POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)

URL: http://www.politeknik.gazi.edu.tr/index.php/PLT/index

Akademik veritabanlarından yazar-makale bağlantı tahmini

Co-author link prediction from academic databases

Yazar(lar) (Author(s)): *Yücel BÜRHAN, Resul DAŞ*

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To link to this article): Bürhan Y. Ve Daş R., "Akademik veritabanlarından yazar-makale bağlantı tahmini", ***Politeknik Dergisi***, 20(4): 787-800, (2017).

Erişim linki (To link to this article): http://dergipark.gov.tr/politeknik/archive

DOI: **10.2339/politeknik.368989**

Akademik Veritabanlarından Yazar-Makale Bağlantı Tahmini

Araştırma Makalesi / Research Article

Yücel BÜRHAN[1*], Resul DAŞ[2]

[1]Munzur Üniversitesi, Tunceli Meslek Yüksekokulu, Bilgisayar Programcılığı Bölümü, Tunceli, Türkiye
[2]Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü, Elazığ, Türkiye

(Geliş/Received : 08.02.2016 ; Kabul/Accepted : 25.03.2017)

ÖZ

Sosyal ağlar: bireylerin tutum ve davranışlarını, jest ve mimiklerini sanal ortamda sembolik eylemlerle sergiledikleri sosyal iletişim platformlarıdır. Hızla gelişmeleri ve kullanım oranının her geçen gün artması sosyal ağların popülaritesini artırmaktadır. İnsanlar birbirleri ile iletişim kurarlar ve fikir, düşünce, fotoğraf, video ve konum gibi bazı verileri buradan paylaşırlar. Bu veriler işlendiğinde kullanıcılar ile ilgili çok önemli bilgiler elde edilir. Bu bilgiler ışığında kullanıcılar ile ilgili önemli tahminler yapmak mümkün olur. Bir ağı analiz edebilmek için öncelikle ağın modeli, graf yapısının özellikleri, ölçütleri ve metrikler hesaplanmalıdır.

Bu çalışmada yazarların çalıştıkları konular arasındaki benzerlik hesaplaması yapılmaktadır. Uygulamada, rastgele seçilen yirmi yazar ile oluşturulan veri seti için komşuluk tabanlı modeller arasından uygun olan beş tanesi (Jaccard Index, Sorensen Index, Ortak Komşu, L. H. Newman Index ve Salton Index yöntemleri) uygulanmaktadır. Hesaplama sonuçlarının değerlendirilebilmesi için; aralarında bağlantı olduğu bilinen yazarlar arasındaki benzerlik hesaplaması sonuçları, referans değer olarak kullanılmaktadır. Elde edilen benzerlik değerlerinin referans değerleri ile kıyaslanması sonucu yeni bağlantılar oluşmaktadır. Oluşan bu bağlantılar yazarların birlikte yayın yapabileceğini ortaya koymaktadır.

Anahtar Kelimeler: Sosyal ağlar, ortak yazar bağlantı tahmini, ağların matematiği, bağlantı tahmini.

Co-Author Link Prediction from Academic Databases

ABSTRACT

Social networks: social communication platforms where individuals exhibit their attitudes and behaviors, gestures and mimics in symbolic actions in a virtual environment. Their swift development and wide usage proportion rises social networks' necessity and popularity. People communicate with each other and share some data like idea, thinking, photo, video and location. When this data is processed, very important information is obtained about the users. In this light, it is possible to make important estimates about the users. To analyze a network, the network model, properties of graph structure, criteria and metrics must be calculated first.

In this study, similarity calculation between the subjects which authors studied has been done. In this application, five suitable models (Jaccard Index, Sorensen Index, Common Neighbor, L. H. Newman Index and Salton Index methods) are applied to the data set generated by randomly selected twenty writers. In order to evaluate the calculation results; the results of the similarity calculation between authors with known linkages are used as reference values. Comparisons of obtained similarity values with reference values result in new connections. These links reveal that authors can publish articles together.

Keywords: Social networks, co-authorship prediction, mathematics of networks, link prediction.

1. GİRİŞ (INTRODUCTION)

Sosyal ağlarda bağlantı tahmini, ağın mevcut durumunun incelenerek gelecekteki durumunun tahmin edilmesi problemidir. Bunun gerçekleştirilebilmesi için de sosyal ağlarda bireyler düğüm, bireyler arası ilişkiler ayrıt olarak düşünüldüğünde ağ yapısı, graf yapısı ile örtüştürülerek tanımlanabilir. Graf teorisinde G=(V, E) şeklinde bir matris yapısı oluşturulabildiği için, kullanıcılar arası ilişkiler bu matris üzerinden tanımlanabilmektedir.

Düğüm niteliklerinin bilindiği ağlarda, mevcut olmayan ama gelecekte oluşabilecek bağlantılar tahmin edilebilmektedir. Ağda yeni ilişkiler oluşması, ağa yeni düğümlerin katılması ihtimalinin yanı sıra ağdan bağlantıların yok olması da mümkündür. Sosyal ağlar dinamik olduğundan dolayı bu tahminlerin yapılması oldukça zordur [1].

Bağlantı tahmini yapılabilmesi için ağdaki bilgilerin nasıl tanımlanacağı, bilgilerin nasıl kullanılacağı konuları temel sorunların başında gelmektedir. Bu tanımlama ve kullanım şekli doğru yapılandırıldığında bağlantı tahmini işleminin etkin ve doğru bir şekilde gerçekleştirilebileceği görülmektedir [2].

Bu çalışmada bazı bağlantı tahmini yöntemleri yazar-makale ağına uygulanmaktadır. Yapılan bu çalışmada, lokasyondan bağımsız olarak gelecekte birlikte yayın yapabilecek yazarların tahminine yönelik bir model geliştirilmiştir. Gerçekleştirilen model, literatürdeki benzer çalışamalardan farklı olarak hazır bir veri seti kullanmamaktadır, belirli bir lokasyonla sınırlandırılmamıştır ve

**Sorumlu Yazar (Corresponding Author)*
e-posta : yucelburhan@munzur.edu.tr

var olduğu bilinen ilişkileri görsel olarak sunmak yerine olası ilişkileri tahmin etmeye odaklanmaktadır.

Çalışmanın bir sonraki bölümünde daha önce bu alanda yapılan çalışmalar anlatılmaktadır. Çalışmanın üçüncü bölümünde bağlantı tahmini problemi irdelenmekte, bağlantı tahmini problemi için kullanılan yöntemler açıklanmaktadır. Dördüncü bölümde bağlantı tahmini problemi ile ilgili yapılan çalışma detaylı olarak açıklanmakta, problem çözümü için gerekli işlemler adım adım anlatılmaktadır. Beşinci bölümde uygulama sonuçları, şekiller ve tablolar ile verilerek sonuçlar detaylı olarak incelenmektedir. Son olarak genel sonuçlar ve değerlendirmeye yer verilmektedir

1.1. Benzer Çalışmalar (Similar Studies)

T. H. Huang ve M. L. Huang [3] önerdikleri çalışmada DBLP ağında geçmiş döneme ait verileri kullanarak ortak yazarlık analizi yapmışlardır. Birlikte yayın yapmış olan yazarları yılları da baz alarak görsel olarak göstermişler; böylece akademik işbirliği anlamında oldukça faydalı sonuçlar elde etmişlerdir.

Sun ve arkadaşları [4] çalışmalarında yine DBLP veritabanını kullanmış, yapılan benzer çalışmalardan farklı olarak yazarlar arasındaki ilişkiyi hesaplarken tek bir faktör ile değil de birkaç değişken faktörü ele alarak analiz yapmışlardır (örn. sadece yazar tipi değil de konumları, konuları ve makale isimleri de analize dahil edilmiştir). Yapılan çalışma sonucunda oldukça iyi sonuçlar alındığı gözlenmiştir.

Pengbin ve arkadaşları [5] yaptıkları çalışmada Google Scholar'dan elde ettikleri veri kümesinde yazarların yayın sayılarını göz önünde bulundurarak merkezilik ölçütüne göre bir analiz yapmışlardır. Elde ettikleri sonuçlar yaptıkları çalışmanın, ağın merkezine yaklaştıkça performansının arttığını göstermektedir.

Li ve Xuezhu [6] yaptıkları çalışmada tanınmış iki üniversite için ortak yazarlık araştırması yapmış ve yapılan çalışmaları karşılaştırarak bir sonuç elde etmişlerdir. Ayrıca üniversite ortak yazarlık ağı gelişimi için de önerilerde bulunmuşlardır.

Anastasios ve arkadaşları [7] yaptıkları çalışmada bir yükseköğretim akademik birim içindeki araştırma ve işbirliği yapılarını 4 yıl boyunca izlemiş, yazarlar arasındaki ilişkileri, kurum içi ve kurum dışı araştırma gruplarına katılan yazarları tespit ve analiz etmiştir.

Bidault ve Hildebrand [8] önerdikleri çalışmada işbirliği yapan asimetrik geçmişe sahip akademisyenler arasında ortak yazarlığın getiri ve kayıplarının dağılımını araştırmışlardır. Bunu yaparken, yazarların önceki yayınlarına göre ortak yazdıkları bir makalenin alıntılarındaki artış ve azalış ile eş zamanlı olarak yayınladıkları makalelerin artış azalışları ayırt ediliyor. Ayrıca hem genç hem de kıdemli yazarlar için bu değerlerin artış azalışı etkileyen faktörler irdelenmiştir.

