,3, \qquad (5)$$

Layer 4: The outputs of this layer are the product of the normalized firing strength coming from layer 3. The nodes compute a parameter function on the layer 3 output. Parameters in this layer are called consequent parameters and can be represented as Eq. 6.

$$\boldsymbol{Q_{4,i} = \bar{w} f_i = \bar{w}(p_i x + q_i y + r_i)}, \ i = 1,2,3, \qquad (6)$$

Where "p", "q" and "r" are called consequent parameters.

Layer 5: In this layer, a single node aggregates all incoming signals from layer 4 to calculate a single output as in Eq. 7.

$$Q_5 = \sum \bar{w} f_i \ , \ i = 1,2,3, \qquad (7)$$

Training and Testing in ANFIS begin by dividing the input data sets into training, testing and checking groups. The training data are normalized to be suitable for the training process using the Min, Max method. It's done by mapping each input value between binary 00, 01 and 10 numbers. After that, the forward path which is the first phase in hybrid algorithm will start to initialize and train the consequent linear parameters in layer 4 based on Linear Regression algorithm. After the Linear Regression process finished, the second phase in the hybrid algorithm which is based on back-propagation algorithm will begin training the nonlinear promise parameters which find in layer 2.

The finishing of training process for the ANFIS can be done by two methods. In the first method, training ANFIS ends when the threshold error is less than the specific limit which is defined at the beginning of the training phase. In the second method, ANFIS stops learning after a specific number of learning iterations. In our case, the ANFIS algorithm is stopped learning based on the limited number of iterations.

3. EXISTING STUDIES

A With the advancements of using reconfigurable FPGA and soft computing techniques within WSNs, many studies have been done on this field. The flexibility and efficiency of FPGA devices motivate the researchers to use them in many soft intelligent techniques such as [3]. The authors used FPGA to implement an adaptive neuro-fuzzy system based on radial basis neural network to reduce and optimize the hardware resources with the fixed-point and simplified floating-point arithmetic.

Another study by [8] shows the ability of implementing the meta-heuristic learning algorithms of neuro-fuzzy system (NFS) on the FPGA based on improved particle swarm optimization (iPSO). According to the authors, Virtex5 FPGA board was used to implement the neuro-fuzzy system. Particle swarm optimization technique was used to train and update neuro-fuzzy parameters. The researchers used this kind of integration to speed up the training algorithm and minimize the usage of hardware resources such as size of memory and number of multipliers.

Another FPGA implementation of Adaptive Neuro-fuzzy Inferences Systems (ANFIS) for controlling temperature and humidity data sensors inside a greenhouse using geothermal energy as a power source for heating system was proposed in [9]. Four techniques are used in this approach; artificial neural network, PI control, fuzzy logic control,

ACTA INFOLOGICA
ISSN: 2602-3563

and adaptive neuro-fuzzy control to control the indoor temperature. The main advantages of using the hardware language approach are rapid prototyping and allowing usage of powerful synthesis controller through the use.

In [5], the implementation on reconfigurable hardware of a Sugeno type for adaptive Neuro-Fuzzy inference system is proposed. The pipeline and parallel pipeline architecture play an important role in modelling the algorithm for the FPGA based implementation. They used Matlab Simulink toolbox to simulate the proposed controllers, implement and simulate the adaptive Neuro-Fuzzy inference system algorithm using FPGA boards based on both high level synthesis tool and VHDL coding. The pipeline structure of the proposed hardware was designed in a way that permits the parameters to update in parallel. The hardware model's processing speed is very high, and allows the controller to be used in real-time applications.

In [19], the adaptive neuro-fuzzy model was implemented based on FPGA.In this article, High Level Synthesis (HLS) tool with C/C++ language from Xilinx was used to design the embedded soft core within FPGA fabric. The authors describe some of optimization methods to implement and realize the adaptive neuro-Fuzzy algorithm. In this article, The authors Use a high level language (C-language) instead of Hardware Description Language (HDL). So, the information about hardware details are not needed. According to the authors, using such kind of embedded IP core can be used for variety real time applications.

In addition to the control systems that appear in previous works, hybrid neuro-fuzzy can be used for power prediction system [13][1]. The authors proposed a power prediction system for wind energy generation. In this work, the sensor networks are used to perform some measurement and send the results to the main station. The main station performs many analysis to the sensing data based on data mining algorithms like Fuzzy C-Means algorithm (FCM) which is used to estimate the fuzzy rules for hybrid neuro-fuzzy systems [1]. The prediction system can be used for Damage prediction for wind turbines using both wireless sensor and actuator networks [13]. Fuzzy control system is used to remove the effect of overheating by forecasting the damage in wind turbines. The authors in this work use the adaptive Neuro-Fuzzy inference system to build a real database based on real temperature, frequency, and the effect of temperature on the natural frequency values. Then, using this database to enhance the performance of Fuzzy control system.

In [7], an online learning recurrent Neuro-Fuzzy classifier algorithm is proposed for use in classification applications. The recurrent network is embedded in the RNFC by adding feedback connections in the second layer, where the feedback units act as memory elements. They show that effective neuro-fuzzy classifier should be able not only to adaptively adjust fuzzy membership functions but also to dynamically adapt fuzzy operators.

In [6], a complete hardware and software system was designed based on embedded FPGA device to develop a soft controller for intelligent environments. A group of embedded IP cores are used to build neuro-fuzzy architecture which represents the hardware portion of the system. MicroBlaze core processor that represents the software part used to drive and control the operation for overall system. Their approach is simplifying an ANFIS model in order to reduce the computational costs of the algorithm and to promote the scalability and modularity of its HW implementation. In addition, we exploit the redundancy inherent in ambient-intelligence data in order to reduce the dimensionality of the system. A high speed, scalability and efficiency can be achieved by using FPGA devices compared with other implementation based soft computing mechanisms. Another implementation of neuro-fuzzy inference system based on FPGA Zynq System on Chip development board with a dual-core ARM Cortex A9 processor is presented in [18]. The ANFIS algorithm is designed and implemented on the FPGA based on the VHDL language. The algorithm is used to correct the corner displacements of the vehicle.

ACTA INFOLOGICA
ISSN: 2602-3563

4. METHODS

The architectural flexibility of FPGA provides high efficiency such as parallelism, and the performance through the development of algorithms. So a new generation of reconfigurable, smart, and low level of energy consumption sensors is developed based on FPGAs. Figure 2 explains the general architecture of sensor node based FPGA [10].

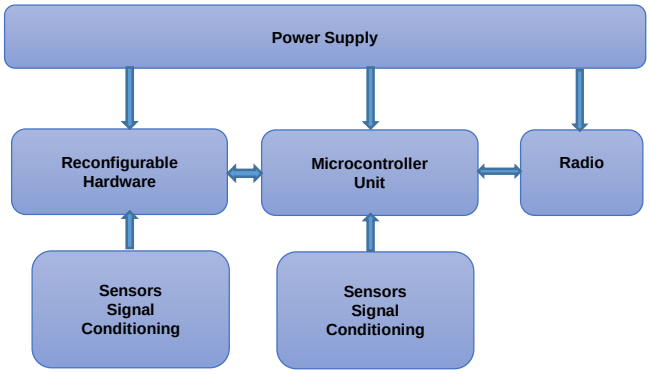


Figure 2. Architecture of FPGA with sensors

In our work, reconfigurable part within a wireless sensor node is used for implementation of multiple and different applications to control of sensing data, decision making, analysis and advanced data management based on ANFIS algorithm. ANFIS algorithm is the best Neuro-fuzzy algorithm that is used to control the nonlinear applications. That's why we will use the ensemble technique to design and build parallel embedded soft FPGA based MicroBlaze microcontrollers within the sensor reconfigurable hardware part. The block diagram in Figure 3 shows the proposed reconfigurable part:

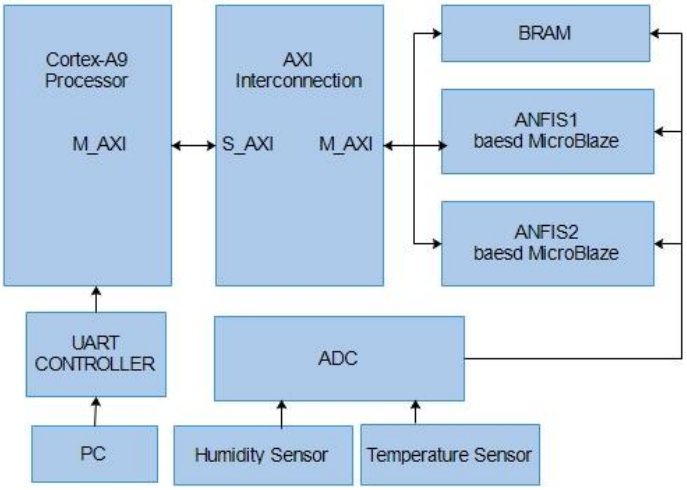


Figure 3. Proposed Reconfigurable Architecture (System)

The ensemble of the design system in figure 3 is composed of hardware and software part. The basic hardware part was designed using Vivado 2016.4 – Block Design which is an important tool from Xilinx to start creating the design system within Zynq®-7000 AP SoC ZC702 Evaluation Kit. The hardware part consists of two ANFIS algorithms based on MicroBlaze soft-core processors. The first ANFIS algorithm is connected to temperature sensor and other algorithm is connected to humidity sensor with their associated BRAM (Block Random Access Memory), UART (Universal Asynchronous Receiver Transmitter) controller, and ADC (Analog to Digital Convertor) controller.