Türker ve Çavuşoğlu [9] çalışmalarında bilimsel işbirliği ağlarında ana parametrelerin evrimsel keşfinden ziyade çok bağlantı özelliklerini çok daha detaylı bir şekilde ortaya çakarmaya çalışmışlardır. Bunun için de eşleştirilen düğümlerin derece koşullarını, derece farklılıklarını, akademik yaş farklılıklarını ve bağlantı ağırlıklarını göz önünde bulundurmuşlardır. Ortak yazarlık ağında bağlantıların çoğunun karşılaştırılabilir derecelerde ve akademik yaştaki düğümleri birbirine bağladığı gözlenmiştir. Bununla birlikte benzer akademik kariyerler arasında güçlü işbirliği faaliyetleri tespit edilmiş ayrıca bağlantı ağırlığı ve derece fark dağılımlarında güç yasası rejimleri gözlenmiştir.

Bruna [10] yaptığı çalışmada ekonomideki ortak yazarlık insidans artışı ile ilgili literatürden yola çıkarak yazarların yayınlardaki dönüşleri optimize etmeye motive olduğu varsayımına dayanarak yazarların ortak yazarlık seçeneklerini analiz etmek için teorik bir model sunmuştur. Model iki maliyet yapısı analiz etmektedir; bunlardan biri yazar sayısı ile orantılıdır diğeri değildir. Araştırmacıların heterojenliği, kural yapmak için en üst seviye çaba göstermek ile daha iyi araştırmacı seçmek arasındaki dengeyi ima eder. Bu modelde düşük kaliteli aralтırmacılar yüksek kaliteli araştırmacılarla işbirliği yapabilmek için fırsat kollarlar. Sonuç olarak ödüllerin yazarlarla orantılı olmadığı gözlenmiştir. Ödüller düşük ise her iki yazar türünün de (yüksek kaliteli ve düşük kaliteli) birbiri ile işbirliği yapabildiği ama ödül yüksek ise tüm yazarların yüksek çabalarını en yükseğe çıkarabilecek bir araştırma organizatirüne yöneldiği gözlenmiştir.

De Stefano ve arkadaşları [11] yaptıkları çalışmada İtalyan akademik istatistikçileri arasındaki ortak yazar ağlarının analizi üzerine üç veri kaynağının (Web of Science, Güncel İstatistik İndeksi ve ulusal olarak finanse edilen araştırma projeleri) etkisini araştırmışlardır. Sonuç olarak İtalyan istatistikçilerde uluslar arası düzeyde elde edilen sonuçlara kıyasla, on yıllık bir gecikme ile, ortak yazarlığı artırmaya yönelik genel bir eğilim gözlenmiştir.

Ortega [12] yaptığı çalışmada yazar ağlarının yapısal özellikleri ve araştırma etkisi arasındaki analiz etmeyi amaçlamıştır. Bibliyografik analizler için uygunluğu test etmek amacıyla Micrisift Academic Search kullanılmıştır. Bu motordan alıntı sayısı ve 500 tek adımlı benlik ağı çıkarıldı. Sonuçlar, küçük ve seyrek ağların, (yüksek arasındalık merkeziliği ve yüksek ortalama yol uzunluğu) yüksek kümeleme katsayısı ve yüksek bir ortalama derece ile tanımlanan, yoğun ve kompakt ağlara göre belge başına daha fazla atıf yapabildiğini göstermektedir. Disiplin farklılıklarına göre, Matematik, Sosyal Bilimler ve Ekonomi ve İşletme, daha seyrek ve küçücük ağlara sahip disiplinler; Fizik, Mühendislik ve Jeosciences, yoğun ve kalabalık ağlarla karakterize edilir.

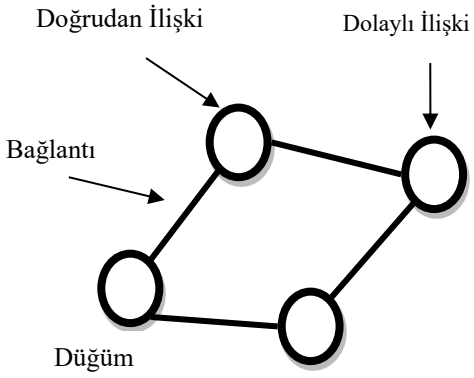
Koseoglu [13] yaptığı çalışmada 1980-1204 yılları arasında Stratejik Yönetim Dergisinde yayınlanan makalelerden yazar işbirliklerinin entelektüel yapısını ve evrimini araştırmayı amaçlamıştır. Bu çalışma, yazarın genel görünüşü, yazarlık kalıpları, yazar üretkenliği, yazarların sıralaması, ortak yazarlık ağının görselleştirilmesi, stratejik yönetim ortak yazarlık ağ niteliklerinin diğer disiplinlerle karşılaştırılması, ana bileşenler ve çekirdek yazarların evrimini içerir. Ayrıca stratejik yönetim

ağının küçük dünya ağ teorisine, merkezilik derecesi, Bonacich'in güç indeksi, yakınlık merkezliliği ve arasındalık merkeziliği gibi bireysel ağ özelliklerine uyup uymadığına ilişkin tartışmalar içerir. Son olarak, yazarlar, sonuçların, sınırlamaların ve gelecek araştırmalar için önerilerin kapsayıcı bir değerlendirmesini sunar.

Sharma ve Sharma [14] çalışmalarında yazarlar arasındaki bağlantıyı tahmin etmek için bir yöntem geliştirmişlerdir. Geliştirdikleri yöntemi C# dilinde ve DBLP veri seti ile gerçekleştirmişler, yöntem olarak yapay sinir ağı tabanlı bir algoritma kullanmışlardır. Yöntemin önceki yöntemlerle kıyaslandığında iyi bir sonuç verdikleri gözlenmiştir.

2. BAĞLANTI TAHMİNİ PROBLEMİ (LINK PREDICTION PROBLEM)

Bağlantı tahmini yapılabilmesi için ağın yapısının tam olarak bilinmesi gerekmektedir. Sosyal ağlarda bağlantı tahmini için ağ, graf yapısına dönüştürülmelidir. Veriler graflardaki köşeler, ilişkiler ise kenarlar olarak tanımlanmaktadır. Ağ yapısı vektörel olarak da ifade edilebilmektedir. Ağdaki köşe ve kenarların yapıları bilinirse buradan daha oluşmamış bazı bağlantılar tahmin edilebilmektedir. Hatta eklenmesi muhtemel köşelerin de yapıları bilinirse aralarındaki bağlantılar bile tahmin edilebilmektedir. Yine aynı şekilde gelecekte kopacak bağlantılar da tahmin edilebilmekte ve graftan silinmesi muhtemel kenarları tespit edilebilmektedir. Bu tahminleri bilmek zor problemlerden biridir çünkü ağ dinamik bir yapıya sahiptir. Ağdaki bilgilerin tanımlanma şekli de önemli bir problemdir. Mevcut bilgilerin hangilerinin hesaplamalarda kullanılması gerektiği, hangi bilgilerin ne kadar etkin rol oynadığı iyi tespit edilmelidir.



Şekil 1. Sosyal bir ağda graf yapısı (Graph structure in a social network)

Graf olarak modellenen sosyal ağda düğümler (aktörler) ve aralarındaki ilişkiyi temsil eden ayrıtlar Şekil 1'de gösterilmektedir. Düğümler arasında direk bağlantı olabileceği gibi dolaylı bağlantılar da olabileceği Şekil 1'de de görülmektedir.

Bağlantı tahminin zorlukları üç kısma ayrılabilir: ilki, ağın yapısının yanı sıra düğümlerin ağ içindeki etkinliğini etkileyen özelliklerinin de bilinmesi ağa yeni eklenecek veya ağdan silinecek düğüm ve bağlantıların tahmin edilmesinde önemli bir unsurdur. Örneğin sosyal ağlarda düğümlere karşılık gelen bireylerin sevdikleri veya sevmedikleri şeylerin bilinmesi bağlantı tahmininde temel kriterler olarak göze çarpmaktadır.

İkincisi, ağ hakkında eldeki bilgiler tutarlı olmayabilmektedir. Bu da ağa bağlantı tahmini için uygulanan algoritmaların etkinliğini düşürmektedir.

Üçüncüsü, ağ çok büyük ise yani düğüm ve kenar sayısı çok fazla ise hesaplamalar zorlaşır, yani büyük bir ağda graf modelinin çıkarılması ve tahmin için uygulanan algoritmaların doğru sonuçlar vermesi ağ büyüdükçe zorlaşmaktadır.

2.1. Bağlantı Tahmininde Kullanılan Yöntemler (The Methods Used In Link Prediction)

Sosyal ağlarda bağlantı tahmini temelde veri madenciliğine dayanmaktadır. Bağlantı tahmini yöntemleri genel olarak graf tabanlı, olasılıksal ve benzerlik tabanlı yaklaşımlar olmak üzere üç gruba ayrılır. Bu yaklaşımlar içinde yer alan algoritmalar kullanılarak sosyal ağın yapısı ve geleceği hakkında fikir sahibi olunabilir.

2.1.1. Graf tabanlı yöntemler (Graph-based methods)

Graf, ağların yapısını modellemek için kullanılan matematiksel bir modeldir. Ağda yollar, merkezilik ölçümleri, köşe dereceleri, kümeleme katsayısı, gibi özellikler ağın yapısı ve geleceği ile ilgili tahminler yapmamızda çok yararlı olacaktır.

Graf tabanlı yaklaşımlarla yapılacak bağlantı tahmininde, köşeler ve köşeler arasındaki kenar sayısı önemli bir yer tutar. Bu sayı derece olarak da adlandırılır. Ağların çoğunda düğüm derecelerinin düzensiz olduğu göze çarpmaktadır. Bazı düğümler beklenenden az bazı düğümler beklenenden fazla düğüme sahip olabilir. Yapılan çalışmalarda derece dağılımına bağlı olarak rastgele graf modelleri önerilmektedir [15].

Derece dağılımının hesaplamasının kolaylığına rağmen gerçek dünya problemlerinde ciddi sorunlarla karşılaşılmaktadır. Özellikle atıf ağları, world wide web ve bazı sosyal ağlarda bu sorunlar net bir şekilde göze çarpmaktadır. Bu sorunun temel sebebi düğüm dereceleri ile ağırlıklandırılmış ağlarda bağlantı tahmini yaparken yeni eklenecek düğümler ve bu düğümlerle birlikte oluşması muhtemel bağlantıların tespitlerinden kaynaklanmaktadır. Örneğin Barabsi-Albert modelinde [16] her yeni düğümü, mevcut düğümlerin sahip olduğu bağlantıların sayısıyla orantılı bir olasılıkla var olan düğümlere bağlar. Matematiksel olarak yeni bir düğümün *i* düğümüne bağlanma olasılığı denklem (1)'de gösterilmektedir:

$$p_i = \frac{k_i}{\Sigma_j k_j} \quad (1)$$

Burada k_i, *i*'ninci düğümün derecesi, önceden var olan *j* düğümlerin derecesine bölünür. Diğer bir deyişle, düşük dereceli düğümlerde yeni bağlantı oluşma olasılığı düşük

iken; yüksek dereceli düğümlerde bu olasılık çok daha yüksektir.

Graf tabanlı yaklaşımlarda ağlardaki gelişim gözle görülebilecek seviyede iken diğer yaklaşımlarda çok daha büyük ağlar analiz edilebilmiştir.

2.1.2. Olasılıksal yöntemler (Probabilistic Methods)

Olasılıksal yaklaşımlar, adından da anlaşılacağı üzere, ağda olabilecek değişikliklerin olasılığını tahmin etmeye çalışır. Olasılıksal yaklaşımlar muhtemel bağlantıların olasılığını tahmin eden modeller ve bir ağın muhtemel yapılarının olasılığını tahmin eden modeller olmak üzere iki gruba ayrılır. Olasılıksal yöntemler çoğunlukla graf tabanlı yaklaşımlara dayalıdır. Son yıllarda sosyal ağlar için üstel rasgele graf modellerine büyük bir ilgi oluşmuştur. Üstel rasgele graf modelleri bir ağın genel özellikleri, köşeler ve kenarlar kullanarak bütün bir ağ için olasılıksal modelleri tahmin etmede kullanılan popüler bir yaklaşımdır. Bu modeller belirli ağ yapılarının oluşup oluşmayacağı hakkında tahminler yapmaya izin veren istatistiksel modellere dayalıdır. Üstel rasgele graflar, sıradan graflardaki sınırlamalarının ortadan kaldırılarak ağlar için makul modeller geliştirmek için kullanılır [17].

2.1.3. Benzerlik tabanlı yöntemler (Similarity-based methods)

Benzerlik, bağlantı tahmini probleminde ağın geleceği hakkında yapılabilecek tahminlerin güçlü olmasında önemli bir ölçüttür. Aralarında bağlantı olmayan iki düğüm birbirine ne kadar benzerse gelecekte bu iki düğüm arasında bağlantı oluşma olasılığı da o kadar yüksektir. Örneğin, Facebook'da arkadaş olmayan iki kişinin ortak özellikleri ne kadar çoksa gelecekte arkadaş olma ihtimalleri o kadar yüksek olacaktır.

Benzerlik ölçütleri semantik ve topolojik olarak ikiye ayrılmaktadır. Semantik ölçütlerde düğümün içeriği benzerlik ölçütü olarak ele alınır. Örneğin yazar işbirliği ağında makalelerin anahtar kelimelerdeki benzerlikle yazarlar arasında gelecekteki etkileşimler tahmin edilebilir [18]. Topolojik ölçütler benzerlik ölçütü olarak ağın yapısını kullanırlar. İki düğüm arasındaki ortak komşuların sayısı topolojik ölçütlere bir örnektir. Topolojik ölçütler literatürde genel olarak komşuluk tabanlı ve yol tabanlı diye kategorize edilmiştir.

2.1.3.1. Komşuluk tabanlı ölçütler (Neighborhood based criteria)

Komşuluk tabanlı ölçütlerde temel fikir x ve y düğümlerinin komşuları Γ(x) ve Γ(y) nın ne kadar ortak özelliği varsa gelecekte aralarında bağlantı olma ihtimali de o kadar yüksektir. Γ(x), x düğümünün ağdaki komşularının kümesini göstermektedir.

• **Ortak Komşular:** x ve y düğümleri için ortak komşuların sayısını ifade etmektedir [19]. Bu ifadenin matematiksel karşılığı denklem (2)'de gösterildiği gibidir.

$$OK(x,y) = |\Gamma(x) \cap \Gamma(y)| \quad (2)$$

• **Jaccard Katsayısı:** x ya da y den rasgele seçilen bir özelliğin hem x hem de y de birlikte bulunma olasılığıdır. Jaccard, ortak komşuların normalleştirilmiş halidir [20]. Matematiksel olarak denklem (3)'teki gibi ifade edilir.

$$JK(x,y) = |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)| \quad (3)$$

- **Salton İndex:** Kosinüs benzerliği de denilen bu yöntemde kx, x düğümünün derecesi ky de y düğümünün derecesi olsun. Bu durumda formül denklem (4)'teki gibi olur [21]:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x + k_y}} \quad (4)$$

- **Sorensen İndex:** Bu yöntem ekolojik topluluk verileri için kullanılır [22]. Matematiksel olarak denklem (5)'te gösterildiği gibi ifade edilir:

$$S_{xy} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \quad (5)$$

- **Leicht-Holme-Newman İndeks:** Ortak komşusu olan düğümlerin benzerlik değerleri bu yönteme göre Ortak Komşu İndeks'inden daha yüksek değer alır [23]. Denklem (6) L. H. Newman İndeks'in matematiksel karşılığıdır.

$$N_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x * k_y} \quad (6)$$

• **Adamic/Adar Katsayısı:** Bu ölçüm iki web sayfasındaki içeriklerin birbirlerine ne kadar yakın olduğunu ölçmektedir. Bunu yapmak için bu sayfaların özellikleri belirlenmelidir [24]. Formülü denklem (7)'de verilmiştir:

$$AA(x,y) = \sum_{z:x \text{ ve } y \text{ nin özelliği}} \frac{1}{\log(frekans(z))} \quad (7)$$

Bu nicelikte ortak özelliklerden nadir olanların ağırlık oranları artırılmaktadır. Bağlantı tahmini problemi için bu formül denklem (8)'deki gibi güncellenmiştir:

$$AA(x,y) = \sum_{z \epsilon \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(z)} \quad (8)$$

Bağlantı tahmini için yapılan çalışmalarda Adamic/Adar diğer ölçütlere oranla daha iyi sonuçlar vermektedir.

• **Tercihli Bağlılık:** Ağda oluşacak yeni bir bağlantının düğümlerinden birinin belirli bir düğüm olma ihtimali, o düğümün komşularının sayısı ile orantılıdır. Yani komşu sayısı çok olan düğümlerin yeni bağlantı oluşturma ihtimali daha yüksektir [25]. Matematiksel formülü denklem (9)'daki gibidir:

$$TB(x,y) = |\Gamma(x)|.|\Gamma(y)| \quad (9)$$

• **Kaynak Paylaştırma İndeksi:** Karmaşık ağlarda kullanılan bir ölçüttür. Birbiriyle doğrudan bağlantısı olmayan düğümler arasındaki bağlantıları ölçer. Düğümler arasında doğrudan bağlantı olmamasına rağmen düğümler komşuları üzerinden bağlantı sağlarlar. Düğümler arasındaki benzerlik aralarındaki veri akışına göre değerlendirilir. Veri akışı yüksek olan düğümler daha benzerdir. Formülde kullanılan k(z), z nin derecesidir [26]. Matematiksel ifadesi denklem (10) da verilmiştir:

$$KP(x,y) = \sum_{z \epsilon \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} \quad (10)$$

2.1.3.2. Yol tabanlı ölçütler (Road based criteria)

Yol tabanlı ölçütler iki düğüm arasındaki en kısa yolların sayısını baz alırlar [27].

• **Katz:** Katz ölçütü düğümler arasındaki en kısa yolların sayısının toplamını baz alır. Benzerlik hesaplamasında l burada yol uzunluğudur ve uzun yolların hesaplama üzerindeki olumsuz etkisini azaltmak için bir parametreye (β^l) üs olarak dahil edilmiştir [28]. Bu ölçüt denklem (11)'de şöyle ifade edilir:

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l . \left|yollar_{x,y}^{\langle l \rangle}\right| \quad (11)$$

$\left|yollar_{x,y}^{\langle l \rangle}\right|$, x ve y düğümleri arasında l uzunluğundaki yolların sayısıdır. $\beta > 0$ olmalıdır. β değeri ne kadar küçük verilirse ortak komşuların değerine o kadar yaklaşılır.