ISSN: 2602-3563

The Xilinx Advanced eXtensible Interface (AXI) which is a kind of smart microcontroller buses within FPGA board is used to join and aggregate the outputs of hardware algorithms (slave AXI) to the main processor (master AXI).

The software part is done using Xilinx Software Development Kit (SDK 2016.4) is represented by compiler, libraries, and application programs. The hardware system is programmed using C language to implement the developed ANFIS algorithms in our hardware. Both hardware and software platforms are transferred into a bit file which is ready to be downloaded on the FPGA device to create the system. Figure 4 summarizes the sequence of operations that did by FPGA processor (Cortex-A9).

Figure 4. Sequence of operations main processor

ACTA INFOLOGICA
ISSN: 2602-3563

The Cortex-A9 processor starts to initialize all drivers and hardware cores, stores the collecting data to BRAM, and then performs the data processing operations that involve:

- Data classification.
- Data normalization
- Data fuzzification
- Data routing: In next step, the processor routes each data set to the equivalent MicroBlaze cores and then, reading backs the results in order to send them to the host computer via serial communication.

5. FINDINGS

In Matlab environment, most of researchers try to use a graphical user interface (anfisedit-GUI) for ANFIS algorithm or a function provided by a Matlab Toolbox, which is "anfis-command" to construct and do the testing and training that makes it simple to use and easy to follow.

In our approach, we've modified ANFIS algorithm and combined it with gradient descent optimization algorithm during learning phase to minimize the error, prevent stuck in local minima and reach a global minima. This optimization algorithm is known as Momentum that speeds up the steps that is taken towards the optimal solution.

In Momentum, a fraction "m" of the previous weight update is added to the current weight. As a result, we provide a faster convergence and speed up the training process. This could not be realized if we apply ANFIS function provided by Matlab Toolbox. In figure 5, the pseudo code explains how to add the momentum factor into the ANFIS algorithm.

Set the type of membership function
*(**Gaussian function**)*
*Load initial premise parameters (**a,b,c**)*
*Load initial consequent parameters (**p,q,r**)*
*Load initial momentum factor (**m**)*
Input the sensor data
Normalize the data

// start forward path

While** (error > threshold) **then

Generates the membership degrees for input
Normalize and aggregate generated data
Calculate the output of the algorithm
Updata consequent parameters
$p_{new}=p_{old}$ ***+ delta(error) * input data***
$q_{new}=q_{old}$ ***+ delta(error) * input data***
$r_{new}=r_{old}$ ***+ delta(error)***

// start backward path

// update premise parameters
$a_{new}=a_{old}$ ***+ delta(error) + momentum factor (m)***
$b_{new}=b_{old}$ ***+ delta(error) + momentum factor (m)***
$c_{new}=c_{old}$ ***+ delta(error) + momentum factor (m)***

End while loop
De-normalize the output data

Figure 5. Pseudo code of the proposed algorithm

ACTA INFOLOGICA
ISSN: 2602-3563

Two types of data sets (temperature and humidity) taken from Beach Weather Stations - Chicago Park are used for both our modified code with momentum factor and Matlab Toolbox for ANFIS [20][21]. Each data set consists of 13917 samples from 5/22/2015 to 12/31/2015. Figure 6 shows the effective of adding the momentum factor to the ANFIS algorithm.

According to the Figure 6, the mean square of error with momentum factor is (0.9058) after 40 iterations, while as equal (23.0417) after the same number of iterations.

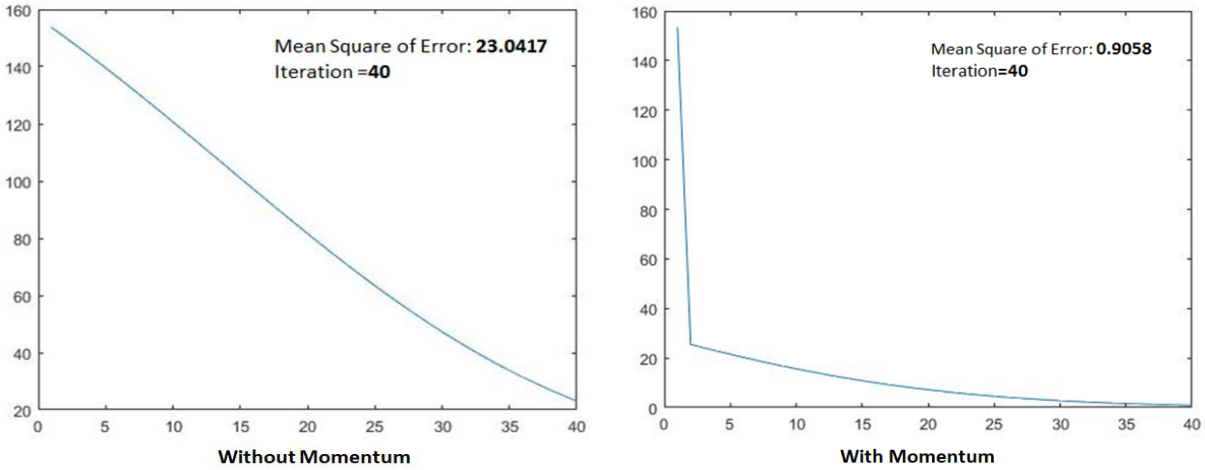


Figure 6. Training and Testing results for Temp. data set %70 training and %30 testing.

Table 1 and Table 2 show the error after training and testing the data sets.

Table 1. Training and Testing results for Temperature data set %70 training and %30 testing.

Technique type	Training	Testing
Matlab Toolbox	3.7802e-11	0.013204
Modified Algorithm	1.3943e-14	0.0010988

Table 2. Training and Testing results for Humidity data set %70 training and %30 testing.

Technique type	Training	Testing
Matlab Toolbox	9.3506e-11	0.062509
Modified Algorithm	7.4105e-15	0.001184

while the table 3 and 4 show the effect of adding the momentum factor on speed in training and testing phases.

Table 3. Training Speed (Execution Time) in Sec.

Execution Time in Sec (Training)		
Data Type	ANFIS Toolbox	Modified Algorithm
Temperature	3.675	1.1921
Humidity	3.6427	1.1895

Table 4. Testing Speed (Execution Time) in Sec.

Execution Time in Sec (Testing)		
Data Type	ANFIS Toolbox	Modified Algorithm
Temperature	0.0124	0.011
Humidity	0.0122	0.012

According to table 3, we can easily see that adding the momentum factor to the modified algorithm reduces the execution time for training phase about 60% compared with Matlab toolbox algorithm in both temperature and

ACTA INFOLOGICA
ISSN: 2602-3563

humidity data sets. While the execution time in table 4 for testing phase is almost the same for both algorithms because there are no iteration processes happened in both algorithms under the test condition.

6. DISCUSSION AND CONCLUSIONS

In this paper, implementation of adaptive Neuro-Fuzzy algorithms using FPGA was proposed. Ensemble of the multiple ANFIS algorithm is used in order to improve the efficiency and accuracy for a sensor node. Because of using different ways to train and simulate Ensemble ANFIS within a single wireless sensor node, we generate a kind of intelligent system. In our proposed method, pre-processing the data collected from different sensors is possible and can be used in diverse applications such as prediction, data cleaning, data smoothing and data compression. As a result, we can eliminate unnecessary data and reduce the traffic on transmission channels.

On the other hand, two methods were applied to simulate the temperature and humidity data sets: a Matlab Toolbox and modified algorithm with momentum factor for ANFIS. Using modified algorithm with momentum provides a best and fast convergence during training process. Our modified algorithm is also more flexible and can be modified to combine with other optimization algorithms. The results have shown the efficiency of using the momentum factor with ANFIS algorithm. In the future, we will try to combine a modified ANFIS with different types of soft computation algorithms for the sensor node.

REFERENCES

[1] Ahmed, E.,Mohamed, S., Khaled, M., Ahmed, A., (2016), A hybrid neuro-fuzzy power prediction system for wind energy generation, International Journal of Electrical Power & Energy Systems,74, 384-395.

[2] Akyıldız, L., Sankarasubramaniam, Y., Su, W., Cayırcı, E. ,(2002), "Wireless sensor networks: A survey", Journal of Computer Networks, 38, 393-422.

[3] Andrzej, P., Meng, J. ,(2016), "The method of hardware implementation of fuzzy systems on FPGA", International Conference on Artificial Intelligence and Soft Computing, ICAISC 2016, 284-298.

[4] Brassai, S., Hajdu, S., Tamas, T., 4(2015), "Hardware implementation of a neuro-fuzzy controller using high level synthesis tool", Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics, 10.1515/macro-2015-0018.

[5] Brassai, S., Hajdu, Sz., Tamas, T., Bakó, L., (2015), "Hardware implemented adaptive neuro-Fuzzy system", Carpathian Control Conference, 10.1109/Carpathian CC.2015.7145046.

[6] Campo, I., Basterretxea, K., Echanobe, J., (2012), "A System-on-chip development of a neuro–fuzzy embedded agent for ambient-intelligence environments", IEEE Transactions on Systems, 10.1109/TSMCB.2011.2168516.

[7] Cheng-Jian, L., Chun-Cheng, P., (2014), "Classification using an efficient neuro-fuzzy classifier based on adaptive fuzzy reasoning method", Conference on Computer, Consumer and Control, 10.1109/IS3C.2014.34.

[8] Cihan, K.arakuzua, F., Mehmet, A. ,(2016), "FPGA implementation of neuro-fuzzy system with improved PSO learning", Journal of the International Neural Network Society,79- 2016, 128-140).