•**Ulaşma zamanı:** Ulaşma zamanı $H_{(x,y)}$, x düğümünden başlayan rastgele gezintinin y düğümüne ulaşıldığında elde edilen adım sayısıdır. Ulaşma zamanı ne kadar düşük olursa x ve y düğümleri birbirine o kadar benzerdir ve aralarında bağlantı olma ihtimali o kadar yüksektir; düşük olması x ve y nin benzer olduğunu aralarında bağlantı olabileceğini göstermektedir. Yönlü ağlarda bu ölçüt simetrik değildir. O yüzden bunun yerine gidiş/dönüş zamanı (commute time), $C_{(x,y)} = H_{(x,y)} + H_{(y,x)}$, kullanılmalıdır. y düğümün çok geniş bir dağılım olasılığına sahip olduğunda $H_{(x,y)}$ çok küçük olacaktır. Bunu dengelemek için ölçüt normalleştirilebilir [29]. Ulaşma zamanı matematiksel olarak denklem (12)'de verilmiştir:

$$UZ(x,y) = -\left(H_{(x,y)}.\pi_y + H_{(y,x)}.\pi_x\right) \quad (12)$$

• **Köklü PageRank:** Ulaşma zamanında x ve y düğümleri arasındaki yollar çok kısa olsa da rastgele yürüyüşlerle x düğümünden y düğümüne ulaşmak için çok fazla düğüm geçilmesi gerekebilir. Bu durumun önüne geçebilmek için rasgele yürüyüş her adımda β parametresindeki olasılık değeri ile yeniden başa döndürülebilir. Böylece ağda rastgele yürüyüşler mümkün olan en kısa yollardan yapılabilir. Rasgele yürüyüşün belli bir olasılıkla yeniden başlatılması web sayfalarındaki PageRank ölçütünün temelidir. Rasgele yürüyüş β [0,1] olasılığı ile başa döner, 1-β olasılığı ile o an bulunan düğümün komşularından rasgele birine gider. Bu işlem her adımda uygulanır. i düğümünün tüm koşuları için diagonal derece matrisinde $D[i,j] = \sum_j A[i,j]$ dir. $N = D^{-1}A$, komşuluk matrisinin satırlarının 1'e normalleştirilmesidir [30]. Formülü denklem (13)'te verilmiştir:

$$KPR = (1-\beta)(I-\beta N)^{-1} \quad (13)$$

• **SimRank:** Bu ölçüt, "iki düğüm benzer düğümler ile bağlantılı ise bu iki düğüm benzerdir" esasına dayanır [31]. Matematiksel ifadesi denklem (14)'te şöyle ifade edilir:

$$Benzerlik(x,y) := \gamma . \frac{\sum_{a \epsilon \Gamma(x)} \sum_{b \epsilon \Gamma(y)} Benzerlik(a,b)}{|\Gamma(x)|.|\Gamma(y)|} \quad (14)$$

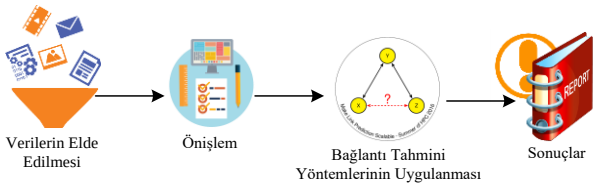
Bu yöntemler dışında pek çok yöntem ve yaklaşım mevcuttur. Bunlardan bazıları sık örüntü madenciliği, rastgele yürüyüş ve yayma yöntemleridir.

3. GELİŞTİRİLEN UYGULAMA (DEVELOPED APPLICATION)

Akademik alanda yapılan çalışmalarda en dikkat eken zorluk, bir akademisyenin çalışmak istediği bir konuda kendine kaynak veya birlikte çalışacağı başka bir akademisyeni bulabilmesidir. Bu çalışmada seçilmiş olan yazarların çalışma konuları ele alınarak yazarlar arasında çalışma konuları bazında oluşan benzerlik hesaplanmıştır. Böylece yazarların çalışma alanlarında kendileri ile benzer konuları çalışan yazarların tespit edilmesi amaçlanmıştır.

Geliştirilen bu araç ile gerçekleştirilmek istenen amaç, yazarların çalışma alanları arasındaki benzerliği hesaplayarak, yazarlar arasındaki ilişkileri analiz etmektir. Bu amaç doğrultusunda, yazarların çalıştıkları alanlara göre benzer konularda çalışan ilgili yazarları tespit etmek; ortak çalışma yapan yazar gruplarını ortaya koymak ve ortaya çıkarılan bu bilgiler kullanılarak elde edilecek sonuçlara göre birlikte yayın yapabilecek yazarları tahmin edecek veya bir yazarın gelecekte çalışması muhtemel konuyu tahmin edecek bir yapının temellerini oluşturmasını sağlamaktır.

Gerçekleştirilen uygulamanın genel adımları Şekil 2'de gösterilmektedir.



Şekil 2. Gerçekleştirilen uygulamanın adımları (steps to implement application)

3.1. Verilerin Elde Edilmesi (Obtaining Data)

Makale Yazar ilişkisinin tespitinde IEEE Explore, AIP, IET, AVS, IBM, Semantic Scholar, CiteSeerX, BibSonomy, MathSciNet, PubMed, RePEc, zbMATH gibi birçok farklı veri tabanları kullanılabilir. Bu veri tabanlarında makaleye ilişkin özet, makale adı, anahtar kelimeler, yayınlanma tarihi, DOI, ISSN ve ISBN bigileri, kontrol ve index terimleri gibi birçok veri elde edilebilir. Ayrıca makale seçimleri, konferans, dergi, kitap, eğitim kursu gibi yayınlanma türüne göre de ayrılabilir. Bu veri kümeleri kullanım amaçlı olarak herkese açıktır. Bu çalışmada IEEE Explore veritabanından rastgele belirlenen 20 adet yazara ait;

- makale isimleri,
- makalelerin yayınlanma yılları,
- makalelerin anahtar kelimeleri
- anahtar kelimelerin kaç kez kullanıldığı

verileri Microsoft Visual Studio. Net programlama aracı yardımıyla elde edilmiştir. Alınan veriler MS SQL Server veritabanına aktarılmıştır.

3.2. Ön İşlem (Pre-Processing)

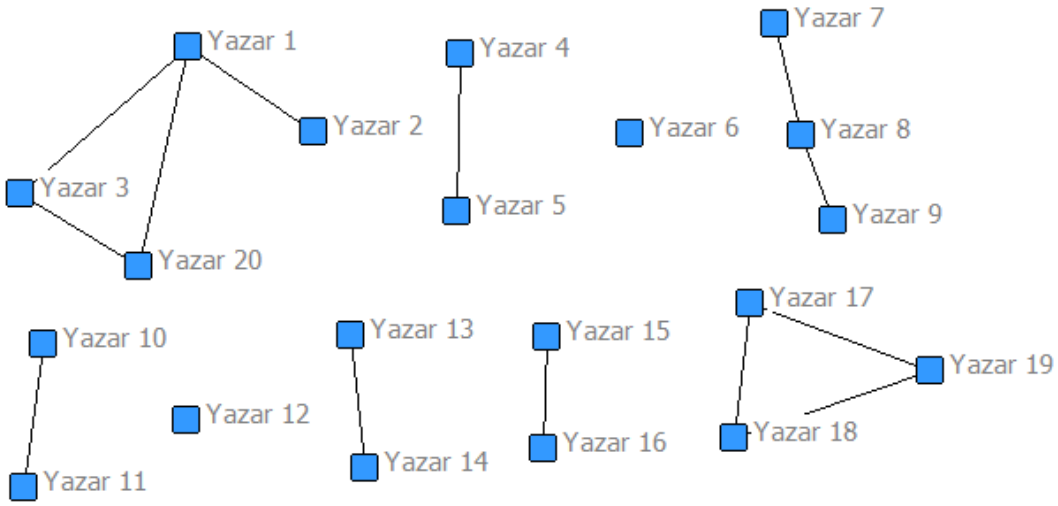
Yazarlara ilişkin veritabanından alınan tüm veriler incelenerek, gerçekleştirilen uygulama için uygun parametreler belirlenmiştir. IEEE Explore sisteminden alınabilecek parametreler; yazar adı, yayın yılı, kurum, yayın başlığı, anahtar kelimeler, kontrol terimleri, özet, ISSN, ISBN'dir. Bu parametrelerden yazar adı ve yayınlanan makalelerin anahtar kelimeleri uygulama için yeterli görüldüğünden filtrelenerek alınmıştır. Özet, ISSN ve ISBN ise benzerlik için hiçbir anlam ifade etmemektedir.

3.3. Bağlantı Tahmini Yöntemlerinin Uygulanması (Application of Link Prediction Methods)

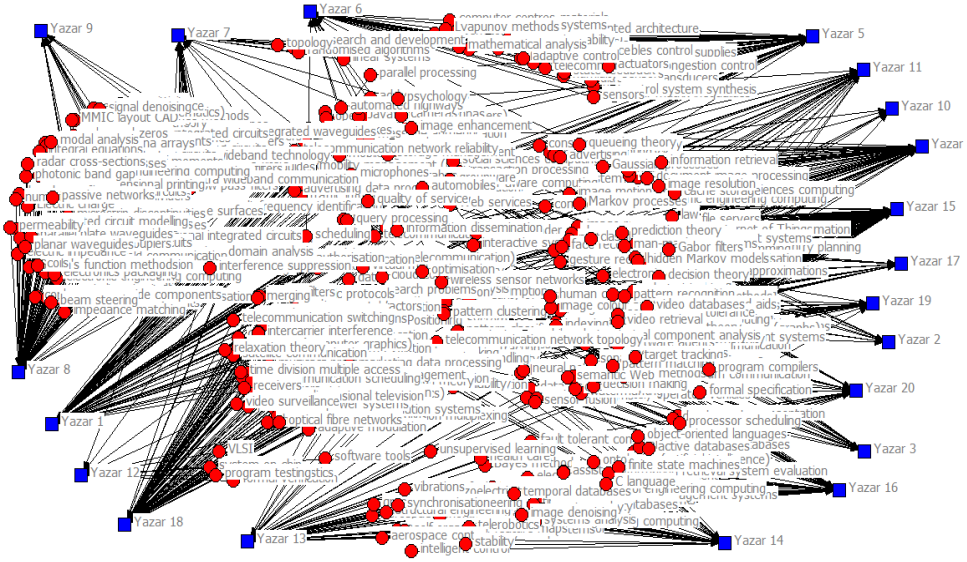
Belirlenen veri kümesine *Jaccard Index, Sorensen Index, Ortak Komşu Index, L. H. Newman Index ve Salton Index* formülasyondaki temsil biçimleri açısından bu yöntemlere uygun olmasıdır [32]. Bu yöntemlerin verilere uygulanmasında, Microsoft Visual Studio .Net ortamında C# programlama dili ile geliştirilen bir program aracı kullanılmıştır. Ayrıca Yazar-Yazar Ağı ve Yazar-Konu Ağı kullanılarak Ucinet program aracı ile çizilen yapı Şekil 3 ve Şekil 4'de gösterilmiştir.