[9] Doaa, M., Hanaa, T., (2017), "Analysis and design of greenhouse temperature control using adaptive neuro-fuzzy inference system", Journal of Electrical Systems and Information Technology, 4(1), 34-48.

[10] François, Ph., (2014), "Runtime hardware reconfiguration in wireless sensor networks for condition monitoring", Universitäts - und Landesbibliothek, Darmstadt.

[11] Janusz, K., (2000), "Introduction to neuro-fuzzy systems", Advances in Soft Computing. Springer-Verlag, ISBN 978-3-7903-1852-9. Berlin/Heildelberg.

[12] Janusz, K. (2002), "Neuro-fuzzy architectures and hybrid learning", Springer-Verlag,ISBN 978-3-7903-1802-4, Berlin/Heildelberg.

ACTA INFOLOGICA
ISSN: 2602-3563

[13] Maicon, M.,Luci, P., Silvana, R., Flavia, C., Claudio, M., Paulo F., Albert, Y., (2017), "Damage prediction for wind turbines using wireless sensor and actuator networks", Journal of Network and Computer Applications. 80,123-140

[14] Meena, S. & Krishna, N., (2014), "Simulation of dynamically reconfigurable wireless sensor node", International Conference on Electronics and Communication System, Coimbatore, 10.1109/ECS.2014.6892795. India.

[15] Melin, P., Soto, J., Castillo, O., Soria, J., (2013), "Time series prediction using ensembles of ANFIS models with genetic optimization of interval type-2 and type-1 fuzzy integrators", Journal of Hybrid Intelligent Systems, 10.3233/HIS-140196.

[16] Potdar, V., Sharif, A., Chang, E., (2009), "Wireless sensor networks: a Survey", International Conference on Advanced Information Networking and Applications Workshops, 10.1109/WAINA.2009.192. UK, Bradford.

[17] Rajasekaran, C., Jeyabharath, R. and Veena, P., (2014), "Hardware-software reconfigurable techniques for wireless sensor network", Journal of Applied Sciences, Engineering and Technology 8(17): 1855-1862.

[18] Rasika, A. and Pathan, M., (2015), "Design and implementation of ANFIS algorithm using VHDL for vehicular system", Journal on Recent and Innovation Trends in Computing and Communication 3, 2: 820-824.

[19] Sándor, T., Szabolcs, H., Tibor, T., (2015), "Embedded adaptive neuro-fuzzy inference system with hardware implemented real time parameter update", International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics, 10.1515/macro-2015-0021.

[20] Chicago. , (2015). "Beach Weather Stations - Automated Sensors - 2015 - Air Temperature | City of Chicago | Data Portal". [online] Available at:https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated- Sensors-2015-Air-/rsk3-iyyk

[21] Chicago. , (2015). "Beach Weather Stations - Automated Sensors - 2015 - Humidity | City of Chicago | Data Portal". [online] Available at: https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors-2015-Humi/4tf5-5fw5

This work was presented at the 4th International Management Information Systems Conference and published in the conference abstract book

ACTA INFOLOGICA, 2018; 2(1): 33-44
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

Validity Issues in Linked Data Driven IS Research

Ziya Nazım Perdahçı*, Mehmet Nafiz Aydın, Kenan Kafkas

ABSTRACT

This research adopts a complex system approach to linked data, which has a trace aspect and to examine validation issues in linked data driven IS research. Thereby a relevant question arises: What are the validity issues in the overall network analysis process applied on such linked data? This research argues that validity issues are vital to research in linked data and requires a complex system approach so that true value of linked data can be discerned and applicable to the real-world cases. Particular emphasis is placed on the validation issues in empirical research on linked data concerned with the educational system. This paper should be considered as a contribution to the efforts of those who are struggling with the validity issues in SNA. The intention of the work is to build a checklist that can be used to check the validity of the data, methods, and algorithms for transdisciplinary research teams who utilize theory of networks in general and SNA in particular in a particular domain, which is an educational system for the focus of this research. The findings may help the school administrators, instructors and student advisors in the decision making processes.

Keywords: Trace Data, Social Network Analysis, Network Science.

İlişkisel Verilere Dayalı Bilişim Sistemleri Araştırmalarında Geçerlik Konuları

ÖZ

Bu araştırmada ilişkisel veri odaklı Bilişim Sistemleri araştırmalarında görülen geçerlilik sorunlarını incelemek amacıyla, ilişkisel veri karmaşık sistem yaklaşımı benimsenmiştir. Bu durumda akla şu soru gelmektedir: Buna bezer ilişkisel veri üzerinde yapılan genel ağ analizi çalışmalarında geçerlilik sorunları nelerdir? Bu araştırmada ilişkisel veri araştırmalarında geçerlilik sorunlarının hayati derecede önemli olduğu ve ilişkisel verinin gerçek değerinin anlaşılması için karmaşık sistem yaklaşımının gerekli olduğu savunulmaktadır. Özellikle eğitim sistemleri ile ilgili deneysel araştırmalarda geçerlilik sorunları üzerinde durulmuştur. Bu çalışma Sosyal Ağ Analizinde geçerlilik sorunlarıyla karşılaşanların çabalarına katkı olarak düşünülmelidir. Disiplinlerarası çalışmalarında ağ teorisi özellikle de eğitim alanında Sosyal Ağ Analizi kullanan araştırma ekipleri için veri, metot ve algoritmaların geçerliliğini kontrol etmek amacıyla kullanılacak bir liste oluşturmak hedeflenmektedir. Elde edilen bulgular okul yöneticileri ve öğretmenlere karar verme süreçlerinde yardımcı olabilir.

Anahtar Kelimeler: İz Verisi, Sosyal Ağ Analizi, Ağ Bilimi.

Information of Author(s):

Ziya Nazım Perdahçı
ORCID: 0000-0002-1210-2448
nz.perdahci@msgsu.edu.tr
Mimar Sinan Fine Arts University

Mehmet Nafiz Aydın
ORCID: 0000-0002-3995-6566
mehmet.aydin@khas.edu.tr
Kadir Has University

Kenan Kafkas
ORCID: 0000-0002-1034-569X
kenankafkas@gmail.com
Kadir Has University

DOI: 10.30801/acin.356598

Submit Date: 21.11.2017
Accept Date: 22.02.2018
Publish Date: 26.06.2018

(*) Contact Author
Address: Mimar Sinan Fine Arts University, Department of Informatics, Bomonti, İstanbul, Turkey
Telephone Number: +90 212 246 00 11 Ext:6102

ACTA INFOLOGICA
ISSN: 2602-3563

1. INTRODUCTION

In the last decade in almost every field, data have become abundant, more accessible, and more diverse. For companies as well as academics, combining enterprise-wide data with open data to generate business intelligence brings up new opportunities and challenges (Behrendt et al., 2014). Recently, the term "linked data" is suggested to refer to bringing together all relevant digital data on the Internet for the sake of open data integration (Dong and Srivastava 2015). The current work extends the very idea of linked data from a typical integration context to trace data, which reveals the business context and complex relations of the things and their interactions. This aspect is crucial to make use of both enterprise and open data where the notion of "linked" emphasizes what and how data are derived from business context. In this regard, this research adopts a complex system approach to linked data, which has a trace aspect and to examine validation issues in linked data driven IS research.

The trace data present in Information Systems has certain characteristics. Among other types there is a data type called event-based data, which is often times enabled by conventional transaction information systems. The events mentioned here are usually records of various interactions between at least two entities. The recorded data turns up to have a linked structure and as a result of this linked structure a complex system emerges. To examine these complex systems, it is necessary to apply network science. Thereby a relevant question arises: What are the validity issues in the overall network analysis process applied on such linked data? The value and importance of the digital trace validation becomes immediately clear, taking into consideration the current studies (Jungherr 2015).

Howison et al. (2011) articulate the validity issues in network analysis of digital trace data and propose a number of issues that researchers should take into account. Addressing validity issues is vital to research in linked data and requires a complex systems approach so that true value of linked data can be discerned and applicable to the real-world cases. To better articulate the validation issues in a real-world context we utilize empirical research on linked data in a school information system as a case.

Although almost all the metadata in network studies in education comes from school management information systems, the crucial linked data is obtained generally by means of face to face surveys. Therefore, Social Network Analysis (SNA) in education does not completely rely on trace data. However, most of the methods and an important portion of the data are common in both papers. For this reason, validity issues match for both. This paper should be considered as a contribution to the efforts of those who are struggling with the validity issues in SNA. The intention of the work is to build a checklist that can be used to check the validity of the data, methods, and algorithms for transdisciplinary research teams who utilize theory of networks in general and SNA in particular in a particular domain, which is an educational system for the focus of this research.

Educators quickly adapted this situation and began to involve data more often in their decision making processes. Data Driven Decision Making (DDDM) concept is introduced in education (Marsh et al., 2006). Many advanced software emerged to meet the needs of educators. Learning Management Systems and School Management Information Systems became widely used in schools by both governments and private enterprises. Management Information Systems (MIS) are being used by schools to support a range of administrative activities including attendance monitoring, assessment records, reporting, financial management, and resource and staff allocation (O'Brien, 1998). As a result, a huge amount of data is collected and stored in relational database systems. Now, from governments to private sector, the education administrators and instructors rely on the information that is obtained by analysis of the data.

Although individuals play a key role in education, they are not isolated entities. Therefore, interactions among individuals also provide valuable information. This type of linked data requires a specific analysis, namely Social Network Analysis. Statistical data analysis in classical sense lacks the ability to capture the essence of complex systems that emerge from intricate human relationships. With SNA algorithms in a school environment for instance, key players in the network can be found or correlations between certain attributes of students can be calculated. The data necessary for this type of analysis may be the friendship ties among students or collaboration ties among groups. Additionally, SNA requires the data that is available in the traditional School Information Systems. For instance, a typical analysis would include the study of the friendship relations among

ACTA INFOLOGICA
ISSN: 2602-3563

students, involving calculation of the correlations between success rates and friendship preferences. This requires combining the relationship data with metadata, referred to as attributes such as gender or test scores. The findings may help the school administrators, instructors and student advisors in the decision-making processes.

2. METHOD AND BACKGROUND

The method of this study involves three steps: First, a model explaining general validity issues linked data driven IS research (Howison et al., model). Second, a specific research case (SNA in education) examplifying and elaborating the issues that might be encountered. Third, a framework on which the issues are explained in detail. Similar to the research design adopted in (Howison et al., 2011), we frame our study of validity issues respect to the decision matters researchers face with in network analysis of linked data driven IS research.

The growth of data sources produced on online interactive platforms have drawn significant attention from IS researchers, but the validity issue in SNA in IS research context has remained an open issue (Whelan et al., 2016). Scholars with exception of (Howison et al., 2011) have addressed this issue within their own research contexts. For instance, (Nia et al., 2010). have examined the validity of Network Analysis in open source projects. For our research purpose, we need a model that should achieve theoretical cohesion (full chain of reasoning across all the phases in SNA), and provides researchers with meta-level issue analysis (by raising abstraction level to overcome limitations of case specific results). In line with these reasons, we adopted the model proposed by (Howison et al., 2011). The model (Figure 1), proposed by (Howison et al., 2011) is composed of six elements connected by five links raising ten issues. These links are the transition areas between steps starting from Information System to the research construct. First, when working with a digital trace data particular attention is to be paid to the information system producing that data. The misuse of the system can cause misinterpretations of the collected data. Another issue that should be considered in this phase is a reliability of the data generated. In the next step, the complication of converting digital trace data into nodes and links should be solved. In order to do that, the researcher should make one of the most crucial decisions, which is determining the type and intensity of the links, as well as deciding on the missing links. Creating a network where the order of the events matter can raise a problem of temporal aggregation. In another step, while using a network to obtain some measures, a researcher should address network tool effects and temporal mismatch. There is a large selection of software tools available for social network analysis (SNA), thus choosing proper software for an analysis is an important step, since these tools can help researchers with avoiding errors as well as at the same time can threaten to validity in their use. The temporal mismatch issue can be addressed by deciding the period of time over which measures derived from that network will be measured. The last issues a researcher should consider emerge when aligning a measure and a construct are data completeness and inference and inappropriate importation.

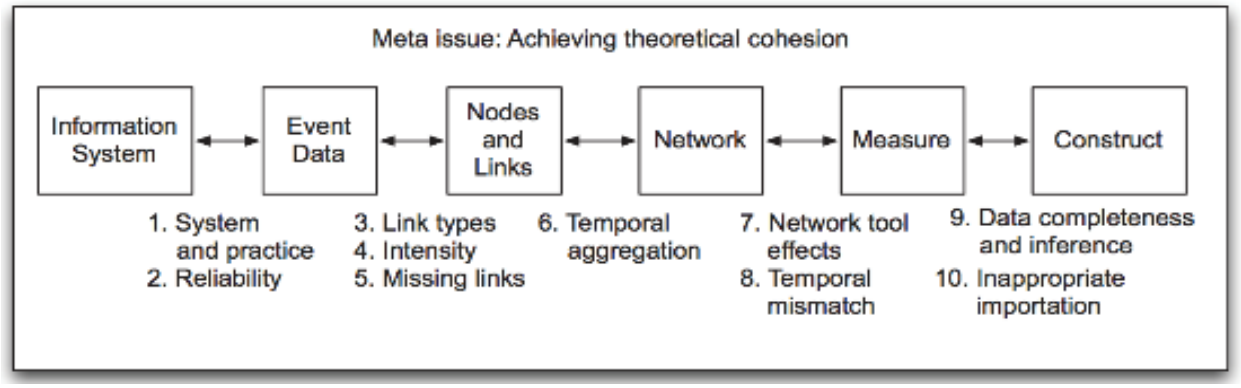


Figure 1. Howison et al. (2011) introduce five links in the chain of reasoning and corresponding validity issues to achieve theoretical cohesion.

Howison et al. (2011) state that "In practice, the process of achieving alignment between a theoretical context and the chain of reasoning underlying valid measurement is an iterative one, most likely involving multiple

ACTA INFOLOGICA
ISSN: 2602-3563

adjustments and decisions and revisiting these to achieve a cohesive logic." Thus, applying the model (Figure 1) has been an iterative process, which brought out three phases.

2.1. Social Network Analysis in Education

From the Network Science perspective, a network consists of two types of simple components, nodes and links. Nodes may represent an individual in a social network or an enzyme in a cell. The connections between nodes are called the links. They may represent kinship between individuals in social network or chemical interaction between enzymes in a cell. The term graph refers to mathematical representation of a network. It is analogous to a wiring diagram. The terms network, node and link are mostly used for referring real world complex systems whereas the terms graph, vertices and edges are used when referring to a mathematical representation of the real-world systems. These are only subtle differences and these terms are often used interchangeably (Barabási and Pósfai, 2016).

A node can have more than one link. Total number of links of a node is called its degree. If links in a network have distinct direction from one node to another, this type of network is called a directed network. In an undirected network links do not have directions. There are two types of degrees in directed networks; in-degree which is the number of links pointing towards a node, out-degree that is the number of nodes pointing out from a node

Social Network Analysis requires node and link data and in education networks, this corresponds to nodes being students or teachers and links being the relationships among nodes. The metadata about the nodes are called node attributes, which can be any data that range from the name or address of the student to test achievement scores or the name of the course taken. The link data is type of data, which defines a relationship between two nodes for instance, if two students study together or take the same course, the two nodes representing the students are linked in the graph. These data are mostly available in School Management Information Systems; however, additional data can be gathered by surveys. After collection, the data is prepared for analysis. The next stage is modelling. Deciding the types of nodes and the links is called modelling. Figure 2. shows a simple model of a network where nodes are students and links are drawn if the two students are best friends.

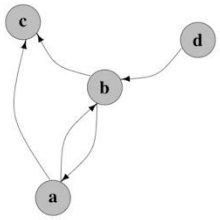


Figure 2. A simple directed network model.

In other words, if student "a" claims that student "b" is his or her close friend, then an edge with an arrow pointing from node "a" to node "b" connects them. Looking at this toy model closely reveals the fact that student "a" accepts student "c" as a close friend, student "c" however, does not perceive him or her as close friend. In these kinds of networks, links have directions which makes the network directed network. In the organization science literature, such a directed network model for friendship has not been examined extensively (Smirnov and Thurner, 2017).

ACTA INFOLOGICA
ISSN: 2602-3563

3. FINDINGS FOR PROPOSED FRAMEWORK

3.1. Framing Validity Issues in Linked Data in Network Science

In this study, we present a framework (see Figure 3) that illustrates general structure of a typical scientific research involving linked data along with digital trace data. Our intention is to address the validation issues that may arise during the research process. The framework contains flow of data through steps of research up to the scientific output. One can see that the framework contains a number of feedback loops to enhance scientific rigor and validation. In essence, it is an operational means to support research inquiry by incorporating relevant theoretical accounts. Noticably, linked data is considered from a complex system point of view, which allows us to bring theory of network and information system research together. In doing so, the scientific output may include descriptive, predictive, and prescriptive IS which in turn enhance researchers to provide feedback to information systems.

Real World System

Information systems collect data from the system, which is in fact the Real World. However, interactions between them is not one way. That is to say, the system and IS constantly interact and shape each other. Customers interacting with sales representatives, employees interacting with each other or the environment, students working mutually on a project, are examples of such activities in a real-world system.

Information System

Information Systems keep track of these activities and a huge amount of data cumulate over time. This event-based trace data has a linked nature. The interactions leave trace of events, which can later be collected. Since events take place among entities of the system, the entities can be linked to each other and this phenomenon can be represented as a network.

Linked Data

Two different types of data are obtained from two different systems. Offline enterprise data is gathered from the IS and depending on the case, ground truth data or open data is gathered from the real-world system. For instance, ground truth may correspond to various observed relationships among students in school environment. On the other hand, open data may correspond to other social interactions gathered out of school premises.

The resulting network is a complex system containing substantial number of interacting components. To examine such systems, network science approach is required.

Network Science and IS Research

The necessary steps to analyze these networks is explained in SNA process section. This section covers the network analysis particularly the Social Network Analysis in order to limit the scope of the paper.

The Scientific Output

The scientific output of the entire process is the description of the system, predictions towards the future and prescriptions of the IS problems. Finally, these outputs are sent to the IS as a feedback.

During the transitions of each step of the process, researchers should check the validity of the data, methods and findings. The validity issues section addresses the possible validity issues and explains them in detail.