3.4. Uygulama Sonuçları (Application Results)

Yapılan uygulama ile on adet yazarın çeşitli konferanslarda yayınladığı makalelerin anahtar kelimeleri arasındaki benzerlikler beş farklı matematiksel model uygulanarak analiz edilmiş ve sonuçları aşağıda 5 farklı tabloda verilmiştir. Yazarlar



Şekil 3. Yazar-yazar ilişki ağı (author – author network)



Şekil 4. Yazar-yazar ilişki ağı (author – subject network)

komşuluk tabanlı bağlantı tahmini yöntemleri uygulanmıştır. Bu yöntemlerin seçilmesindeki en önemli etmenler, yöntemlerin uygulanma kolaylığı ve eldeki verilerin rastgele seçilmiş, IEEE Explore sisteminde yayınlanmış olan konferans makaleleri veri seti hazırlamak için tercih edilmiştir.

Tablo 1: Jaccard İndeks uygulaması sonuçları (application results for Jaccard Index)

Jaccard	Yazar 1	Yazar 2	Yazar 3	Yazar 4	Yazar 5	Yazar 6	Yazar 7	Yazar 8	Yazar 9	Yazar 10	Yazar 11	Yazar 12	Yazar 13	Yazar 14	Yazar 15	Yazar 16	Yazar 17	Yazar 18	Yazar 19	Yazar 20
Yazar 1	-	0,10976	0,24390	0,06818	0,05128	0,01852	0,01869	0,01685	-	0,02299	0,08491	0,02198	0,03968	0,04630	0,06780	0,02362	0,03261	0,11765	0,01031	0,13978
Yazar 2	0,10976	-	0,16000	0,01493	0,02041	-	-	-	-	-	0,02439	-	0,01754	0,02564	0,01370	-	-	0,01739	-	0,03125
Yazar 3	0,24390	0,16000	-	0,02597	0,03390	0,02128	0,02174	-	-	0,03846	0,06000	0,03333	0,02985	0,04082	0,04959	-	0,06452	0,05785	0,02857	0,33333
Yazar 4	0,06818	0,01493	0,02597	-	0,08696	0,01163	0,02381	-	-	0,06452	0,06977	-	0,02857	0,03448	0,22059	0,11458	0,04348	0,07742	0,11940	-
Yazar 5	0,05128	0,02041	0,03390	0,08696	-	0,02985	0,01493	-	-	0,02128	0,02778	0,01961	-	-	0,01370	-	0,03846	0,04930	0,01786	0,01563
Yazar 6	0,01852	-	0,02128	0,01163	0,02985	-	0,03774	-	-	0,02941	0,05172	-	0,02667	-	0,03053	-	0,05128	0,03817	0,02326	-
Yazar 7	0,01869	-	0,02174	0,02381	0,01493	0,03774	-	0,13514	0,07895	0,03030	0,05263	-	-	-	0,01515	-	0,02564	0,01504	0,02381	-
Yazar 8	0,01685	-	-	-	-	-	0,13514	-	0,14141	-	-	0,00917	-	-	-	-	-	0,01471	-	-
Yazar 9	-	-	-	-	-	-	0,07895	0,14141	-	-	-	0,04167	-	-	-	-	-	-	-	-
Yazar 10	0,02299	-	0,03846	0,06452	0,02128	0,02941	0,03030	-	-	-	0,14286	-	-	-	0,03636	0,07843	0,11111	0,02679	0,15000	-
Yazar 11	0,08491	0,02439	0,06000	0,06977	0,02778	0,05172	0,05263	-	-	0,14286	-	0,02326	0,03797	0,03226	0,06061	0,06579	0,04545	0,09302	0,11364	0,05556
Yazar 12	0,02198	-	0,03333	-	0,01961	-	-	0,00917	0,04167	-	0,02326	-	0,01695	-	0,00855	0,01724	-	-	-	0,06061
Yazar 13	0,03968	0,01754	0,02985	0,02857	-	0,02667	-	-	-	-	0,03797	0,01695	-	0,50943	0,05405	0,05435	-	0,01948	0,04839	0,02817
Yazar 14	0,04630	0,02564	0,04082	0,03448	-	-	-	-	-	-	0,03226	-	0,50943	-	0,03759	0,03947	-	0,00725	0,02174	0,03774
Yazar 15	0,06780	0,00870	0,04959	0,22059	0,01370	0,03053	0,01515	-	-	0,03636	0,06061	0,00855	0,05405	0,03759	-	0,15672	0,03448	0,08040	0,06034	0,01550
Yazar 16	0,02362	-	-	0,11458	-	-	-	-	-	0,07843	0,06579	0,01724	0,05435	0,03947	0,15672	-	0,01667	0,03311	0,06667	-
Yazar 17	0,03261	-	0,06452	0,04348	0,03846	0,05128	0,02564	-	-	0,11111	0,04545	-	-	-	0,03448	0,01667	-	0,12037	0,16000	0,02778
Yazar 18	0,11765	0,01739	0,05785	0,07742	0,04930	0,03817	0,01504	0,01471	-	0,02679	0,09302	-	0,01948	0,00725	0,08040	0,03311	0,12037	-	0,10714	0,03937
Yazar 19	0,01031	-	0,02857	0,11940	0,01786	0,02326	0,02381	-	-	0,15000	0,11364	-	0,04839	0,02174	0,06034	0,06667	0,16000	0,10714	-	-
Yazar 20	0,13978	0,03125	0,33333	-	0,01563	-	-	-	-	-	0,05556	0,06061	0,02817	0,03774	0,01550	-	0,02778	0,03937	-	-

Tablo 2: Sorensen Indeks Uygulama sonuçları (application results for Sorensen Index)

Sorensen	Yazar 1	Yazar 2	Yazar 3	Yazar 4	Yazar 5	Yazar 6	Yazar 7	Yazar 8	Yazar 9	Yazar 10	Yazar 11	Yazar 12	Yazar 13	Yazar 14	Yazar 15	Yazar 16	Yazar 17	Yazar 18	Yazar 19	Yazar 20
Yazar 1	-	0,19780	0,39216	0,12766	0,09756	0,03636	0,03670	0,03315	-	0,04494	0,15652	0,04301	0,07634	0,08850	0,12698	0,04615	0,06316	0,21053	0,02041	0,24528
Yazar 2	0,19780	-	0,27586	0,02941	0,04000	-	-	-	-	-	0,04762	-	0,03448	0,05000	0,01724	-	-	0,03419	-	0,06061
Yazar 3	0,39216	0,27586	-	0,05063	0,06557	0,04167	0,04255	-	-	0,07407	0,11321	0,06452	0,05797	0,07843	0,09449	-	0,12121	0,10938	0,05556	0,30000
Yazar 4	0,12766	0,02941	0,05063	-	0,16000	0,02299	0,04651	-	-	0,12121	0,13043	-	0,05556	0,06667	0,36145	0,20561	0,08333	0,14371	0,21333	-
Yazar 5	0,09756	0,04000	0,06557	0,16000	-	0,05797	0,02941	-	-	0,04167	0,05405	0,03846	-	-	0,02703	-	0,07407	0,09396	0,03509	0,33077
Yazar 6	0,03636	-	0,04167	0,02299	0,05797	-	0,07273	-	-	0,05714	0,09836	-	0,05195	-	0,05926	-	0,09756	0,07333	0,04545	-
Yazar 7	0,03670	-	0,04255	0,04651	0,02941	0,07273	-	0,23810	0,14634	0,05882	0,10000	-	-	-	0,02985	-	0,05000	0,02963	0,04651	-
Yazar 8	0,03315	-	-	-	-	-	0,23810	-	0,24779	-	-	0,01818	-	-	-	-	-	0,02899	-	-
Yazar 9	-	-	-	-	-	-	0,14634	0,24779	-	-	-	0,08000	-	-	-	-	-	-	-	-
Yazar 10	0,04494	-	0,07407	0,12121	0,04167	0,05714	0,05882	-	-	-	0,25000	-	-	-	0,07018	0,14545	0,20000	0,05217	0,26087	-
Yazar 11	0,15652	0,04762	0,11321	0,13043	0,05405	0,09836	0,10000	-	-	0,25000	-	0,04545	0,07317	0,06250	0,11429	0,12346	0,08696	0,17021	0,20408	0,10526
Yazar 12	0,04301	-	0,06452	-	0,03846	-	-	0,01818	0,08000	-	0,04545	-	0,03333	-	0,01695	0,03390	-	-	-	0,11429
Yazar 13	0,07634	0,03448	0,05797	0,05556	-	0,05195	-	-	-	-	0,07317	0,03333	-	0,67500	0,10256	0,10309	-	0,03822	0,09231	0,05479
Yazar 14	0,08850	0,05000	0,07843	0,06667	-	-	-	-	-	-	0,06250	-	0,67500	-	0,07246	0,07595	-	0,01439	0,04255	0,07273
Yazar 15	0,12698	0,01724	0,09449	0,36145	0,02703	0,05926	0,02985	-	-	0,07018	0,11429	0,01695	0,10256	0,07246	-	0,27097	0,06667	0,14884	0,11382	0,33053
Yazar 16	0,04615	-	-	0,20561	-	-	-	-	-	0,14545	0,12346	0,03390	0,10309	0,07595	0,27097	-	0,03279	0,06410	0,12500	-
Yazar 17	0,06316	-	0,12121	0,08333	0,07407	0,09756	0,05000	-	-	0,20000	0,08696	-	-	-	0,06667	0,03279	-	0,21488	0,27586	0,35405
Yazar 18	0,21053	0,03419	0,10938	0,14371	0,09396	0,07353	0,02963	0,02899	-	0,05217	0,17021	-	0,03822	0,01439	0,14884	0,06410	0,21488	-	0,19355	0,07576
Yazar 19	0,02041	-	0,05556	0,21333	0,03509	0,04545	0,04651	-	-	0,26087	0,20408	-	0,09231	0,04255	0,11382	0,12500	0,27586	0,19355	-	-
Yazar 20	0,24528	0,06061	0,30000	-	0,03077	-	-	-	-	-	0,10526	0,11429	0,05479	0,07273	0,03053	-	0,05405	0,07576	-	-

Tablo 3: Ortak Komşu uygulama sonuçları (application results for Common Neighbour Index)

Ortak Komşu	Yazar 1	Yazar 2	Yazar 3	Yazar 4	Yazar 5	Yazar 6	Yazar 7	Yazar 8	Yazar 9	Yazar 10	Yazar 11	Yazar 12	Yazar 13	Yazar 14	Yazar 15	Yazar 16	Yazar 17	Yazar 18	Yazar 19	Yazar 20
Yazar 1	-	9	20	9	6	2	2	3	-	2	9	2	5	5	12	3	3	20	1	13
Yazar 2	9	-	4	1	1	-	-	-	-	-	1	-	1	1	1	-	-	2	-	1
Yazar 3	20	4	-	2	2	1	1	-	-	1	3	1	2	2	6	-	2	7	1	11
Yazar 4	9	1	2	-	8	1	2	-	-	4	6	-	3	3	30	11	3	12	8	-
Yazar 5	6	1	2	8	-	2	1	-	-	1	2	1	-	-	2	-	2	7	1	1
Yazar 6	2	-	1	1	2	-	2	-	-	1	3	-	2	-	4	-	2	5	1	-
Yazar 7	2	-	1	2	1	2	-	15	3	1	3	-	-	-	2	-	1	2	1	-
Yazar 8	3	-	-	-	-	-	15	-	14	-	-	1	-	-	-	-	-	3	-	-
Yazar 9	-	-	-	-	-	-	3	14	-	-	-	1	-	-	-	-	-	-	-	-
Yazar 10	2	-	1	4	1	1	1	-	-	-	5	-	-	-	-	-	2	3	3	-
Yazar 11	9	1	3	6	2	3	3	-	-	5	-	1	3	2	8	5	2	12	5	3
Yazar 12	2	-	1	-	1	-	-	1	1	-	1	-	1	-	1	1	-	-	-	2
Yazar 13	5	1	2	3	-	2	-	-	-	-	3	1	-	27	8	5	-	3	3	2
Yazar 14	5	1	2	3	-	-	-	-	-	-	2	-	27	-	5	3	-	1	1	2
Yazar 15	12	1	6	30	2	4	2	-	-	4	8	1	8	5	-	21	4	16	7	2
Yazar 16	3	-	-	11	-	-	-	-	-	4	5	1	5	3	21	-	1	5	4	-
Yazar 17	3	-	2	3	2	2	1	-	-	2	2	-	-	-	4	1	-	13	4	1
Yazar 18	20	2	7	12	7	5	2	3	-	3	12	-	3	1	16	5	13	-	12	3
Yazar 19	1	-	1	8	1	1	1	-	-	3	5	-	3	1	7	4	4	12	-	-
Yazar 20	13	1	11	-	1	-	-	-	-	-	3	2	2	2	2	-	1	5	-	-

Tablo 4. L. H. Newman İndeks Uygulama Sonuçları (application results for L. H. Newman index)

L. H. Newman	Yazar 1	Yazar 2	Yazar 3	Yazar 4	Yazar 5	Yazar 6	Yazar 7	Yazar 8	Yazar 9	Yazar 10	Yazar 11	Yazar 12	Yazar 13	Yazar 14	Yazar 15	Yazar 16	Yazar 17	Yazar 18	Yazar 19	Yazar 20
Yazar 1	-	0,02439	0,02439	0,00372	0,00357	0,00174	0,00181	0,00074	-	0,00697	0,00665	0,00443	0,00249	0,00393	0,00274	0,00152	0,00563	0,00452	0,00152	0,01321
Yazar 2	0,02439	-	0,04444	0,00377	0,00542	-	-	-	-	-	0,00673	-	0,00454	0,00717	0,00208	-	-	0,00412	-	0,00926
Yazar 3	0,02439	0,04444	-	0,00339	0,00488	0,00357	0,00370	-	-	0,01429	0,00909	0,00909	0,00408	0,00645	0,00561	-	0,01538	0,00648	0,00625	0,04583
Yazar 4	0,00372	0,00377	0,00339	-	0,00661	0,00121	0,00251	-	-	0,01937	0,00616	-	0,00208	0,00328	0,00950	0,00777	0,00782	0,00377	0,01692	-
Yazar 5	0,00357	0,00342	0,00488	0,00661	-	0,00348	0,00181	-	-	0,00697	0,00296	0,00443	-	-	0,00091	-	0,00750	0,00316	0,00305	0,00203
Yazar 6	0,00174	-	0,00357	0,00121	0,00348	-	0,00529	-	-	0,01020	0,00649	-	0,00292	-	0,00267	-	0,01099	0,00331	0,00446	-
Yazar 7	0,00181	-	0,00370	0,00251	0,00181	0,00529	-	0,01122	0,01587	0,01058	0,00673	-	-	-	0,00138	-	0,00570	0,00137	0,00463	-
Yazar 8	0,00074	-	-	-	-	-	0,01122	-	0,02020	-	-	0,00184	-	-	-	-	-	0,00056	-	-
Yazar 9	-	-	-	-	-	-	0,01587	0,02020	-	-	-	0,01299	-	-	-	-	-	-	-	-
Yazar 10	0,00697	-	0,01429	0,01937	0,00697	0,01020	0,01058	-	-	-	0,04329	-	-	-	0,01068	0,02381	0,04396	0,00794	0,05357	-
Yazar 11	0,00665	0,00673	0,00909	0,00616	0,00296	0,00649	0,00673	-	-	0,04329	-	0,00551	0,00371	0,00391	0,00453	0,00631	0,00932	0,00673	0,01894	0,00758
Yazar 12	0,00443	-	0,00909	-	0,00443	-	-	0,00184	0,01299	-	0,00551	-	0,00371	-	0,00170	0,00379	-	-	-	0,01515
Yazar 13	0,00249	0,00454	0,00408	0,00208	-	0,00292	-	-	-	-	0,00371	0,00371	-	0,03555	0,00305	0,00425	-	0,00113	0,00765	0,00340
Yazar 14	0,00393	0,00717	0,00645	0,00328	-	-	-	-	-	-	0,00391	-	0,03555	-	0,00301	0,00403	-	0,00060	0,00403	0,00538
Yazar 15	0,00274	0,00208	0,00561	0,00950	0,00091	0,00267	0,00138	-	-	0,01068	0,00453	0,00170	0,00305	0,00301	-	0,00818	0,00575	0,00277	0,00818	0,00156
Yazar 16	0,00152	-	-	0,00777	-	-	-	-	-	0,02381	0,00631	0,00379	0,00425	0,00403	0,00818	-	0,00321	0,00193	0,01042	-
Yazar 17	0,00563	-	0,00625	0,00782	0,00750	0,01099	0,00570	-	-	0,04396	0,00932	-	-	-	0,00575	0,00321	-	0,01852	0,03846	0,00641
Yazar 18	0,00452	0,00412	0,00648	0,00377	0,00316	0,00331	0,00137	0,00056	-	0,00794	0,00673	-	ssss	0,00060	0,00277	0,00193	0,01852	-	0,01389	0,00386
Yazar 19	0,00152	-	-	0,01695	0,00305	0,00446	0,00463	-	-	0,05357	0,01894	-	0,00765	0,00403	0,00818	0,01042	0,03846	0,01389	-	-
Yazar 20	0,01321	0,00926	0,04583	-	0,00203	-	-	-	-	-	0,00758	0,01515	0,00340	0,00538	0,00156	-	0,00641	0,00386	-	-

Tablo 5: Salton İndeks Uygulama Sonuçları (application results for Salton Index)