ACTA INFOLOGICA
ISSN: 2602-3563

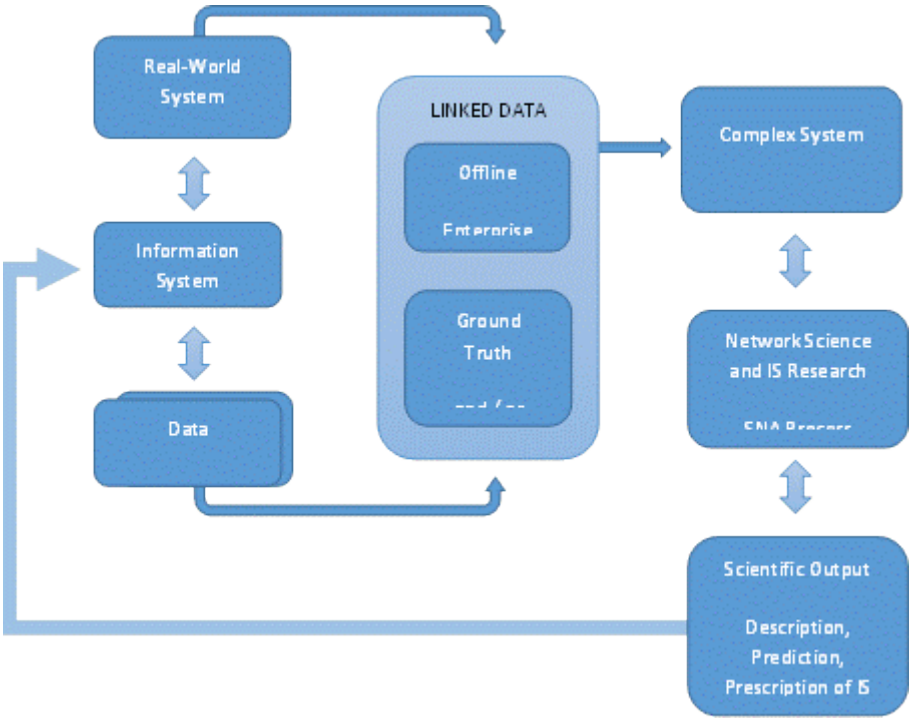


Figure 3. Framework for Validation Issues in Linked Data Driven IS Research

3.2. Network Analysis of Linked Data

The SNA Process

Four basic steps are taken into consideration when approaching a social network research problem from Network Science perspective: Data preparation, network modelling, network analysis and interpretation (Aydın and Perdahçı, 2014). Researchers follow these steps when looking for answers to their research problems. The process has an iterative structure. The data is obtained from various sources depending on the necessities of the social network that is subject of the study. In the literature, self-reported friendship data collection is one of the well-known techniques (Labun et al., 2016). These data are then prepared to suit the requirements of the analysis and visualization tools so that they can be appropriately imported into working environments.

Following the data preparation, the modelling step takes place where the nodes and the links of the network are decided. After this step, analysis begins which is mainly visualizing the constructed network and applying suitable SNA algorithms. The findings that are obtained in this step are scrutinized and the process moves to the next step. In the interpretation step, if the findings do not satisfy the research objectives, the process enters in a loop where the previous steps are repeated by going back to modelling step. The network model is adjusted for new requirements and the process moves to next iteration. For instance, the first iteration would be a network model in which nodes are students and undirected links are reciprocated friendships among students. Should the analysis fail to capture the true nature of the friendship relations a second iteration would involve remodeling the friendship relations as a directed network (one student sees the other student as a friend, but the other student does not.)

Then the SNA process continues, and the findings are reported. In some situations, researchers may decide that there is a problem with the data they collected in the beginning. For instance, they might decide to add a node attribute. In that case, process moves back to the beginning and the entire process is repeated with the changes made.

ACTA INFOLOGICA
ISSN: 2602-3563

Modelling

The choices made at the modelling step are critical to representing a complex system as a graph. For instance, individuals who are regularly interacting with each other can be connected to create a professional network. On the other hand, the relationships among individuals who are calling and mailing each other define an acquaintance network. The analysis of first type of networks for instance, can play role in a company's success in terms of management. The analysis of second type of networks can play a key role in marketing products or services (Aydın and Perdahçı, 2014). Likewise, when analyzing social networks in education modelling step can be critical.

There are several options as there are several types of interactions in a school environment. These interactions can be acquaintance, close friendship among students or teaching relationship between teachers and students. Additionally, sharing activities such as projects, classes, team sports and various clubs can create interactions. The relationship should be determined by choosing the best option, which fits the research objectives. Furthermore, multiple relationships can be selected as links. In that case, the network becomes multi layered. To illustrate a multi layered model, let the close friendship relations among a set of students constitute a friendship network, while linking the same set of students according to their shared projects, make up a collaboration network. The combination of these two networks in one network in which some of the links represent friendship and others represent collaboration, results in a multi layered or, in other words, multiplex network. The actors in a network can also vary in terms of node selection such as, students, instructors, administrators. However, selecting nodes is usually simpler with respect to link selection.

Table 1. Possible inputs and outputs of SNA in school networks depending on the link type

	Friendship Network	**Collaboration Network**	**Course Affiliation Network**
Input	Who is friends with whom	Who studies with whom	Who takes same courses with whom
	Test achievement scores	Content of the project	Course subject
	Gender	Degree of contribution	Social clubs
	Class	Project grade	Team sports
	Social Background	Study major	Success rate of the activities
Used Metrics	Clustering	Clustering	Clustering
	Assortativity	Assortativity	Assortativity
	Modularity Opt.	Modularity Opt.	Modularity Opt.
	Centrality measures	Centrality measures	Centrality measures
Outcome	Correlations based on attributes	Correlations based on the specific project	Better groups, teams etc.
	Class Compositions	Modularity Opt.	Homopohily
	Broker nodes	Sustainability of Interconnectedness	Correlations based on the specific programs

Network Analysis Findings

In this step, the modeled network is visualized as a sociogram (Moreno et al., 1932; Freeman, 2004) by utilizing graph visualization tools. A sociogram is a map of social relations in which individuals are depicted as circles and related individuals are linked with each by lines. Here, the students are represented as nodes and the nodes are linked according to chosen model in the previous stage. This graph representation gives visual insight into the social network, which shows how the students are connected with each other. After visualization, basic

ACTA INFOLOGICA
ISSN: 2602-3563

analysis takes place where basic structural properties of the network are calculated. These findings for instance, show the density of the network (i.e. how densely they are connected to each other). And, how much the students are clustered (how closely they are grouped). Further analysis finds the communities in the network or how students with different attributes mix with each other. Additionally, the students who form a link between communities can be found with SNA algorithms.

Although in an education setting there are several relationships that can be subject of study, the current study focuses on three types of networks: Friendship networks, collaboration networks, and affiliation networks. While the friendship networks focus on the friendship relations among students, collaboration networks focus on the shared activities of the students such as working on the same project. Attending the same course or being in the same social club generally represent the links of an affiliation network. These network types nececitate different inputs and the anlysis of these networks offers different outputs. Table 1. Shows some of the possible inputs for a research that utilizes SNA as a method and the possible outputs of that analysis.

4. DISCUSSIONS OF VALIDITY ISSUES FOR LINKED DATA IN A SCHOOL INFORMATION SYSTEM

In the following, we discuss our findings on validity issues encountered in linked data driven Information Systems research.

4.1. Reliability Issues from System Generated Data

SNA in school networks obtain node attributes such as student grades, name of the taken courses etc. from information systems. These are not system generated data and since such data is heavily checked by the administrators, teachers, students and parents, they can be considered reliable. For instance, if the age, gender or grade of a student is entered incorrectly into the system, the students instantly demand a correction. Therefore, in a school network the validity of the data gathered from Information System should not be an issue for statistical conclusions derived from this data.

4.2. Aligning Digital Data and Nodes & Links

A researcher, at the beginning of a study, has to decide how to use the data to build a network. The first desicion is to determine which entities in the data will constitute nodes and links. There are many options for this decision step, for instance, in an e-commerce sales data, products might be the nodes as well as the sellers or the buyers. The desicion should align the data and the network according to the research objective both contextually and theoreticaly. “In Social Network Analysis, however (emphasis on the word Social), nodes are almost always people, although at different levels of analysis they might be individuals, groups, or organizations” (Howison et al., 2011). Similarly, in education settings nodes are almost people namely, students, instructors or administrators. There may be instances where education tools or classrooms make up the nodes of the network along with students and instructors. Since these are all solid entities in education environments, they should not raise significant validity issues.

4.3. Choosing Multiple or Single Link Types

Choosing link types in modelling step of SNA is a validity issue that concerns the construct of the network. The number of link types in a network should align with the research interests and requirements. In a school, social network there are several relationships that can be studied: affiliation (course or lab attendance), collaboration (study or project groups), friendship (best friends or acquaintance). A network can be modelled using single link type or multiple types (multiplex network). An analysis of a single link type network may leave out the effects of other relationships among the nodes. On the other hand, a multiplex network analysis may lack the ability to detect the certain interactions taking place in the network. In that sense, number of link types should align with the research objectives.

ACTA INFOLOGICA
ISSN: 2602-3563

4.4. Defining a link (Link Intensity)

Only existence of a link between two nodes is sufficient for some research problems however, in some cases link's intensity or strength is also a key factor. For example, in a student collaboration network, links can be established in a binary mode where a link has only two states: zero or one. In other words, a links indicates whether two students studied together or not. However, if the same students collaborated multiple times or they collaborated for longer periods, we might need a stronger link betweeen them so that the degree of their collaboration is represented as link intensity. Therefore, the links should have an attribute indicating the intensity of that relationship called weight of a link. Metrics such as centrality measures or community detection algorithms produce different results depending on the type of the links.