Salton	Yazar 1	Yazar 2	Yazar 3	Yazar 4	Yazar 5	Yazar 6	Yazar 7	Yazar 8	Yazar 9	Yazar 10	Yazar 11	Yazar 12	Yazar 13	Yazar 14	Yazar 15	Yazar 16	Yazar 17	Yazar 18	Yazar 19	Yazar 20
Yazar 1	-	0,94346	1,98030	0,75794	0,34100	0,19069	0,19157	0,22299	-	0,21200	0,83925	0,20739	0,43685	0,47036	0,87287	0,26312	0,30779	1,45095	0,10102	1,26267
Yazar 2	0,94346	-	0,74278	0,12127	0,14142	-	-	-	-	-	0,15430	-	0,13131	0,15811	0,09285	-	-	0,18490	-	0,17408
Yazar 3	1,98030	0,74278	-	0,22502	0,25607	0,14434	0,14586	-	-	0,19245	0,41208	0,17961	0,24077	0,28006	0,33241	-	0,34816	0,61872	0,16667	1,65831
Yazar 4	0,75794	0,12127	0,22502	-	0,80000	0,10721	0,21567	-	-	0,49237	0,62554	-	0,28868	0,31623	2,32845	1,06341	0,35355	0,92859	0,92376	-
Yazar 5	0,54100	0,14142	0,25607	0,80000	-	0,24077	0,12127	-	-	0,14434	0,23250	0,13868	-	-	0,16440	-	0,27217	0,57346	0,13245	0,12403
Yazar 6	0,19069	-	0,14434	0,10721	0,24077	-	0,26968	-	-	0,16903	0,38411	-	0,22792	-	0,34427	-	0,31235	0,42875	0,15076	-
Yazar 7	0,19157	-	0,14586	0,21567	0,12127	0,26968	-	1,33631	0,46852	0,17150	0,38730	-	-	-	0,17277	-	0,15811	0,17213	0,15250	-
Yazar 8	0,22299	-	-	-	-	-	1,33631	-	1,31701	-	-	0,09535	-	-	-	-	-	0,20851	-	-
Yazar 9	-	-	-	-	-	-	0,46852	1,31701	-	-	-	0,20000	-	-	-	-	-	-	-	-
Yazar 10	0,21200	-	0,19245	0,49237	0,14434	0,16903	0,17150	-	-	-	0,79057	-	-	-	0,37463	0,53936	0,44721	0,27975	0,62554	-
Yazar 11	0,83925	0,15430	0,41208	0,62554	0,23250	0,38411	0,38730	-	-	0,79057	-	0,15076	0,33129	0,25000	0,67612	0,55556	0,29488	1,01058	0,71429	0,39736
Yazar 12	0,20739	-	0,17961	-	0,13868	-	-	0,09535	0,20000	-	0,15076	-	0,12910	-	0,09206	0,13019	-	-	-	0,33806
Yazar 13	0,43685	0,13131	0,24077	0,28868	-	0,22792	-	-	-	-	0,33129	0,12910	-	3,01869	0,64051	0,50767	-	0,23943	0,37210	0,23408
Yazar 14	0,47036	0,15811	0,28006	0,31623	-	-	-	-	-	-	0,25000	-	3,01869	-	0,42563	0,33753	-	0,08482	0,14586	0,26968
Yazar 15	0,87287	0,09285	0,33241	2,32845	0,16440	0,34427	0,17277	-	-	0,37463	0,67612	0,09206	0,64051	0,42563	-	1,68676	0,36515	1,09119	0,63117	0,17434
Yazar 16	0,26312	-	-	1,06341	-	-	-	-	-	0,53936	0,55556	0,13019	0,50767	0,33753	1,68676	-	0,12804	0,40032	0,50000	-
Yazar 17	0,30779	-	0,34816	0,35355	0,27217	0,31235	0,15811	-	-	0,44721	0,29488	-	-	-	0,36515	0,12804	-	1,18182	0,74278	0,16440
Yazar 18	1,45095	0,18490	0,61872	0,92859	0,57346	0,42875	0,17213	0,20851	-	0,27975	1,01058	-	0,23943	0,08482	1,09119	0,40032	1,18182	-	1,07763	0,43519
Yazar 19	0,10102	-	0,16667	0,92376	0,13245	0,15076	0,15250	-	-	0,62554	0,71429	-	-	0,14586	0,63117	0,50000	0,74278	1,07763	-	-
Yazar 20	1,26267	0,17408	1,65831	-	0,12403	-	-	-	-	-	0,39736	0,33806	0,23408	0,26968	0,17474	-	0,16440	0,43519	-	-

Yazar-Yazar ağı incelendiğinde bazı yazarların birlikte yayın yaptıkları görülmektedir. Yapılan bu analiz çalışmasında yazarlar arasındaki benzerlikler irdelendiğinde ortak yayın yapmamış bazı yazar çiftlerinin benzerlik değerlerinin ortak yayın yapmış olan yazarlar arası benzerlik değerinden daha yüksek olduğu görülür. Bu da bu yazarlar arasında da bir ilişki kurulabileceği (örneğin ortak yayın yapabilecekleri, yeni ilişki kurulan yazarların çalışmaları muhtemel konuların tespiti) anlamına gelir.

Yapılan uygulamada Tablo 1, yazarlar arasındaki benzerliğin Jaccard İndex yöntemi ile hesaplanan sonuçları içermektedir. Tablo 2, Sorensen İndex yöntemi ile; Tablo 3, Ortak Komşu İndex yöntemi ile; Tablo 4, L. H. Newman İndex yöntemi ile ve Tablo 5, Salton İndex yöntemi ile hesaplanan sonuçları içermektedir. Şekil 3'te yazarlar arasındaki mevcut bağlantılar gösterilmektedir. Aralarında bağlantı olan yazarlar birlikte yayın yapmış olan yazarlardır. Yazarlar arasında oluşması muhtemel yeni bağlantıları tahmin etmek için Şekil 3'te birbiri ile bağlantılı olan yazar çiftlerinden her tablo için bir yöntem kullanılarak benzerlik değerleri hesaplanmış ve her tabloda bu yazarlar arasındaki en büyük değer eşik değer olarak belirlenmiştir.

Tablo 1-5 dikkate alındığında yazar 1 ile yazar 2, yazar 1 ile yazar 3 ve yazar 1 ile yazar 20 arasında oluşan benzerlik değerleri incelendiğinde yazar 1 ile yazar 3'ten elde edilen değerler en büyük değerler olduğundan eşik değerleri olarak belirlenmiştir. Tablolar ayrı ayrı analiz edildiğinde beş tabloda da belirlenen eşik değerlerinden büyük tek değerin yazar 1 ile yazar 18 arasında oluşan benzerlik değeri olduğu görülmektedir. Uygulanan beş yöntemde de bu değerlerin eşik değerlerinden yüksek çıkması yazar 1 ile yazar 18 arasında yeni bir bağlantı oluşabileceği anlamına gelmektedir.

Aynı şekilde Şekil 3'ten yola çıkarak, yazar 2 sadece yazar 1 ile bağlantılı olduğundan eşik değerleri beş yöntemde de yazar 2 ile yazar 1 arasında oluşan benzerlik değerleri olur. Tüm tablolar ayrı ayrı incelendiğinde Tablo 1, Tablo 2, Tablo 4 ve Tablo 5'te yazar 2 ile yazar 3 arasındaki benzerlik değerlerinin eşik değerlerden büyük olduğu görülmektedir. Bu durumda Jaccard İndex, Sorensen İndex, L. H. Newman İndex ve Salton İndex yöntemlerine göre yazar 2 ile yazar 3 arasında yeni bir bağlantı oluşabilmektedir. Sadece Tablo 3'te dolayısı ile Ortak Komşu yöntemine göre bu iki yazar arasında bağlantı oluşmamaktadır.

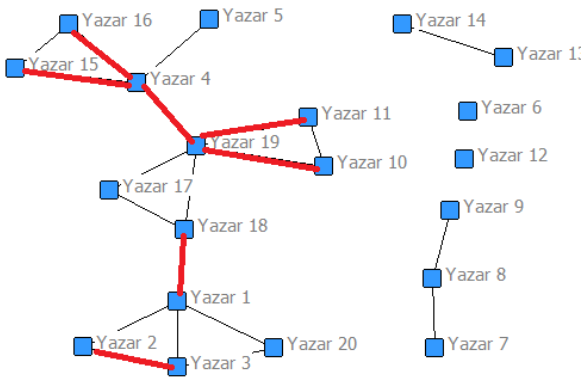
Yazar 4 ile sadece yazar 5 arasında bağlantı olduğundan eşik değerleri yazar 4 ile yazar 5 arasında hesaplanan değerlerdir. Tablolar incelendiğinde Yazar 4 ile yazar 15, yazar 4 ile yazar 16 ve yazar 4 ile yazar 19 arasındaki benzerlik değerlerinin tüm tablolarda eşik değerlerinden büyük olduğu görülmektedir. Bu durumda Jaccard İndex, Sorensen İndex, Ortak Komşu İndex, L. H. Newman İndex ve Salton İndex yöntemlerine göre yazar 4 ile yazar 15, yazar 4 ile yazar 16 ve yazar 4 ile yazar 19 arasında da bağlantılar oluşabileceği söylenebilir.

Yazar 10 ile sadece yazar 11 bağlantılı olduğundan eşik değerleri bu iki yazar arasında hesaplanan değerler olur. Tablolar incelendiğinde yazar 10 ile yazar 19 arasındaki benzerlik değerlerinin Tablo 1, Tablo 2, Tablo 4 ve Tablo 5'te eşik değerlerinden büyük olduğu görülür. Sadece Ortak Komşu İndex yönteminde bu değer sağlanmamaktadır. Bu durumda yazar 10 ile yazar 19 arasında da bir bağlantı olabileceği söylenebilir.

Yazar 15 ile sadece yazar 16'nın bağlantısı olduğundan burada da eşik değerleri yazar 15 ile yazar 16 arasında hesaplanan değerlerdir. Tablolar analiz edildiğinde yazar 15 ile yazar 4 arasındaki benzerlik değerlerinin beş yöntemin tamamında eşik değerlerinden yüksek olduğu gözlenmektedir. Bu durumda yazar 15 ile yazar 4 arasında da bağlantı oluşabileceği söylenebilir.