4.5. Defining a non-link (Missing Links)

In school social network analysis, often link data is obtained via face to face surveys. In a school environment, social networks consist of small number of nodes. Therefore, reaching to the students, instructors or administrators is not a challenge. This eliminates the validity issue from missing node aspect. However, this does not eliminate the non-link issue which means that knowing if a link exists between two nodes is important, in the same sense, knowing if a link does not exist is also important. In other words, lack of a link indicates lack of interaction between two nodes where in fact, this might be due to missing data, which can cause serious errors for example, in a friendship network, students are asked about their friends and the links of the network is established with that data. However, in the constructed network if two nodes are not linked this does not necessarily mean that they are not friends. If information flow in the network is needed to be measured, making sure that two students are exactly friends or not friends is critical. Betweenness centrality is a network measure that quantifies the brokerage position of a node. The nodes that have high betweenness centrality play a vital role in bridging communities in a network. For papers that examine this metric in school networks (Grunspan et al., 2014) absence of a node is as important as existence of it.

4.6. Temporal Aggregation

In dynamic analysis of a social network, some situations can cause validity problems. For example, an affiliation network considers two students as linked if they take the same course. However, if the students did not attend that course at the same time, this means that there has not been an interaction between those nodes in the network. This issue happens when links aggregate in time and can lead to faulty analysis results especially for the algorithms that depend on the path calculations. There are solutions to this problem and the techniques to deal with this issue are mentioned in (Howison et al., 2011).

4.7. Network Tool Effects

There are various tools to visualize the network map and calculate the network metrics (Csardi and Nepusz, 2006; Hagberg et al., 2008). Using these tools helps researcher to standardize the implementation of the algorithms. Otherwise, reimplementation of the algorithms by different researchers could cause other validity issues. Additionally, using common tools provide some shared ground for the work to be repeated by other researchers. "Nonetheless, the convenience these tools provide can also mask threats to validity in their use. First, programs use subtle variations of algorithms and slightly different names for the same algorithm, potentially leading to confusion and misinterpretation of results." (Howison et al., 2011). To overcome this problem, the same algorithms are calculated manually (pen-and-paper calculation) on a toy network and compared with the network tool's calculation results.

4.8. Temporal Mismatch

In most cases, social networks are analyzed in a static manner. In other words, obtained data is a snapshot of the network, which represents a short time interval. This can lead to validity issues where a network metric which changes over time as the network evolves may mislead the researchers (Huisman and Snijders, 2003; Leskovec et al., 2005). For example, betweenness centrality of a student may be high when the data is collected, then it may change due to the nature of social interactions. Moreover, not only measured values change but also the

ACTA INFOLOGICA
ISSN: 2602-3563

links in the network may change over time. For instance, a student may decide to partner with a different student in a collaboration network. Similarly, change of friends or even best friends is frequently seen among high school students. Researchers should make sure that the observed measurements span the entire time frame.

4.9. Questions of Data Completeness

Data about the relationships among actors of a network gathered in one environment no matter how complete may lack the information of interactions that happen out of that environment. For instance, researchers may study friendships among students and its correlations to numerous factors by collecting the relationship data via surveys or interviews. However, due to access or privacy issues, their information most likely will lack their social media interactions. This is another validity issue that is caused by incompleteness of the data. Depending on the subject of the study, incomplete link data will reduce the reliability of the SNA results.

4.10. Inappropriate Importation of Network Measure Interpretation.

Network measures and algorithms utilized in SNA are mostly transferred from other disciplines of science. For example, Pearson's correlation formula was originally used to describe a biological phenomenon (Pearson, 1896). It was later imported by several other disciplines such as economy, physics, chemistry etc. when it is needed to examine linear relationships between two quantitative variables. In network science, this metric is used to find out the mixing behaviors among nodes called assortativity (Newman, 2002). In specific case of social networks, this metric is used for instance, to measure mixing behavior among individuals (Bearman et al., 2004). The interpretation of the results might differ in different venues. Therefore, researchers should pay attention to the interpretations of their findings for avoiding validity issues that is caused by importation.

5. CONCLUSION

This paper addresses the validity issues that researchers face when conducting Social Network Analysis. Although its scope is SNA studies in general, education domain applications is used to exemplify the validity issues. Prior to addressing these issues, network concept is briefly explained. Furthermore, a conceptual model is presented which covers Network Science processes and how the linked data advances starting from the real-world system and IS to complex systems and finally analyzed to produce scientific output.

The validation issues mostly arise between the phase transitions. Data reliability at the beginning when deciding the nodes and links are not likely to cause serious validation problems since these entities are well defined in education networks. However, decisions about the link types, weights or even non-existence of a link are potentially critical validation checkpoints during the SNA process. Another type of validation issue arises due to temporal issues when deciding the analysis to be static or dynamic. The interpretation of the algorithm results should involve the effects of time over the network. Additionally, as in every scientific research, the utilized tools have validity issues as well as the measures. Researchers should be aware of the strengths and weaknesses of their tools and metrics.

The Network Science is relatively a new discipline. Therefore, researchers should be informed about the pitfalls throughout the processes. This paper does not claim to address all possible issues rather; it is intended to be used by researchers as a starting point to avoid these issues and as a validation checklist after their research. To that end, issues are collected and examined in detail furthermore; practical solutions are offered to facilitate the researchers in their network analysis efforts.

ACTA INFOLOGICA
ISSN: 2602-3563

REFERENCES

Aydın, M. N., & Perdahçı, N. Z. (2014). Ağ Bilimi Yaklaşımı Ve Çevrimiçi Etkileşimli Sağlık Platformunun Bir Örnek Olarak İncelenmesi: Informa Yönetim Bilişim Sistemleri Dergisi, 1(2), 60-80.

Barabási, A. L., & Pósfai, M. (2016). *Network science*. Cambridge university press.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American journal of sociology*, *110*(1), 44-91

Behrendt, S., Richter, A., and Trier, M. (2014). Mixed methods analysis of enterprise social networks. *Computer Networks*, *75*, 560-577.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1-9.

Dong, X. L., & Srivastava, D. (2015). Big data integration. *Synthesis Lectures on Data Management*, *7*(1), 1-198.

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, *1*.

Grunspan, D. Z., Wiggins, B. L., & Goodreau, S. M. (2014). Understanding classrooms through social network analysis: A primer for social network analysis in education research. *CBE-Life Sciences Education*, *13*(2), 167-178

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States)

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767.

Huisman, M., & Snijders, T. A. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological methods & research*, *32*(2), 253-287.

Jungherr, A. (2015). Analyzing political communication with digital trace data. *Cham, Switzerland: Springer*.

Labun, A., Wittek, R., & Steglich, C. (2016). The co-evolution of power and friendship networks in an organization. *Network Science*, *4*(3), 364-384.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005, August). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 177-187). ACM.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education.

Moreno, J. L., Whitin, E. S., & Jennings, H. H. (1932). Application of the group method to classification: National Committee on Prisons and Prison Labor. *Psychology Abstracts*, *6*, 2872.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, *89*(20), 208701

Nia, R., Bird, C., Devanbu, P., & Filkov, V. (2010, May). Validity of network analyses in open source projects. In *Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on* (pp. 201-209). IEEE.

O'Brien, J. A. (1998). *Management information systems: Managing information technology in the networked enterprise*. McGraw-Hill Professional.

ISSN: 2602-3563

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, *187*, 253-318

Smirnov, I., & Thurner, S. (2017). Formation of homophily in academic performance: Students change their friends rather than performance. *PloS one*, *12*(8), e0183473.

Whelan, E., Teigland, R., Vaast, E., & Butler, B. (2016). Expanding the horizons of digital social networks: Mixing big trace datasets with qualitative approaches. *Information and Organization*, *26*(1-2), 1-12.

This work was presented at the 4th International Management Information Systems Conference and published in the conference abstract book.

ACTA INFOLOGICA, 2018; 2(1): 45-51
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

A Real Life Web Based Marketing Optimization Framework With External Data

Şadi Evren Şeker*

ABSTRACT

Big data and data science studies in recent years are booming exponentially, parallel to the data collected and increased processing speeds. As an inevitable consequence, most of the web-based companies are migrating their business models to novel technologies based on the big data and data science research. This paper is based on a real life experience based on one of the web stores with highest volume sales in Turkey. The project was building a data science model on big data technologies to make estimations based on the external data, such as weather conditions, customer demography, news at newspapers, current product alternatives, financial facts (like currency exchange rate or stock market values) and most importantly the sentimental analysis and opinion mining on social network, blogs and news. In the paper, details of problems and possible solution alternatives and methodology for problem solving and solutions and outcomes of the study are explained in the given order.

Keywords: Marketing, opinion mining, targeted marketing, big data, data science, customer behavior.

Information of Author(s):

Şadi Evren ŞEKER
ORCID: 0000-0002-7323-3695
sadievrenseker@sehir.edu.tr
İstanbul Şehir Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Yönetim Bilişim Sistemleri Bölümü

DOI: 10.30801/acin.356344

Submit Date:	20.11.2017
Accept Date:	03.01.2018
Publish Date:	26.06.2018

(*) Contact Author
Address: İstanbul Şehir Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Yönetim Bilişim Sistemleri Bölümü, İstanbul, Türkiye ● **Telephone Number:** +90 0216 4444 034

ACTA INFOLOGICA
ISSN: 2602-3563

1. INTRODUCTION AND PROBLEM DEFINITION

E-marketing is an increasing trend and the advances in e-marketing is directly affecting the web-based sales [1]. An Internet company in Turkey, has invested for a research project to collect data and develop a data science project to increase marketing success. By definition of the project, marketing success criteria is the percentage of sales for ads displayed. Also the data is grouped into two categories, internal and external as below:

- internal data is defined as the data in the local databases of the company, like customer information, product information, sales information
- external data is defined as any data source, which might affect the customer behavior about the advertisement, such as the weather conditions, financial data (like currency rates or stock market values), opinions and sentimental effects of daily news or blogs or social networks.

The project has three major questions:

1. What are the correlated parameters and which parameters have the highest affect on marketing success?
2. What is the best technology for implementing the project?
3. What is the best data science solution for the decision-making and customer / advertisement matching optimization?

During the project all three questions above are solved and a working real life project is implemented [2]. For the first question, some temporary research oriented systems are installed and some prototype coding is developed on the temporary systems. Although the system was not built on a full scaled technology in this step, data is collected by using some sampling algorithms and some discovery correlation algorithms such as Kendall's-Tau, Spearman's-Rho or Goldman-Kruskal's-Gamma [3] is executed to understand the correlation between the parameters and the marketing success.

For the second major question, the prototype implementation is executed on some technologies like Apache Hadoop, Microsoft Azure, Apache Spark and for the data science layer, MLLib, AzureML and Mahout libraries are implemented for test purposes.

Finally for the third question, several data science solutions are tested. Because the data was collected from man different unstructured sources, during the preprocessing, some imputation, data cleaning and noise reduction algorithms are implemented. Although some of the dimension reduction algorithms tested, the success of these algorithms was not high enough to take into consideration for the case. For example, Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA) algorithms are tested an the contribution of the algorithms was less than 1% for most cases and had negative affect in some cases. After the data preprocessing and extracting the feature vectors, a limited set of machine learning algorithms are tested. The reason for limitation is the number of suitable algorithms in big data world, which means they can run on map-reduce based, is limited within the libraries.

As a summary, the problem is defined on the selection of best technology, best algorithm on most important parameters based on both external and internal parameters. The project outputs the best alternative advertisement for any customer.

2. METHODS

The project can be divided into three sub projects as defined in the introduction and problem definition section. For all subparts of the project different methods are tested and details of these methods are provided within this section. So the section is organized in three subsections, which are the preprocessing and parameter selection with correlation factors, technology decision and data science problems and solutions.

ACTA INFOLOGICA
ISSN: 2602-3563

2.1. Preprocessing and Parameter Selection with Correlation Factors

Before dealing with the data science layer problems, some preprocessing algorithms applied and feature vectors are extracted.

During the preprocessing phase, the noisy data is cleaned and some of the missing data is imputed. The data cleaning is only applied on the external data, since the data source is unstructured and data collection is implemented by some background processing running on servers and the data source is located under some other organizations. For the internal data, where it is structured and gathered via database queries, there is no such problem as noisy data but there are some missing values for some of the data sets. For example, database might not hold the whole information about customer demographics like age or birthplace of the person. This information is collected from some of the customers by their own will. So, there is no imputation for the external data and no data cleaning for the internal data because of their structures.

For the imputation phase, k-nearest-neighborhood (k-nn) algorithm is deployed with k=3, and the missing values are predicted by the rest of the data set with using 3 most common samples. For the data cleaning, the row-wise deletion technique is implemented for higher confidence.

Just after the preprocessing phase, the correlation factors are calculated over the feature vectors extracted.

Pearson's Rho function is simply division of covariance of two parameters to the multiplication of standard deviation of each parameter as demonstrated in equation (1).

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \quad (1)$$

Pearson's Rho function yields a value between -1 and +1 and 0 means there is no correlation while +1 means they are highly correlated and -1 means there is a high correlation in the negative direction.

Kendall's Tau function is based on concordant and non-concordant pairs and concordant pairs are considered as the agreement between two parameter features. In below equation the count of non-concordant is subtracted from the number of concordant samples and the value is divided to a normalization factor, which is the total number of pairs within the given data set. The formulation of Kendall's Tau is demonstrated in equation (2).

$$\tau = \frac{\sum_{i=0}^{n} concurdant_i - \sum_{i=0}^{n} nonconcurdant_i}{n(n-1)/2} \quad (2)$$

Similar to Pearson's Rho, Kendall's Tau also yields a value between -1 and +1 and again the 0 output means no correlation, +1 means high correlation in same direction and -1 means high correlation in negative direction.

Finally, Goldman Kruskal's Gamma is based on the probability of randomly selected pair of observation in the same order (Ps) or opposite order (Pd) and can be easily calculated by the division of difference to summation between these two parameters. This formula is given in equation (3).

$$\gamma = \frac{P_s - P_d}{P_s + P_d} \quad (3)$$

Similar to the Pearson's Rho and Kendall's Tau correlations, the Goldman Kruskal's Gamma is also between -1 and +1 and again the 0 output means no correlation, +1 means high correlation in same direction and -1 means high correlation in negative direction.

After calculating all three correlation coefficient [4] [5], we go for the arithmetic average of methods and try to rank the correlation between each parameter and the output parameter.

ACTA INFOLOGICA
ISSN: 2602-3563

2.2. Deciding Technology

Decision on the technology for the project depends on many different criteria, such as the current knowledge and experience of employees on a certain technology, current software license agreements, current technologies already installed on the servers and so on. The fortunate part of the project was, the company was investing its first big data / data science project and they were open for any technology decision. The technology decision is divided into three layers during the technology management of the project:

- Decision on data science technologies
- Decision on big data technologies
- Decision on the server technologies depending on the above decisions

So the first step of technology decision was deciding on the final step of the project. During the project, the data science layer was determinant for the rest of the technology decisions. The reason was, for any alternative technology in the data science layer with less success would create a negative motivation for the project, so data science layer is accepted as the determiner of the technology for the project.

The data science level algorithms are decided at the first step and details of the algorithms are explained in subsection 2.3 of this paper. After the decision of the algorithms, performance tests are executed and the performance of technologies are compared on the private cloud servers. Company was already holding a private cloud with Linux and Windows operating systems and core database of the company was built on MS SQL. For the apache technologies, Hadoop and spark, the tests ran on the Linux servers and for azure, the tests ran on the Microsoft cloud. After the benchmark test comparison, company decided to go on the spark technology and searched for a suitable server alternative. The best alternative with the cost/performance and easy maintenance was amazon web services (AWS), that the company would reach. Although AWS offers some plans and technology alternatives, another decision is done for the simple storage service (S3) and running spark in the AWS cloud [6].

2.3. Data Science Problems and Solutions

The first data science experiments are executed on sampled data with Rapid Miner software and the results achieved during the experiments are carried on to the production phase. During the experiments, five basic problems are researched:

- Feature extraction technique for the opinion mining
- Machine Learning Algorithm for the opinion mining
- Sales Forecasting
- Ad Matching
- Customer Segmentation

During the feature extraction for opinion mining phase, three major techniques are tested, term frequency – inverse document frequency (TF-IDF), Word-to-vector and Bag-of-Words [7]. During the machine learning for opinion mining, the classification algorithms, such as, k-nn, decision tree, random forest and naïve bayes algorithms are tested. For the forecasting, linear regression with, least squares, Lasso and ridge alternatives are tested, also isotonic regression and random forest algorithms are tested at this step too. Finally for the ad matching problem, recommender algorithms are tested based on content based filtering and collaborative filtering.

As a conclusion of the section 2, the technology decision can be demonstrated as in Figure 1.

ACTA INFOLOGICA
ISSN: 2602-3563

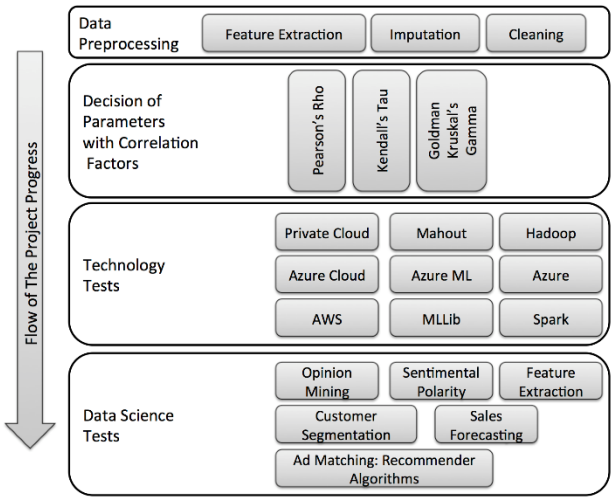


Figure 1. Overview of the Project Progress

The best results achieved on TF-IDF during the feature extraction, random forest for the opinion mining, logistic regression for the customer segmentation and k-nn algorithm based on the user/product similarity. Finally for the customer segmentation problem, a decision tree algorithm is implemented over the x-means algorithm. X-means algorithm ran with unknown k parameter for k-means, bounded between 2 to 120 and best solution achieved for the 38 segments of customers. Again, for the decision tree on the customer segmentation, best results achieved by the random forest algorithm. Details of only best practice algorithms are provided below.

TF-IDF is one of the text mining methods used for feature extraction from natural language data sources [8], [9] [10] [11].

The TF-IDF calculation is provided in equation (1).

$$TF - IDF(t, d, D) = tf(t, d) x\ idf(t, D) \qquad (4)$$

Where t is the selected term, d is the selected document and D is all documents in the corpus. Also TF-IDF calculation in above formula is built over term frequency (TF) and inverse document frequency (IDF), which can be rewritten as in equation (5).

$$tf(t, d) = \frac{f(t, d)}{\max\{ f(w, d) : w \in d \}} \qquad (5)$$

where f is the frequency function and w is the word with maximum occurrence. Also the formulation of IDF is given in equation (6).

$$d(t, D) = log \frac{|D|}{|\{d \in D : t \in d\}|} \qquad (6)$$

where |D| indicates the cardinality of D, which is the total number of documents in the corpus.

Another crucial algorithm mostly applied on similar data sets is the random forest (RF) algorithm [12]. Random forest is based on the decision tree approach and the method applies bagging [13] [14], random feature selection [15], and shape quantization [16] methods together.