Yazar 19 ile yazar 17 ve yazar 18 arasında bağlantı bulunmaktadır. Tablolar analiz edildiğinde yazar 19 ile yazar 17 arasında hesaplanan değerlerin daha büyük olduğu gözlenir. Yani eşik değerleri bu iki yazar arasında hesaplanan değerlerdir. Tablolar incelenince yazar 19 ile yazar 10 arasındaki benzerlik değerlerinin Tablo 1, Tablo 2, Tablo 4 ve Tablo 5'te; yazar 19 ile yazar 4 ve yazar 19 ile yazar 11 arasındaki benzerlik değerlerinin ise yöntemlerin tamamında eşik değerlerinden büyük olduğu gözlenmektedir. Bu durumda yazar 19 ile yazar 4, yazar 19 ile yazar 10 ve yazar 19 ile yazar 11 arasında yeni bağlantı oluşabileceği söylenebilir.

Bu durumda yeni Yazar-Yazar ağı şekil 5 te gösterilmiştir, yeni oluşan bağlantılar kırmızı renktedir:



Şekil 5. Yeni Yazar-Yazar ağı (New aouthor-author network)

4. SONUÇ VE ÖNERİLER (CONCLUSIONS AND RECOMMENDATIONS)

Özellikle sosyal ağların gelişmesi ile birlikte bağlantı tahmini problemi popüler bir hal almıştır. Günümüzde ticari anlamda yaygın olarak kullanılan bağlantı tahmini probemi akademik olarak;

- Sosyal ağlarda kullanıcıların arkadaşlık ilişkilerini tahmin etme,
- Yazar-makale ağlarında yazarların gelecekte yapacakları yayın konularını tahmin etme ve yazarların ortak yayın çıkarma olasılığını belirleme,

- Hastalık-ilaç ağlarında hastalıkları tedavi etmede kullanılabilecek ilaçların belirlenmesi gibi alanlarda yoğun bir şekilde kullanılmaktadır.

Bağlantı tahmininin başarım oranının yüksek olabilmesi için seçilecek kriterlerin ve parametrelerin doğru belirlenmesi gerekmektedir.

Ayrıca bağlantı tahmini problemini çözme yöntemleri graf yapıları üzerinde uygulandığından graf yapılarının da gayet iyi bilinmesi gerekmektedir.

Bu çalışmada bağlantı tahmini yöntemleri akademik alanda yazarların çalışma konularına uygulanmış ve çalışma konuları arasındaki benzerlikler ortaya konmuştur. Özellikle yerel indeksleme yöntemlerinden Jaccard Index, Sorensen Index, Ortak Komşu, L. H. Newman Index, Salton Index yöntemlerinin analiz için uygun olduğu ve başarılı sonuçlar vereceği düşünülerek analiz bu yöntemlerle yapılmıştır.

Yapılan analizler sonucunda veri setindeki yazarların çalışma alanları arasında benzerlikler tespit edilmiştir. Özellikle bu benzerliklerden yola çıkarak hangi yazarların birbiri ile birlikte çalışma yapabileceği hakkında tahminler yapılmıştır. Tahmin için yapılan hesaplamalarda görülmüştür ki, ortak yayın veya ortak anahtar kelime sayısı (ortak komşu indeks yönteminde net olarak görülmektedir) tek başına doğru sonuçlar vermemektedir.

Yapılan çalışmada; yazar adı, yayın yılı, kurum, yayın başlığı, anahtar kelimeler, kontrol terimleri, özet, ISSN, ISBN parametreleri IEEE Explore veritbanından çekilebilmektedir. Bu parametreler arasından yazar adı ve anahtar kelimeler benzerlik ölçümü için en baskın terimler olarak tespit edilmiştir. Özet, yayın başlığı, ISSN ve ISBN'nin benzerlik ölçümü için yararlı özellikler olmadığı ve yayın yılı, kurum, kontrol terimlerinin ise yazar adı ve anahtar kelimeler ile birlikte kullanılarak benzerlik ölçümünün daha hassas bir şekilde tahmin yapılması sağlanabilir. Bu parametrelerin eklenmesinin dezavantajı ise işlem yükünün artması ve eklenecek parametrelerin (düğümlerin) ağırlıklarının hesaplanma zorluğudur. Gelecekteki çalışmalarda bu parametreler eklenerek daha detaylı çalışmalar yapılması amaçlanmaktadır.

KAYNAKLAR (REFERENCES)

[1] Lü L., and Zhou T., "Physica A: statistical mechanics and its applications", ***Physica A: Statistical Mechanics and its Applications***, 390(6): 1150-1170, (2011).

[2] Tuğal İ., "Sosyal Ağlarda Hastalık İlaç Bağlantı Tahmini", ***Yüksek Lisans Tezi***, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, (2013).

[3] Huang T. H. and Huang M. L., "Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of ındividual researchers", ***Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*** (CGIV'06), Sydney, Australia, (2006).

[4] Sun Y., Barber R., Gupta M., Aggarwal C. C. And Han J., "Co-author relationship prediction in heterogeneous bibliographic networks", ***International Conference on Advances in Social Networks Analysis and Mining***, Kaohsiung, Taiwan, 121-128, (2011).

[5] Pengbin G., Weiwei W. and Bo Y., "Co-authorship network analysis in ımprovisation theory research", ***International Conference on Information Management, Innovation Management and Industrial Engineering***, Sanya, China, 244-248, (2012).

[6] Li L. and Xuezhu G., "Innovation performance of university co-authorship network", ***International Conference on Information Management, Innovation Management and Industrial Engineering***, Sanya, China, 410-413, (2012).

[7] Anastasios T., Sgouropoulou C., Papageorgiou E., Terraz O. and Miaoulis G., "Co-authorship networks in academic research communities: the role of network strength", ***16th Panhellenic Conference on Informatics***, Piraeus, Greece, 150-155, (2012).

[8] Bidault F., and Hildebrand T., "The distribution of partnership returns: Evidence from co-authorships in economics journals", ***Research Policy***, 43(6): 1002-1013, (2014).

[9] Türker İ. and Çavuşoğlu A., "Detailing the co-authorship networks in degree coupling, edge weight and academic age perspective", ***Chaos, Solitons and Fractals***, *91*: 386-392, (2016).

[10] Bruno B., "Economics of co-authorship", ***Economic Analysis and Policy***, 44(2): 212-220, (2014).

[11] De Stefano D., Fuccella V., Vitale M. P. and Zaccarin S., "The use of different data sources in the analysis of co-authorship networks and scientific performance", ***Social Networks***, 35(3): 370-381, (2013).

[12] Ortega J. L., "Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search", ***Journal of Informetrics***, 8(3): 728-737, (2014).

[13] Koseoglu M. A., "Growth and structure of authorship and co-authorship network in the strategic management realm: evidence from the strategic management journal", ***BRQ Business Research Quarterly***, 19(3): 153-170, (2016).

[14] Sharma D. and Sharma U., "Link prediction algorithm for co-authorship networks using neural networks", ***IEEE ICRITO***, (2014).

[15] Newman M. E. J., "Scientific collaboration networks", ***İ.Network Construction And Fundamental Results. Physical Review E***, 64: 016131, **(**2001**).**

[16] Barabási A. L., **"**From networks to human behavior", ***In Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval***, 435, **(**2009).

[17] Wang C., Satuluri V. and Parthasarathy S., "Local probabilistic models for link prediction", ***In Proceedings of 7th IEEE International Conference on Data Mining,*** 322–331, (2007).

[18] Xiang E. W., "A survey on link prediction models for social network data", ***PhD thesis***, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, (2008).

[19] Newman M., E., J., "Clustering and preferential attachment in growing networks", ***Phys. Rev. E***, 64(2): (2001).

[20] Müller B., Hagelstein A. and Gübitz T. "Life science ontologies in literature retrieval: A comparison of linked data sets for use on semantic search on a heterogeneous corpus", In ***Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management***, Bologna, Italy, (2016).

[21] Kushwah A. K. S., and Manjhvar A. K., "A review on link prediction in social network", ***International Journal of Grid and Distributed Computing***, *9*(2): 43-50, (2016).

[22] Sesli M. and Yegenoglu E. D., "Comparison of similarity coefficients used for cluster analysis based on RAPD markers in wild olives", ***Genet Mol. Res.*** 9: 2248–2253, (2010).

[23] Leicht E. A., Holme P. and Newman M. E., "Vertex similarity in networks", ***Physical Review E***, 73(2): 026120, (2006).

[24] Adamic L. A. and Adar E., "Friends and neighbors on the web", ***Social networks***, 25(3): 211-230, (2003).

[25] Barabási A. L. and Albert R., "Emergence of scaling in random networks", ***Science***, *286*(5439): 509-512, (1999).

[26] Zhou T., Lü L. and Zhang Y. C., "Predicting missing links via local information", ***The European Physical Journal B***, 71(4): 623-630, (2009).

[27] Newman M., "Networks: an introduction", ***Oxford University Press***, (2010).

[28] Arik G., Varan H. D., Yavuz B. B., Karabulut E., Kara O., Kilic M. K., ... and Halil M., "Validation of Katz index of independence in activities of daily living in Turkish older adults", ***Archives of Gerontology and Geriatrics***, 61(3): 344-350, (2015).

[29] Fouss F., Pirotte A., Renders J. M., and Saerens M., "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation", ***IEEE Transactions on Knowledge and Data Engineering***, 19(3): 355-369, (2007).

[30] Guns R. ,"Predictive Characteristics of Co-authorship Networks: Comparing the Unweighted, Weighted, and Bipartite Cases", ***Journal of Data and Information Science***, *1*(3): 59-78, (2016).

[31] Lu J., Gong Z. and Lin X., "A novel and fast simrank algorithm", ***IEEE Transactions on Knowledge and Data Engineering***, (2016).

[32] Liben-Nowell D. and Kleinberg J., "The link prediction problem for social networks", ***Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM),*** NewYork: ACM Press, 556–559, (2003).