Random forest algorithm simply creates random decision trees and combines the decision trees depending on their success rates. The similar approach can be easily applied on the other algorithms as a meta-classifier. In this approach the decision tree divides the problem space and the majority voting is applied for each portion of data. Although some techniques like boosting may show better success rates, bagging has an advantage of avoiding the over-fitting problem. By a general definition, bagging reduces the variance and avoids the over-fitting [17].

ACTA INFOLOGICA
ISSN: 2602-3563

In this study, we have applied bagging both directly by using random forest and applying techniques as an ensemble classifier over other algorithms.

The second attempt was implementing the random forest algorithm and the most important parameter for the algorithm is the number of trees. This parameter indicates the number of randomly generated decision trees and the random forest algorithm will work as a meta-decision tree over these random decision trees. We have defined the number of random trees as 10, which means random forest algorithm will generate 10 random decision trees and another decision tree over these 10 decision trees as a meta-decision tree.

3. FINDINGS

As an output of the research for different alternatives in section 2, for the problem statement in section 1, the best solution and the overview of the project can be demonstrated as in Figure 2.

The project has two major data source as input. The first group of data source is the internal data source with customer and product statistics. The second group data source is the external data source, which includes the weathers, social networks [18], news and blogs. In order to collect external data, some resident processes are deployed and a temporary database is implemented. All the information collected from both internal and external data is gathered in S3 server in an AWS service for long time storage and online processing as a big data solution. Spark server is running on top of simple storage service (S3) and machine-learning algorithms are running on top of spark server with full map-reduce advantages. So the spark server can load balance and use scaling and cost reduction advantages.

Finally, project provides the required reports and the optimized matching algorithm between the customer and the products and/or advertisements.

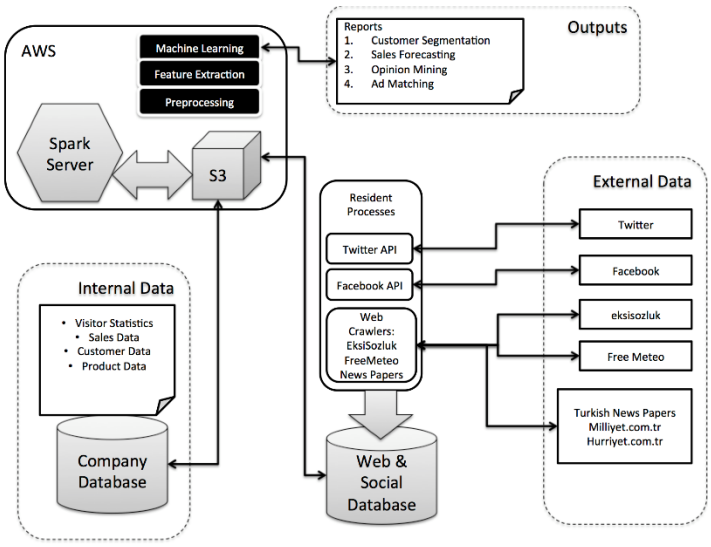


Figure 2. Overview of the Project and Problems Deployment

4. DISCUSSION AND CONCLUSION

The project was about deciding the data science solution alternatives and technology investment for a big data project of a web based company in Turkey. During the project, the determinant factor was the data science policy and technology for the rest of the project investment. After achieving a certain success improvement in the data science prototyping, company researched for the best big data investment. Also from the project it was certain that, the external data sources are affecting the recommender system together with the internal data sources. Project has some unique properties like the first time application of recommender system based on multiple external data sources, working on big data platform and specialized on web marketing in Turkey. Of course, the project will be

ACTA INFOLOGICA
ISSN: 2602-3563

improved by time considering different feature extraction methods like time series analysis or implementing new machine learning algorithms on the big data world but it can be considered as one of the first steps in the area.

REFERENCES

[1] Sadi Evren Seker, "Real Life Machine Learning Case on Mobile Advertisement: A Set of Real-Life Machine Learning Problems and Solutions for Mobile Advertisement," in *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, 2016.

[2] Mehmet Lutfi Arslan, Sadi Evren Seker, and Cevdet Kizil, "Innovation driven emerging technology from two contrary perspectives: A case study of Internet," *Emerging Markets Journal*, vol. 3, no. 3, p. 87, 2014.

[3] Sadi Evren Seker, *Weka ile Veri Madenciliği*.: Draft2Digital, 2015.

[4] Sadi Evren Seker, Cihan Mert, Nuri Ozalp, and Ugur Ayan, "Time series analysis on stock market for text mining correlation of economy news," *Int. J. Soc. Sci. Humanity Stud*, vol. 6, no. 1, pp. 66-91, 2013.

[5] Sadi Evren Seker, Yavuz Unal, Erdinc H Kocer, and Zeki Erdem, "Ensembled correlation between liver analysis outputs," *International Journal of Biology and Biomedical Engineering*, vol. 8, pp. 1-5, 2014.

[6] Abhishek Gupta and Dejan Milojicic, "Evaluation of HPC Applications on Cloud," in *Open Cirrus Summit (OCS) 2011*, vol. 6, 2011, pp. 22-26.

[7] Sadi Evren Seker and Cihan Mert, "A Novel Feature Hashing for Text Mining," *Journal of Technical Science and Technologies*, vol. 2, no. 1, pp. 37-40, 2013.

[8] Marc-André Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *HICSS '04 Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, vol. 3, 2004, pp. 64-73.

[9] Robert P. Schumaker and Hsinchun Chen, "Textual analysis of stock market prediction using breaking financial news:The AZFin Text system," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 2, pp. 1-19, 2009.

[10] Saman Halgamuge, Y Zhai, and Arthur Hsu, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," in *Advances in Neural Networks - ISNN 2007 (Lecture Notes in Computer Science)*, vol. 4493, 2007, pp. 1087-1096.

[11] Gabriel P. C Fung, Jeffrey X Yu, and Wai Lam, "News sensitive stock trend prediction," *Lecture Notes in Computer Science*, vol. 233, pp. 481– 493, 2002.

[12] Breiman L, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[13] Breiman L, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49-84, 1996.

[14] Breiman L, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[15] Ho TK, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognitio*, pp. 278-282, 1995.

[16] Amit Y and Geman D, "Shape quantization and recognition with randomized trees," *Neural Computing*, vol. 9, no. 7, pp. 1545-1588 , 1997.

[17] Watanachaturaporn P and Varshney PK, Arora MK Xu M, "Decision tree regression for soft classification of remote sensing data:," *Remote Sensing of Environment* , vol. 9, no. 3, pp. 322-336 , 2005.

[18] Sadi Evren Seker and Atik Kulakli, "Macroeconomic ICT Facts and Mobile Telecom Operators via Social Networks and Web Pages," *Journal of Business Economics and Management*, vol. 4, no. 2, pp. 99 - 104, 2016.

This work was presented at the 4th International Management Information Systems Conference and published in the conference abstract book.

YAYIN POLİTİKASI

Acta Infologica, İstanbul Üniversitesi Enformatik Bölümü bünyesinde Haziran ve Aralık aylarında olmak üzere yılda iki kez, makalelerin tam metin olarak yayımlandığı akademik hakemli elektronik bir dergidir. Bilimsel içeriğin doğru ve uygun olarak yayımlanabilmesi için tüm çalışmalar konusunda uzman en az iki hakem tarafından değerlendirilmekte olup değerlendirme sürecinde kör hakemlik yöntemi uygulanmaktadır. Değerlendirme sonucunda yayımlanması uygun görülen makaleler yayın sıralamasına alınarak makale sahibine bildirilir. Derginin yayın dili Türkçe ve İngilizce'dir.

Derginin hedef kitlesi veri-enformasyon-bilgi kavramlarını, bilgi- iletişim teknolojileri ve uygulamalarını temel alarak öncelikle Enformatik olmak üzere bu konularda disiplinlerarası alanda da çalışma yapan bilim insanları, araştırmacılar, uzman kişi ve kuruluşlardır. Değerlendirilmek üzere gönderilen tüm çalışmalar bilimsel yayın yapma etiğine uygun olarak hazırlanmalı ve aynı anda başka bir dergi, kongre, konferans vb. yerde değerlendirme sürecinde olmamalıdır. Makale eğer lisansüstü tezlerden üretilmiş veya tezin bir bölümü ise bu durum makale yazarı tarafından makaleye dipnot olarak belirtilmelidir.

Dergiye gönderilen makaleler Editörler tarafından alana özgün katkısı, bilimsel yöntem, anlatım özellikleri ve yazım kuralları açısından incelenir ve benzerlik tarama programından geçirilir. Bu kural ve koşullarla bağdaşmayan yazılar yayınlanmaz.

Makalenin dergide yayınlanmasından sonra makalenin kullanılması, kopyalanması, yayınlanmasına dair tüm telif hakları ACIN'a aittir. Yayınlanmış çalışmaların yazarlarına telif ücreti ödenmez. ACIN ve yazar ismi kaynak gösterilmeden makalelerden alıntı yapılamaz. Derginin tüm hakları saklıdır. Derginin hiçbir bölümü Yayın Sahibi Temsilcisi'nin izni olmaksızın ticari veya başka amaçla elektronik veya mekanik formatta çoğaltılamaz. ACIN ücretsizdir ve Dergipark (dergipark.gov.tr/acin) üzerinden dergiye erişilebilmektedir. ACIN'de yayımlanan yayınlardaki bilimsel içerik ve ifadelere ilişkin tüm sorumluluk yazarlara ait olup dergi yayın yönetimine, editörlere veya İstanbul Üniversitesi Enformatik Bölümü'ne ait değildir.